TRAJECTORY OF MEDICAL STUDENTS' RESEARCH INTEREST BY GENDER, RACE/ETHNICITY, RESEARCH EXPERIENCE, AND PROGRAM: A LONGITUDINAL ANALYSIS

A Dissertation

Presented to

The Faculty of the Curry School of Education

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Xiaoqing Kong, B.A., M.Ed.

May 2014

ABSTRACT

Science, technology, engineering, and mathematics (STEM) education is considered critical to a nation's economic competitiveness and national security (Congressional Research Service, 2012; National Academy of Sciences, 2007). A major concern in STEM education is students' persistence in the STEM pipeline (National Science Board, 2012). This focus is a particular issue for the biomedical research community where more diverse and expert physician-scientists are needed (Ley & Rosenberg, 2005). Most previous studies focus on post-secondary students' general program completion in the STEM related fields. This study seeks to address to some degree the paucity of research on the persistence of students' research interest[1] in the biomedical field based on a longitudinal design with a large sample size. The research questions addressed in this study were:

(1) Does medical students' reported research interest differ among students with different characteristics of gender, race/ethnicity, previous research experiences, or matriculated program prior to their entry to medical schools?

(2) Does medical students' reported research interest change in general across time from prior to their entry to medical schools, to when they are matriculated in medical schools, and to when they graduate from medical schools?

(3) Are patterns of change in medical students' reported research interest across time associated with gender, race/ethnicity, previous research experiences, or matriculated program?

[1] Throughout this dissertation, "research interest" refers to interest in research.

The data used in this study were derived from three questionnaires taken by 39,839 medical school graduates and one student record system data set assembled through Project TrEMUR (Transitions in the Education of Minorities Underrepresented in Research)[2]. After appropriate covariance structures and mean models were selected, longitudinal data analyses (Fitzmaurice, Laird, & Ware, 2004) were conducted to address the research questions. Results indicated that medical students' reported research interest differed among students with different characteristics of gender, race/ethnicity, previous research experiences, and matriculated program prior to their entry to medical schools. After considering all the variables included in the models, medical students' reported research interest decreased significantly from prior to their entry to medical schools to when they were matriculated in medical schools, and such decrease was significantly offset after their matriculation in medical schools until their graduation from medical schools. The patterns of change in medical students' reported research interest across time were significantly associated with gender, race/ethnicity, previous research experiences, and matriculated program.

[2] The three questionnaire data and the student record system data were all provided by the Association of American Medical Colleges (AAMC).

DEDICATION

For my parents, Dali Kong and Jie Liang,
who always love me and believe in me through my life,
and make me a person with positive personality.

For my grandmother, Guizhi Yao,
who cares for me with unconditional love
especially in my first 18 years when I lived with her.

For my husband, Yuchen Zhou,
who loves me and understands me,
and supports me both emotionally and professionally.
We enjoy life together and work hard together.

ACKNOWLEDGEMENTS

I would like to acknowledge the following people for their time, help and support.

Robert H. Tai, my advisor, who brought me in the doctoral program, and led me into the real world of research. While in his research group, I have gained a lot of experience from conducting research to thinking critically as a researcher, thanks to his time, guidance, and mentoring. I will forever be grateful for his tremendous help and support.

Timothy Konold, Heather Wathington, and Jennifer Chiu, my dissertation committee members, who also served on my qualifying paper committee and comprehensive exam committee. They have provided invaluable suggestions and encouragement in each big progress that I have made in my Ph.D. study.

Christine Q. Liu, who served on my dissertation committee. Her detailed and helpful comments and advice on my dissertation have helped me very much in my data analyses and dissertation writing.

The Association of American Medical Colleges (AAMC), who provided the comprehensive longitudinal data sets that led to this dissertation.

Donna Jeffe, and Dorothy Andriole, who helped me with their expert knowledge in medical education.

John Almarode, Kate Dabney, Devasmita Chakraverty, Nathan Dolenc, and Daniel Read, my colleagues in the research group, who give me advice and invite me to work with them on projects.

My parents, my grandmother, my grandfather, my aunts and uncles, and my cousins, who make up a very happy and harmonious big family that support me with their unconditional love from the other end of the earth during my Ph.D. study.

Yuchen Zhou, my husband, who always stands by my side ever since the very beginning when I started considering to apply for a Ph.D. Throughout my Ph.D. study, my communications with him about statistics and programming make me think deeper and smarter.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

Table

LIST OF FIGURES

Page

Figure

CHAPTER 1

INTRODUCTION

Development in the fields of science, technology, engineering, and mathematics (STEM) is considered a key factor of improving a nation's economic competitiveness and national security (Congressional Research Service [CRS], 2012; National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, 2007). With an increasing number of STEM institutions developed, the growth in STEM employment opportunities is faster than that in other fields, which leads to an urgent need to encourage broader participation in the STEM workforce (National Science Board [NSB], 2010; U.S. Department of Labor, 2007). Considered critical to national prosperity and power, STEM education has become the primary source of the STEM labor (CRS, 2012). In an effort to meet STEM workforce demand, the STEM education system in the United States requires considerable attention and investment (National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, 2007).

Among a series of stages in STEM education, doctoral education is an essential form of investment in human resources contributing to science, engineering, research, and scholarship in the society that pressingly needs scientific knowledge creation and technology innovation (National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, 2010). The number of science and engineering doctorate recipients has doubled over the last four decades, and nearly three-quarters of all the research doctorates awarded in 2011 were these in science and engineering fields

(National Science Foundation [NSF], 2012). However, there are still some challenges in the STEM graduate education. For example, women and underrepresented minority groups (Blacks, Hispanics, and American Indians/Alaska Native Americans) are still underrepresented in the STEM doctorate recipient population (NSF & Division of Science Resources Statistics, 2011). In addition, more than one quarter of the doctorates in science and engineering were awarded to temporary residents in 2009 (NSB, 2012). Consequently, it is important to further improve graduate education in the STEM related fields.

Among graduate education programs in the STEM related fields, this research study focuses on physician-scientists in the biomedical research field. Biomedical research is a broad area of science that provides a comprehensive understanding of how to prevent, diagnose, and treat disease (National Research Council [NRC], 2011a). Physician-scientists, including MD and MD/PhD degree holders who pursue research-based careers, are vital members of the biomedical research enterprise, since their scientific questions arise based on their experience of taking care of patients (Kaushansky, 2003; Rosenberg, 1999; Thier et al., 1980; Varki & Rosenberg, 2002). To be more precise, physician-scientists are defined as individuals with medical training who perform biomedical research as their primary professional activity (Ley & Rosenberg, 2005; Varki & Rosenberg, 2002). So far, the medical community has made an effort to improve the physician-scientist training (Ahn, Watt, Man, Greeley, & Shea, 2007; Rosenberg, 2002; Varki & Rosenberg, 2002).

There are three major sources of physician-scientists: those who pursue the MD/PhD dual degree upon matriculation to medical school; those who pursue MD and

PhD degrees separately through single-degree programs (who are not the focus of this study); and those who pursue the MD degree but later become engaged in extended research training, and are thus called "late bloomers" (Varki & Rosenberg, 2002). Most physician-scientists do not hold a PhD degree; however, the MD/PhD dual degree holders play an important role in biomedical research (Ahn et al., 2007; Clark & Hanel, 2001). MD/PhD degree holders accounted for less than 3% of total medical graduates, but more than 25% of NIH grant applications for clinical research between 1997 and 2002 (Kotchen, Lindquist, Malik, & Ehrenfeld, 2004). On the other hand, after all, the MD degree holders who primarily involve themselves in biomedical research account for the majority of the physician-scientists. The two groups of physician-scientists share many aspects in terms of research-based careers. For example, the two groups have similar success rate of obtaining grants from the NIH, and obtain similar proportions of awards among all applicants (Ley & Rosenberg, 2005). Therefore, this study includes both groups and compares the two groups in terms of their research interest change over time.

In the STEM related fields, student attrition has become a major concern in higher education. Researchers have discussed that high rates of attrition among STEM majors are an essential challenge for undergraduate STEM education in the United States (National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, 2007). Other researchers have suggested policy makers increase support for students' graduate study and post-doctoral research (CRS, 2012). After all, the number of grade school students who are interested in science is larger than the number of people who persist in the STEM pipeline, and is even much larger than the number of people who eventually become scientists discovering new knowledge (McGee & Keller, 2007).

Similarly, in the biomedical research field, the population of physician-scientists has become smaller and older (Cech et al., 2001; Guelich, Singer, Castro, & Rosenberg, 2002; Garrison & Deschamps, 2013; Ley & Rosenberg, 2005; Sung et al., 2003). Andriole, Whelan and Jeffe (2008) observed an attrition rate of 28.5% from the PhD portion of the MD/PhD program among all MD/PhD program enrollees nationally who graduated from medical schools between 2000 and 2006. In other words, compared to MD programs (which usually takes about 4 years to complete), academic pursuit in MD/PhD programs is lengthy (which usually takes 7-8 years on average to complete), which may lead to the challenging problem of attrition. With regard to this serious problem, the National Institutes of Health (NIH), and private foundations have taken the initiative by developing various awards and programs (Gallin, Le Blancq, & Clinical Research Fellowship Program Leaders, 2005; Nathan & Wilson, 2003; Neilson, 2003).

The attrition problem may also apply to the medical students who are primarily involved in research. Ley and Rosenberg (2005) found that although the number of patient care physicians increased over the last three decades, the number of research-focused physicians remained constant, even with a little decline over time. Results of a recent survey asking for nearly five hundred MD/PhD students' opinions about their educational experiences show that about a quarter of the participants reported that they had seriously considered leaving the MD/PhD program (Ahn et al., 2007). Meanwhile, the results might underestimate the actual percentage of students who once considered leaving the program since the students who already left the program were not included in the study (Ahn et al., 2007). Therefore, it is essential to retain research interest of medical

students so as to engage perspective physician-scientists to persist in the biomedical research field.

Gender is considered an important factor that is related to persistence in the STEM fields, including the biomedical research field. It has long been noted that women are disproportionately represented in the STEM education and workforce. Although the number of female doctorates awarded has increased faster than that of male counterparts, and female doctorates accounted for 42% of all the STEM doctorates awarded in 2011; females are still underrepresented in some particular STEM fields, such as physical sciences and engineering, accounting for less than 30% (NSF, 2012). What is more, the attrition rate of female students in the STEM related fields is even higher than male students. There is a similar situation in the biomedical research field (Andrews, 2002). Both fewer female applicants and more attrition among female students may explain the phenomenon that there are much fewer female physician-scientists than males (Andrews, 2002). It is imperative to increase the number of female students who are involved in research in medical schools and even more important to understand when to take the initiative to increase female students' research interest—before they apply for medical schools or while they are in medical schools.

In addition to the gender issue, underrepresented minority (URM) groups are another group of people that should be concerned in the STEM pipeline, including the pipeline of the biomedical research field. The URM groups mainly include Black, Hispanic, and American Indian/Alaska Native American people (NSF & Division of Science Resources Statistics, 2011), since these groups of people were disproportionately underrepresented in the STEM education system and workforce (NSF, 2012). The current

student-age population is more racially and ethnically diverse than previous generations in the United States (Cole & Espinoza, 2008; National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, 2011), which urges the nation's STEM enterprise to recruit more URM group labors. That is to say, it is important to broaden the participation of minority groups in STEM education and employment. In the medical field, as promoting diversity in physician workforce has been connected to improvement of health care equality, it is also an urgent and challenging task to recruit and retain URM students, especially those who are interested in doing biomedical research (Cooper, 2003; Fang, Moy, Colburn, & Hurley, 2000). In order to develop effective strategies to retain students in the biomedical research field, a key element is to have a good knowledge of when and how the URM students' research interest changes.

STEM research experience preparation is also important to students' future persistence in the STEM pipeline. Evidence shows that poor mathematics and science preparation is considered a great challenge in the STEM education system (National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, 2011). Research studies indicate that well-designed high school and college programs may help attract and maintain students in the STEM majors and also potentially maintain those individuals in the STEM career path in the future (Lopatto, 2007; Rohrbaugh & Corces, 2011). In the medicine field, students with undergraduate research experiences are more likely to anticipate an advanced degree (Winkleby, 2007). Although much evidence indicates the importance of previous research experiences to students' entering the medical school, little is known about whether such experiences can help sustain students' continued pursuit of research in their future study and career path.

Further, it is also essential to maintain the students who are currently studying in the STEM graduate programs. Research suggests that a good study and work environment may provide strong supports for STEM graduate students (Maton, 2004). Factors that contribute to students' success in science and engineering include academic and social integration, as well as monitoring and advising (Gardner, 2009). Programs with different foci may also influence students' academic pursuit in their respective majors. Biomedical research has historically relied mostly on physician-scientists that include both MD and MD/PhD degree holders who pursue research-based careers. Since MD students and MD/PhD students study and work in different programs, their research experience within the programs might be different, even though they both consider research-oriented careers in the biomedical field. As a result, it is worth examining whether different programs may train students in different ways so that they may have different levels of interest in doing research over time.

Purpose of the Study

The purpose of this study is to examine the trajectory of students' research interest in medical schools over time. To be more specific, this study first explores whether the research interest differs among medical students with different characteristics of gender, race/ethnicity, previous research experiences, or matriculated program at the beginning when they just register the MCAT exam. Then the study applies a longitudinal perspective to investigating whether medical students' levels of research interest change over time in general from when they register the MCAT exam to when they are matriculated in medical schools, and to when they graduate from medical schools. Finally, the study examines whether the patterns of change in research interest levels over time

vary among medical students with different characteristics of gender, race/ethnicity, previous research experiences, or matriculated program. The research questions to be addressed in this study are:

(1) Does medical students' reported research interest differ among students with different characteristics of gender, race/ethnicity, previous research experiences, or matriculated program prior to their entry to medical schools?

(2) Does medical students' reported research interest change in general across time from prior to their entry to medical schools, to when they are matriculated in medical schools, and to when they graduate from medical schools?

(3) Are patterns of change in medical students' reported research interest across time associated with gender, race/ethnicity, previous research experiences, or matriculated program?

Significance of the Study

First of all, this study offers evidence about the trajectory of medical students' research interest and projected research involvement in their careers over time through a national sample. Medical students' research interest in biomedical field is an important indicator of their future pursuit in biomedical research as physician-scientists and thus should be valued by policy makers, administrators, program heads and researchers. With a good understanding of the trend of medical students' research interest over time, corresponding implementations may be suggested to better serve medical students in terms of maintaining their research interest and experiences before, during, and after medical schools. Especially, this longitudinal study may have the potential to predict

students' later participation in postdoctoral research experience based on the trend of their research interest until graduation.

Besides, this study can provide important evidence concerning the research interest change over time among the underrepresented population (female and URM group students in medical schools). This study also compares underrepresented groups to the majority groups (male and White students) in terms of longitudinal research interest and projected research involvement in the careers. After analysis, the study has the ability to provide a comprehensive understanding of the current situation of research interest among underrepresented population in medical schools, and to suggest informative evidence about when to increase or maintain the research interest of underrepresented groups of medical students.

After comparing students with different levels of previous research activity participation, the study may indicate whether and how a variety of high school and undergraduate research oriented programs help increase and maintain medical students' research interest. More importantly, this study will present whether the previous research experiences have a long-term impact on students persistent interest in biomedical research. Accordingly, certain activities and programs are suggested to be developed and improved in order to recruit more physician-scientists in the biomedical research field.

In addition, this study also explores the difference in levels of research interest between MD program enrollees and MD/PhD program enrollees, which may suggest in what ways different programs should be improved in terms of training aspiring physician-scientists. Although the two programs both recruit students with focus on biomedical

research, they should be differentiated in terms of the curricula, strategies and policies to direct those students into the biomedical research workforce.

In summary, this study provides some analysis of how medical students' research interest changes over time and how such changes may vary across different groups of students. Educational researchers may further examine why medical students present such changes that vary across different groups. Policy makers and administrators can develop corresponding strategies to support medical students' program completion and, more importantly, to sustain their interest in biomedical research careers.

CHAPTER 2

REVIEW OF LITERATURE

Although attention has been paid to improving the biomedical research community, only a few studies discuss medical students' attitudes towards the importance of research to their academic pursuit and career path (Ahn, Watt, Greeley, & Bernstein, 2004; Ahn et al., 2007; Garrison & Deschamps, 2013; Newton & Grayson, 2003; Pheley, Lois, & Strobl, 2006; Watt, Greeley, Shea, & Ahn, 2005), and even fewer topics focus on whether medical students' research interest maintains over time (Guelich et al., 2002). However, many researchers have been exploring post-secondary students' experiences in the general STEM related fields, including the medical field, and examining whether those students can persist in their programs over time. As a result, this review of literature provides a comprehensive overview of the main factors that are associated with post-secondary students' persistence in the STEM related fields, especially in the medical field, and certainly the limited research concerning the factors that are related to medical school students' interest in research. To be more specific, researchers in previous studies have found that motivation (McGee & Keller, 2007; Sobral, 2004), demographic factors (Sax, 2001; Tsui, 2007), parental background (Bleeker & Jacobs, 2004), educational experiences (Lopatto, 2004; Russell, Hancock, & McCullogh, 2007), and institutional characteristics (Griffith, 2010) are related to students' persistence in the STEM related fields, including the medical field.

This review of literature starts with the relationship between students' motivation and their persistence in their academic and career pursuit. The focus of this study, research interest, can be considered as an important aspect of motivation which has a close connection with medical students' persistence in the biomedical research field (Lloyd, Phillips, & Aber, 2004). Then the review discusses the other main factors respectively that are associated with students' persistence in post-secondary education in the STEM related fields, including the medical field: demographic factors (i.e., gender and race/ethnicity), parental background, educational experiences, and institutional characteristics. Since the review contains extensive factors that are related students' persistence in the general STEM related fields, it should be noted that some of the topics discussed in the review may not be included in the analysis with the focus only on the biomedical research field. At last, the review discusses the social cognitive career theory (SCCT) which lays the foundation of the assumptions of this study.

Motivation

Motivation is an essential construct of affective dimensions of science learning (Simpson, Koballa, Jr., Oliver, & Crawley, 1994). According to Osborne, Simon, and Collins (2003), motivation can be categorized into two groups: intrinsic motivation and extrinsic motivation. Intrinsic motivation typically refers to an individual's inherent interest in particular activities, whereas extrinsic motivation is typically defined as the intention to engage in activities in order to obtain particular outcomes (Deci, Vallerand, Pelletier, & Ryan, 1991; Hidi & Harackiewicz, 2000). Previous research showed that intrinsic motivation was associated with academic performance and intended career path in the medical field (McManus, Linvingston, & Katona, 2006; Sax, 1994). For example,

motivation to participate in clinical research can positively help maintain clinical investigators in the medical field (Lloyd et al., 2004)

A longitudinal study conducted by Grandy (1998) investigated the factors that were related to the phenomenon that some high-ability minority students stayed while others left in science and engineering fields. Participants were solely 1620 minority students who were enrolled in college and completed the Postsecondary Experience Survey (PES). Results suggested that students' motivation in science was most significantly associated with students' commitment to science and engineering in college. The researcher also discussed that positive attitudes and enjoyment in science were more important for persistence than academic achievement.

In the medical field, motivation is also important to students' academic and career pursuits. Sobral (2004) investigated the patterns of 297 undergraduate students' motivation in a medical program, and their relationships with students' intentions of persistence in their studies. Results indicated that students with higher intrinsic academic motivation were significantly more likely to report a stronger intention to continue their studies. In another quantitative study, researchers examined the relationships between personality traits and intrinsic academic motivation among medical school students (Tanaka, Mizuno, Fukuda, Tajima, & Watanabe, 2009). Researchers conducted regression analyses on 119 students from a graduate school of medicine by including "Temperament and Character Inventory" and "Intrinsic Motivation Scale toward Learning" in the models. They found that dimensions of persistence, self-directedness were consistently and positively associated with students' intrinsic academic motivation, which was consistent with the findings discussed by Sobral (2004).

A qualitative study by McGee and Keller (2007) explored what compelled students to persist into PhD and MD/PhD training with the intent to do research and whether there were different motivations among different demographic groups of students. Researchers interviewed 26 college students from Summer Undergraduate Research Fellowship (SURF) and Initiative for Minority Student Development (IMSD) programs. Compared to other students who left the biomedical research pipeline, the themes of students who went on to PhD and MD/PhD training were various dimensions of motivation to do research: curiosity to discover the unknown, enjoyment of problem solving, and helping others indirectly through research.

Gender

In the past several decades, an increasing number of women have earned bachelor's, master's, and doctoral degrees in the STEM related fields, but women are still underrepresented among STEM degree recipients (NSF & Division of Science Resources Statistics, 2011; Sax, 2001), and account for less than a quarter of STEM labors (U.S. Department of Commerce, 2011). It seems that female students are still less likely to get enrolled in doctoral programs than their male peers (Mullen, Goyette, & Soares, 2003). In terms of the medical field, although there is a big increase in the fraction of MD/PhD program matriculants that females account for, the female student attrition rates during the programs are higher than those of males (Rosenberg, 2008). Besides, fewer female students participated in research than male students in medical schools (Lloyd et al., 2004).

Although many researchers have been examining issues such as why women are less interested in STEM than men and how women decide their undergraduate majors,

few have explored factors that are related to women's persistence in STEM after undergraduate education (Szelenyi & Inkelas, 2011). Students' early math and science preparation, study environment, as well as advice and encouragement from parents, peers and teachers are important to women's enrollment and persistence in STEM graduate education (Rayman & Brett, 1995; Sax, 2001; Szelenyi & Inkelas, 2011), while grades and self-esteem are surprisingly not (Rayman & Brett, 1995; Szelenyi & Inkelas, 2011). Besides, parental involvement was important to female students' decisions to study STEM related fields, but was not significantly associated with female students' persistence in the STEM pipeline (Featherman & Hauser, 1978; Moen, 1989). After a review of previous works about the difficulty for women to participate and persist in STEM professions, Wyer (2003) summarized four levels of barriers for these female students: systems barriers (i.e., biases in political and economic systems); institutional barriers (i.e., biases in the workplace, family and educational settings); interpersonal barriers (i.e., interaction experience with other people); and personal barriers (i.e., individuals' beliefs and values). However, some studies found that there was no significant gender difference in graduate degree aspirations in the STEM fields (Sax, 2001; Wyer, 2003).

Instead of comparing male and female students, some other studies focus on examining factors that were related to students' persistence in the STEM pipeline only among the female group. In a study conducted by Rayman and Brett (1995), the researchers explored factors associated with female students' persistence in science after college. Results based on 547 female survey respondents who graduated from a single-sex undergraduate institution indicated that greater likelihood to stay in science after

graduation was significantly related to more parental encouragement, and having received advisor's or other faculty members' career advice. Another study identified factors that positively or negatively affected women's progress towards the doctoral degree, and investigated whether these influences varied between women who finished their degrees relatively quickly and those who took longer time to finish their degrees (Maher, Ford, & Thompson, 2004). Female students' persistence was found to be associated with commitment to timely degree completion, working relationships with faculty, funding opportunities, family issues, and research experiences (Maher et al., 2004).

In the medical field, there are also several studies examining gender differences in the persistence in programs and career paths in medicine. An exploration of the relationship between gender and full-time faculty appointment indicated that women were more likely to have held full-time faculty appointments than men (Andriole & Jeffe, 2012). Another study conducted by Guelich et al. (2002) investigated medical students' research intention by gender. Analyses of nationwide data indicated that the research interest of both males and females' declined in medical schools, and females were significantly lower in research interest than males. There was a large and persistent gender gap in research intentions in medical school.

Race/Ethnicity

Race/ethnicity was another important focus in many research studies. Recent research has indicated that the undergraduate attrition rates are even higher especially among the underrepresented minorities (URM; including Black, Hispanic, and American Indian/Alaska Native American; Anderson & Kim, 2006; Berkes, 2008; Higher

Education Research Institute, 2010; Huang, Taddese, & Walter, 2000; Hurtado et al., 2007). According to Grandy (1998), it appears that the percentage of URM students who reported early interest in math and science is as high as the percentage of White students. However, the percentage of URM students who actually earned a bachelor's degree in the STEM related fields is much smaller than the percentage of White students. In addition, compared to the percentages of URM in the United States population, they account for a much smaller portion of the students who complete STEM doctoral degrees (Hoffer et al., 2006; Olson & Fagen, 2007). The completion of STEM degrees among minority students may be associated with family background, pre-college preparation, financial aid and employment, institutional characteristics, self-efficacy, and STEM support programs (Museus, Palmer, Davis, & Maramba, 2011).

Previous research indicated that students from the URMs tended to leave the STEM programs during undergraduate study, which led to fewer opportunities for them to pursue STEM doctoral degrees (Summers & Hrabowski, 2006). Accroding to Tsui (2007), three major barriers may explain URM students' low participation in the STEM related fields: social expectations, historical laws and regulations, and discriminatory policies and practices. When it comes to the specific field of medicine, there is also a lack of URM physicians in clinical and academic medicine (Cregler, 1993; Winkleby, 2007). Further, a study by Garrison, Mikesell, and Matthew (2007) indicated that the URM students spent longer time to complete the degree and had higher attrition rates in medical schools. As a result, in order to ensure a diverse STEM workforce, especially physician-scientist workforce, corresponding strategies should be developed to engage and maintain students to be interested in science, especially students from the URM groups.

A qualitative study examines the effects of Historically Black Colleges and Universities (HBCUs) on the pursuit of African American women for STEM careers (Perna et al., 2009). By using a case study on participants from an HBCU college, the researchers found that academic, psychological and financial barriers limited the persistence of Black women in the STEM related fields. In addition, institutional characteristics were also key factors for their persistence in STEM. Similar themes also emerged among medical students in another qualitative study conducted by Odom, Roberts, Johnson, and Cooper (2007). In addition to the barriers discussed above, facilitators experienced by URM medical students in their academic pursuits were scholarships, social support, professional exposure, and personal characteristics (Odom et al., 2007).

Jeffe, Yan, and Andriole (2012) investigated potential mediators for racial/ethnic differences among Asian/Pacific Islander, URM and White groups in full-time faculty appointments. Participants were 1994-2000 MD program matriculants who graduated before 2005. Mediation and logistic regression analyses indicated several significant mediators of racial/ethnic disparities in full-time faculty appointments: participation in postsecondary research activities, academic achievement, and faculty career intentions at graduation.

Parental Background

In the past years, the increase of graduate degrees outpaced the rate of growth in undergraduate degrees. However, it is obvious that more research focuses on undergraduate education than graduate education. Correspondingly, many studies indicate a strong relationship between parental background and students' decisions to

attend college (Hossler, Schmit, & Vesper, 1999), but some other studies suggested that such relationship did not apply in the enrollment in graduate programs (Mare, 1980; Stolzenberg, 1994). However, Ethington and Smart (1986) did find an indirect effect of parental background on students' graduate education decision making. As a result, it could be indicated that the influences of parental background, such as parental involvement, on students' persistence in graduate education are complicated according to previous literature.

It is worth noticing that parents of recent doctorate recipients are better educated than those of doctorate recipients in previous generations. In terms of ethnic groups, it appears that URM doctorate recipients are less likely to have at least a parent with bachelor's degree or above than their Asian and White peers, indicating to some extent that parental education may be another important factor that intertwines with the underrepresentation of minority students in doctorate education (NSF, 2012). Mullen et al. (2003) examined parents' education, academic achievement, and postgraduate educational enrollment. Based on the data from the Baccalaureate and Beyond Longitudinal Study (B&B), parents' education background had a strong influence on their children's entrance into a doctoral program, compared to other postgraduate programs.

A study conducted by George and Kaplan (1998) examined the joint influences of teachers and parents on students' attitudes towards science. Data analyses from the National Education Longitudinal Study of 1988 (NELS: 88) suggested that students' attitudes towards science were significantly associated with parental education, parental involvement, and participation in science activities. In a longitudinal study, Bleeker and

Jacobs (2004) examined the relationships between parents' perceptions of their children and the career plans of adolescents 12 years later. Results indicated that mothers' expectations of their children's success in math career were significantly associated with adolescents' math and science career self-efficacy, as well as adolescents' perceptions of math and science ability.

Educational Experiences

Another important effort should be put on development of numerous science focused programs involving hands-on research experience, since original research practice is considered a critical innovation element during the doctoral degree study (Carter, Mandell, & Maton, 2009). Research experience is able to better prepare students for graduate school and research careers, and to maintain those students to persist in their education and career paths (Hathaway, Nagda, & Gregerman, 2002; Russell et al., 2007; Schultz, Estrada-Hollenbeck, & Wood, 2008). Especially, it is important for medical students to enhance research background during undergraduate education (Rosenberg, 1999), since structured research opportunities may be positively related to long-term, continued research interest among medical school students (Hiatt & Sutton, 2000; Solomon, Tom, Pichert, Wasserman, & Powers, 2003). Therefore, one feasible approach to maintaining students to stay in the STEM pipeline is their involvement in original research projects (Lopatto, 2004). Students are encouraged to participate in various types of science related research-based activities or programs during high school and undergraduate study. Though quite a few studies focus on high school research activities, some researchers believed that high school research experiences could help prepare students for research opportunities in undergraduate study that would lead to students'

engagement in the STEM career path (Rohrbaugh & Corces, 2011; Zaikowski, Lichtman, & Quarless, 2007).

Rohrbaugh and Corces (2011) conducted a study to evaluate the Research Internship and Science Education (RISE) program in a university. The RISE program was dedicated to providing high school students, especially underrepresented students, with opportunities to participate in original research projects in the STEM related fields. Preliminary analyses suggested that the RISE program was effective in preparing high school students for majoring in science related fields in college and continued biomedical research involvement. Another study was conducted to assess Student Education Enrichment Programs (SEEP), which targeted for both high school and college students (Cregler, 1993). The researcher tracked participants in the biomedical science career pipeline, and found that early initiatives and role models could affect students' future decisions. Meanwhile, minority students who participated in enrichment programs or interventions were more successful in being accepted by professional schools. Finally, the researcher concluded that the enrichment programs or interventions could improve biomedical science career recruitment and meanwhile increase the number of minorities in the biomedical science pipeline.

Compared to high school research experience, there are much more undergraduate research opportunities in the STEM related fields. According to Wenzel (1997), undergraduate research is defined as "an inquiry or investigation conducted by an undergraduate in collaboration with a faculty mentor that makes an original intellectual or creative contribution to the discipline". This definition emphasizes the importance of both originality quality of research and student-mentor collaboration. With this definition in

mind, researchers summarized a series of benefits of undergraduate research, among which fostering professional growth and advancement was an important indicator of students' persistence in the discipline (Osborn & Karukstis, 2009). Undergraduate research can not only enhance students' research ability and student-professor connection, but also increase the probability of their enrollment in post-baccalaureate programs.

Carter et al. (2009) evaluated the effect of the Meyerhoff Scholarship Program (MSP) on STEM doctoral students' academic pursuits. In the study, they examined whether undergraduate research experiences were related to PhD degree pursuit outcome, and whether the relationships varied across different intended majors and different types of undergraduate research experiences. The primary goal of the MSP was to increase the number of students who obtain doctoral degrees and pursue a career in the STEM related fields. Participants of the study were 13 cohorts of 441 non-Hispanic students. Results suggested that participation in on-campus, academic year research was associated with higher probability of pursuing a STEM PhD or MD/PhD degree. The relationships varied across different types of undergraduate research experiences, but did not significantly vary across different intended majors. In another study, the effectiveness of the MSP was examined on 395 African-American participants (Maton, Domingo, Stolle-McAllister, Zimmerman, & Hrabowski, 2009). The analysis was conducted to compare doctoral degree against medical/master's/no post-college degrees. Logistic regression results showed that students who pursued STEM doctoral degrees reported significantly higher levels of pre-college research excitement than those in other programs (i.e., medical, master's, and no post-college degrees).

Winkleby (2007) investigated the outcome of a biomedical program, Stanford Medical Youth Science Program, which provided academic enrichment in medical sciences for diversity in the health professions. The study followed 405 program participants for over 18 years. Results indicated that the program was able to prepare low-income students for medical and other science related careers, and that programs in the biomedical pipeline could be successfully implemented.

Instead of concentrating on only one particular undergraduate research program or intervention, Lopatto (2007) attempted to explore students' perceptions of summer undergraduate research experiences in general. In the study, the researcher developed a measurement, named the Survey of Undergraduate Research Experience (SURE), to evaluate the undergraduate research experiences of 1135 undergraduates from 41 universities and colleges. Results suggested that most of the participants who reported prior research experiences had plans for science education beyond the undergraduate study, and the leading two categories were medical degree and doctoral degree in biology. Students with lower gains from undergraduate research experiences were less likely to continue their study after undergraduate education. According to Russell et al. (2007), undergraduate research opportunities (UROs) increase understanding, confidence and awareness; clarify interests in STEM careers; and increase the anticipation of a doctoral degree. Students with UROs because of their real interest in research were more likely to have positive outcomes, such as growing interest in STEM careers. As a conclusion, enthusiasm in research was the key element to fuel students' interest in advanced degree pursuit and career path in the STEM related fields.

Besides quantitative evidence showing the importance of previous research experiences to students' continuous pursuits in the science related fields, some qualitative studies also support such evidence. Previous qualitative research has summarized benefits of research experiences (Hunter, Laursen, & Seymour, 2006; Seymour, Hunter, Laursen, & Deantoni, 2004). The most frequent theme that undergraduate interviewees mentioned was that they gained confidence in the ability to do research and to make contributions to science, followed by increasing conceptual understanding of research, confirming education and career paths, and preparing for graduate schools and future careers (Seymour et al., 2004). The researchers later found that these themes were considered as personal and cognitive growth in students' perspectives, while faculty members treated these gains as a process of professional socialization into science (Hunter et al., 2006).

Institutional Characteristics

In addition to personal characteristics, parental background and educational experiences, previous research also discussed the importance of institutional structures, policies, and practices to students' attainment in the STEM related fields (Manson, 2009; Perna et al., 2009). The majority of doctorate recipients graduate from research universities classified as "very high research activity" institutions by Carnegie Foundation (NSF, 2012). Some researchers emphasize the importance of the characters of current programs to students' persistence in graduate education (DeValero, 2001; Griffith, 2010). Similarly in the medical field, it is essential to develop effective and attractive training programs so as to engage and maintain aspiring physician-scientists interested in research (Kaushansky, 2003; Lloyd et al., 2004; Rosenberg, 1999). Therefore, undergraduate students are suggested to become familiar with graduate study in advance.

From a qualitative perspective, students' relationships with their faculty advisors and their integration into departmental community are considered two most important reasons behind the graduate attrition (Lovitts & Nelson, 2000).

Previous research has focused on the importance of faculty mentors and role models to students' decision making about their majors, especially for women and racial/ethnic minority groups (Kaushansky, 2003; Golde, 2002; Lovitts & Nelson, 2000). Larger proportion of female and URM faculty members in a university or a college might encourage more students of the same gender or race/ethnicity to involve in the STEM related fields. The effects of faculty members as role models are mixed. Some researchers reported that female faculty members played an important role in female students' decisions to select a major (Ashworth & Evans, 2001; Carell, Page, & West, 2010; Rask & Bailey, 2002), while others did not find a significantly positive relationship between faculty role models and students' major choice (Canes & Rosen, 1995; Griffith, 2010).

Departmental climate is also considered important to students' persistence in respective disciplines. Higher percentage of female and minority graduate students in STEM fields was related to higher likelihood of persistence for women and minorities in STEM fields (Griffith, 2010). The institution type and students' educational experiences at institutions were associated with their likelihood of persisting in the STEM related major (Golde, 2002; Griffith, 2010). Additionally, the students' integration into departments' social and professional life is closely related to successful completion of a doctorate degree (Lovitts & Nelson, 2000).

DeValero (2001) examined departmental factors (policies, practices, advising, and climate) related to time-to-degree and completion rates of doctoral students in various programs. Participants were 876 students enrolled in doctoral programs in a top 25 doctorate-granting institution between 1986 and 1990. Through quantitative and qualitative analyses, the researcher found several factors that were associated with students' persistence in their doctoral program study: financial support, departmental orientation and advising, relationship between course work and research skills, requiring significant results in the dissertation, student-committee relationship, student-advisor relationship, student participation, and peer support.

Similar to the general STEM related fields, medical school research experience is also associated with later research involvement (Lloyd et al., 2004). The study conducted by Lloyd et al. (2004) evaluated factors that might influence individuals' current involvement in clinical research. The sample of the study was 428 individuals who graduated from a college of medicine between 1985 and 1995. Results indicated that funded investigators with more time doing clinical research were more likely to report that research carried out during medical school was an important influence on their current clinical research involvement. In addition, there was also a visible gender gap in medical schools: fewer females than males participated in clinical research.

Social Cognitive Career Theory

Based on Bandura's (1986) general social cognitive theory, Lent, Brown, and Hackett (1994) generated social cognitive career theory (SCCT), a theoretical framework which emphasizes how individuals determine career choices and the internal and external factors that impact on this process. In particular, the framework explains the processes of

how individuals develop career interests, enact career choices, and achieve performance outcomes. Meanwhile, Lent et al. illustrated three social cognitive essentials—"self-efficacy beliefs", "outcome expectations", and "goal representations" (p. 83), and how they interrelate with other internal and external factors. Self-efficacy is defined as "people's judgments of their capabilities to organize and execute courses of action required to attain designated types of performances" (Bandura, p. 391). Outcome expectations can be explained as an individual's perceptions about possible outcomes. According to Bandura, a goal may refer to the determination to participate in a specific activity.

In order to study the relationships among self-efficacy, outcome expectations, goal mechanisms and other factors, Lent et al. (1994) proposed three models, each with several propositions. In the model of career interest development, researchers maintain that individuals' career interests are influenced by their concurrent self-efficacy, expected outcomes, as well as their occupationally relevant abilities through interest. In the model of career choice, researchers mainly demonstrate that self-efficacy and outcome expectations both affect individuals' career choice goals and actions through interest. In the model of career performance, researchers highlight the connections among the three social cognitive essentials and task attainment level. On the one hand, self-efficacy beliefs, influenced by ability and past performances, may affect expected outcomes, career goals, and performances. On the other hand, outcome expectations, partially determined by self-efficacy and past performances, may influence career goals and performances as well. In these three models, person and contextual factors also play an important role.

Many researchers conduct quantitative studies to evaluate the models in the SCCT across different domains. Fouad, Smith and Zao (2002) designed a study to examine the fit of the SCCT model for typical academic subjects of the high school curriculum: mathematics and science, English, social studies, and art. Through structural equation modeling analyses, researchers concluded that the results for all the subjects analyzed were consistent with the SCCT model of the links between self-efficacy, outcome expectations, career interests and career goals. Another quantitative study also examined relationships among high school students' learning experiences, gender, self-efficacy, outcome expectations, career interests and aspirations (Tang, Pan, & Newmeyer, 2008). By using structural equation modeling to analyze the data, researchers verified Lent et al.'s (1994) model and concluded that individuals' self-efficacy beliefs, influenced significantly by learning experiences, were strongly related to career interests across genders.

The SCCT hypothesis that self-efficacy is the pivotal variable in individuals' career choices and development processes is discussed as theoretical framework in many studies. However, Armstrong and Vogel (2009) attempted to interpret the interest-efficacy association based on the Holland's (1959, 1997) theory—individuals could be classified in six RIASEC types: Realistic, Investigative, Artistic, Social Enterprising, and Conventional. Researchers analyzed the data collected from around 600 college students in a large Midwestern university through structural equation modeling, and found that the links between interest and self-efficacy were reciprocal and both predict students' career choice. Despite of some arguments on the relationship between self-efficacy and interest,

it appears that researchers agree on the model where individuals' interest can predict their career interest and career goals.

Although the SCCT is generated across disciplines, many research studies on the SCCT model have focused on the domain of in STEM subjects (Fried & MacCleave, 2009; Luzzo, Hasper, Albert, Bibby, & Martinelli, 1999; Nauta & Epperson, 2003). One important reason is that the SCCT model emphasizes the domain-specific nature (Lent et al., 1994). Bandura (1986) also argued that self-efficacy would be domain specific rather than universal in nature. A study conducted by Lent, Brown, and Gore (1997) shows that only the learning experiences in a specific career domain can influence the self-efficacy, interest and outcome expectations in that particular domain and ultimately shape career goals and choices in that domain.

The SCCT has been applied and extended in many research studies related to STEM choice and persistence in post-secondary education (Byars-Winston & Fouad, 2008; Herrera & Hurtado, 2011; Lent et al., 2003; Lent et al., 2005). The SCCT asserts that individuals' primary interest in a field is consistent with their career goals to enter that field (Lent et al, 1994). In this study, it is thus assumed that medical students with primary interest in research tend to hold a career goal in the biomedical research field; meanwhile, medical students' levels of research interest is assumed to be consistent with their expected amount of research involvement in their future careers in medicine. In that case, to explore the trend of medical students' research interest may be considered as to explore the trajectory of medical students' expectation of research involvement in their future careers.

Summary of Existing Research

This review discusses a series of previous studies related to the factors contributing to students' persistence in post-secondary STEM education, especially in the medical field, as well as the social cognitive career theory. Five main factors are examined in this review: motivation, demographic factors (i.e., gender and race/ethnicity), parental background, educational experiences, and institutional characteristics. It seems that whether students stay in the program or leave the program is determined by a set of factors that interact with each other through complicated processes.

Motivation is considered as a key factor of students' commitment in the biomedical research field (Lloyd et al., 2004). Especially, intrinsic motivation (i.e., inherent interest in particular activities) is positively associated with students' academic pursuit and career path in the medical field (McManus et al., 2006). Both quantitative and qualitative studies indicate that students with higher intrinsic academic motivation are more likely to be admitted to and persist in medical schools (McGee & Keller, 2007; Sobral, 2004).

Despite of an increasing number of female doctoral students in the STEM graduate programs, women are still underrepresented (CRS, 2012). Various barriers and challenges may contribute to lower level of persistence in the STEM pipeline among female graduate students (Wyer, 2003). Similarly, students from the URMs also present a disproportionately low completion rates for the post-baccalaureate degrees (Garrison et al., 2007; Higher Education Research Institute, 2010). Therefore, programs and interventions are suggested to target for women and URMs so as to recruit and maintain

more women and URMs in the STEM pipeline, especially in the biomedical field (Guelich et al., 2002; Jeffe et al., 2012; Perna et al., 2009; Szelenyi & Inkelas, 2011).

Parental background is also related to students' decision making in their academic pursuits and career aspirations. As parents of recent doctorate recipients are better educated than those in previous generations, it appears that parental background is positively associated with students' enrollment and persistence in the post-secondary programs (NSF, 2012). Parental background, including parental education and parental involvement, is found to have a significant impact on students' perceptions of learning science (George & Kaplan, 1998).

Another important aspect that may well prepare students for their future pursuits in the STEM related fields is previous educational background, which mainly refers to previous research related experiences that may provide students with authentic and original research participation experiences (Russell et al., 2007). Various high school and undergraduate research programs have been evaluated to have positive effects on students' continuous pursuits in academic and career paths (Carter et al., 2009; Rohrbaugh & Corces, 2011; Winkleby, 2007).

The last factor discussed in this review is the characteristics of the institutions that graduate students currently attend, including students' relationships with faculty advisors and their academic and social experiences in departmental community (Lovitts & Nelson, 2000). Faculty mentors and role models play an important role in students' decisions to choose majors in the STEM related fields, especially among female and URM students (Carell et al., 2010). Meanwhile, departmental environment is also essential to students' persistence in post-secondary STEM programs (Griffith, 2010).

Additionally, the social cognitive career theory (SCCT) is elaborated as a framework for understanding three models of career development: interest development, career selection, and performance and persistence in academic and occupational pursuits (Lent et al., 1994). So far, the SCCT has been applied in many disciplines, including STEM post-secondary education (Byars-Winston & Fouad, 2008; Herrera & Hurtado, 2011). According to the SCCT, this study assumes that medical students' levels of interest in research are consistent with their expected levels of research involvement in the future careers.

Limitations of Existing Research

The body of previous research studies has some major limitations. First, most studies focus on post-secondary students' experiences in the general STEM related fields, and factors of their persistence in the STEM pipeline. Few researchers narrowed the focus down to the medical field, not to mention the comparison between different programs within the medical school (Rosenberg, 2008). For example, many researchers have evaluated the impact of various undergraduate research programs on students' further academic pursuit in graduate degrees in the STEM related fields (Carter et al., 2009; Maton et al., 2009; Rohrbaugh & Corces, 2011). However, there is a paucity of research on the effect of undergraduate research programs on students' entry and persistence in the biomedical research field.

Second, many of previous studies explore factors that are associated with post-secondary students' persistence in STEM education. By persistence, most researchers mean staying in the program or completion of a program (Andriole et al., 2008; Maher et al., 2004). However, in the biomedical research field, it is even more important to

examine students' interest in doing research and whether such interest persists during the program. With decreasing interest in research, even if a medical student graduates, he or she may not become a physician-scientist which will still be a loss and a problem for the biomedical research workforce.

Third, although researchers apply quantitative analysis and qualitative analysis to investigating students' experiences in the STEM pipeline, the sample size for most studies is small. In many studies, participants were students from only one or several institutions (Ahn et al., 2007; Carter et al., 2009; Sobral, 2004). This study seeks to examine the medical students' persistence in their biomedical research interest based on nationwide large-scale data.

Most importantly, almost all researchers analyzed cross-sectional data in previous quantitative studies. Although there are some studies examining longitudinal data (Bleeker & Jacobs, 2004), the researchers focused on students' perspectives of the STEM related fields in general. No studies applied a longitudinal design to take a further look at the trend of medical students' research interest over time until graduation, which may potentially predict their levels of research involvement in later steps (such as postdoctoral training) along their career paths. It thus seems necessary and essential to investigate such topic in a longitudinal perspective (Ahn et al., 2007).

CHAPTER 3

METHODOLOGY

This study seeks to address the following research questions through a series of quantitative analyses including descriptive analyses and longitudinal data analyses.

(1) Does medical students' reported research interest differ among students with different characteristics of gender, race/ethnicity, previous research experiences, or matriculated program prior to their entry to medical schools?

(2) Does medical students' reported research interest change in general across time from prior to their entry to medical schools, to when they are matriculated in medical schools, and to when they graduate from medical schools?

(3) Are patterns of change in medical students' reported research interest across time associated with gender, race/ethnicity, previous research experiences, or matriculated program?

The data in this study come from the quantitative part of Project TrEMUR (Transitions in the Education of Minorities Underrepresented in Research), a mixed method research project that examines the transition experiences of aspiring physician-scientists. The quantitative component in Project TrEMUR consists of data from student record system and three questionnaires administered to perspective students (2001-2006), matriculated students (2001-2007), and graduated students (2005-2011) in medical schools in the United States, all provided by the Association of American Medical Colleges (AAMC). The description of Project TrEMUR including the AAMC data sets

below explains the mixed method project development, data collection, participant information, as well as the sample of this study. This chapter also presents variables of interest in this study, including dependent variables, independent variables, and control variables. Additionally, analytic approach, hypotheses, and missing data are also discussed.

Project TrEMUR

Project TrEMUR (Transitions in the Education of Minorities Underrepresented in Research) is a mixed method research study initiated by my advisor, Robert H. Tai, and has been funded by National Institutes of Health (NIH) since 2011. Project TrEMUR uses a mixed method research design to examine the transition points in the development of the physician-scientist workforce, especially the underrepresented demographic groups in the biomedical research field; and to explore the factors that are related to aspiring and current physician-scientists' persistence in the biomedical research-based career path. Project TrEMUR focuses on four transition points: (1) students who planned a biomedical research career and who considered an MD, MD/PhD, or PhD programs; (2) students who were matriculated into medical schools with biomedical research interest and who did or did not enroll in MD/PhD programs; (3) students matriculated in MD/PhD programs who did and did not complete the programs; and (4) MD and MD/PhD program graduates who are involved or not involved in research-based careers.

The data of the qualitative component of Project TrEMUR mainly consist of interviews conducted by Robert H. Tai research group. Participants were current and former MD, MD/PhD, and PhD students, as well as post-graduate professionals in the biomedical research field (i.e., current faculty members and researchers, and former

faculty members and researchers in biomedical research). The semi-structured interview protocols explored interviewees' educational experiences and career choices in the biomedical research field.

At the very beginning, the research group examined a comprehensive list of medical schools with MD, MD/PhD, or PhD programs throughout the United States, and then randomly selected medical schools in each geographic area. Then the research team solicited current and former doctoral students' consent to voluntary participation in the interview by asking deans of those medical schools to send an announcement, and by sending out posters and flyers to each of the medical schools selected. The research group also used snowball sampling to contact more individuals: after the interviews, participants were asked if they could provide the information of other people for potential participation in the study.

After solicitation and snowball sampling, the research group has so far conducted 217 face-to-face or phone interviews with the individuals who voluntarily participated in the project. Each interview lasted 45 minutes on average. All the interviews were transcribed and stripped of any identifying information. Only members of the research team are privy to the recordings, which are stored in secure files.

AAMC Data Sets

The quantitative component of Project TrEMUR consists of three questionnaire data sets and one student records system (SRS) data set which were provided by the Association of American Medical Colleges (AAMC). The three questionnaires are the Premedical Student Questionnaire (PMQ; the name was changed to the Pre-MCAT Questionnaire in 2008, and later to the Post-MCAT Questionnaire in 2013, but the

acronym did not change[3]), the Matriculating Student Questionnaire (MSQ), and the Graduation Questionnaire (GQ). Descriptions below introduce the respondents, data collection, and typical questions of the four data sets.

Every year, the PMQ is administered by the AAMC to the individuals who register the MCAT (Medical College Admission Test) exam a few weeks before or after their MCAT date. The PMQ respondents may not actually take the MCAT exam. Meanwhile, it should be noted that not all individuals who register the MCAT exam complete the survey. The PMQ data set only has information of those who voluntarily complete the survey. The response rate for the PMQ is about 40% for the calendar year of 2013 (AAMC, 2013a). In the PMQ, typical questions include respondents' career preference, parental education and profession, previous educational experiences, and factors of decision to study medicine. Project TrEMUR has the access to the PMQ data collected from 2001 to 2006.

The MSQ is administered by the AAMC to the students who are matriculated in medical schools in the first year. To be more specific, in terms of the timeline, MSQ is open between May and September in the year when the respondents are matriculated in medical schools. Same as the PMQ, the MSQ surveys are completed by individuals who voluntarily participate. The response rate is high, though, about 70% - 80% in the recent three years (AAMC, 2010; AAMC, 2011a; AAMC, 2012a). Typical questions in the MSQ include premedical experiences, career interest, educational debts and financing, as well as opinions about medicine-related statements. Project TrEMUR has the access to the MSQ data of the 2001-2006 PMQ respondents collected from 2001 to 2011.

[3] More information can be found in the AAMC website at https://www.aamc.org/data/pmq/faq/.

The GQ is another important online questionnaire administered by the AAMC annually. The questionnaire is for the students who are graduating in the academic year to complete. For example, students who are graduating in the academic year 2013-2014 (from July 1, 2013 to June 30, 2014) will be asked to voluntarily complete the 2014 GQ which will be open online from February to June in 2014. The response rate is also high, around 80% in the recent three years (AAMC, 2011b; AAMC, 2012b; AAMC, 2013b). Typical questions in the GQ include career preference, expected involvement in research in future careers, and debts and financing. Project TrEMUR has the access to the GQ data of the 2001-2006 PMQ respondents collected from 2005 to 2011.

The SRS data set captures the progress information from matriculation through graduation of the medical student population in the United States. The SRS data were collected by the American Medical College Application Service (AMCAS) and medical school registrars in the United States. The SRS includes information about medical students' demographics, programs that they are matriculated in and graduate from, as well as time of their matriculation and graduation, etc. The SRS data in this study are used to provide medical students' demographic information and actual years and programs of matriculation and graduation.

The four data sets discussed above are merged to one comprehensive data set with all the information in it so that the generated comprehensive data set can be treated as a longitudinal data set. In that case, one PMQ respondent, as one case, can maximally have information from all the four data sets. With the only one comprehensive data set, it is also clearer to observe individuals' information from the three time points (i.e., when they registered the MCAT exam, when they were matriculated in medical schools, and

when they graduated from medical schools). Quantitative analyses are conducted on this longitudinal data set.

In summary, the first three data sets, PMQ, MSQ, and GQ, are the surveys completed by individuals who considered medicine related careers, who were matriculated in medical schools, and who graduated from medical schools. These three survey data sets contain self-reported data reflecting students' reported opinions about each survey question. On the other hand, since the SRS data are collected by institutions, the data set only provides objective and factual information of medical students, such as gender, race/ethnicity, age, and matriculated program. In the explanation about variables below, more details about how these four data sets are applied for analysis are illustrated.

Participants

The sample of this study consisted of 39,839 medical school graduates who completed all the three questionnaires (PMQ, MSQ, and GQ). To be more specific, these individuals completed the PMQ when they registered the MCAT exam between 2001 and 2006, then completed the MSQ when they matriculated in medical schools between 2001 and 2007, and finally completed the GQ when they graduated from medical schools between 2005 and 2011.

This longitudinal data set provided a series of information about individuals who registered the MCAT exam between 2001 and 2006, such as their demographic information, whether they were matriculated into medical schools, whether they graduated from medical schools, and whether they took MSQ and GQ, etc. However, the data presented a wide admission year range (from 2001 to 2012) among these individuals. Therefore, it was unable to know whether these individuals actually applied for medical

schools, as well as in what year all the matriculants in medical schools (between 2001 and 2007) registered the MCAT exam and completed the PMQ (perhaps sometime beyond the 2001-2006 range) based on the data that our we have the access to. With all the existing information, it was impossible to locate the exact population to correspond the individuals in this longitudinal data set. As a result, weight was not created and used in this research study to extrapolate the population to the national medical school student population. That is to say, the target population of this study referred to the 2001-2006 PMQ respondents who entered medical schools between 2001 and 2007 and who at last graduated from medical schools between 2005 and 2011.

Additionally, within the target population discussed above, not all the students completed the MSQ or the GQ. It can be found in Table 3-1 that there are individuals who completed both, either, and neither. With a further look at Table 3-2, the gender and race/ethnicity composition within each subgroup is similar except for the last group (i.e., the "MSQ N; GQ N" group) where the percentages of White and Asian/Pacific Islander groups are a little different from those in other subgroups. It appears that White students were more willing to complete these questionnaires, while Asian/Pacific Islander students had relatively lower tendency to complete questionnaires. Since this research study followed individuals at three time points, the sample consisted of the 39,839 individuals with all the three questionnaires available (i.e., the "MSQ Y; GQ Y" group in Table 3-2). It can be observed that the White group was a little overrepresented, while the Asian/Pacific Islander group was a little underrepresented.

Dependent Variables

There are multiple dependent variables in a longitudinal study, since each time point has one corresponding dependent variable. As this study focused on medical students' research interest change over time, the dependent variable that was indicated across the three time points was students' reported attitudes towards importance of research in their medical academic pursuit and career path. Unfortunately, there was no exactly the same question that was asked across all the three questionnaires (PMQ, MSQ, and GQ). However, each questionnaire included one to two related variables suggesting students' reported research interest, which were thus considered as dependent variables.

In the PMQ when students just registered the MCAT exam, they were asked how important they considered research interest in their decision to study medicine (P_INTEREST_RSC; Figure 3-1). This variable was treated as a continuous variable with four values: "not important" is coded as 1, "slightly important" is coded as 2, "moderately important" is coded as 4, and "extremely important" is coded as 5. The reason why Option 3 is omitted will be discussed in the next paragraph.

In the MSQ when students were just matriculated in medical schools, they were asked a similar question (FAC_RESEARCH; Figure 3-2)—how important they consider research in their choice of medicine as a career goal. However, considered as a continuous variable, this question had five options: "not important" (coded as 1), "slightly important" (coded as 2), "somewhat important" (coded as 3), "moderately important" (coded as 4), and "extremely important" (coded as 5). It can be observed that the MSQ variable had one more option than the PMQ variable—"somewhat important" which was in the middle of the option range, and that all the other four options in the

MSQ variable were exactly the same as the options in the PMQ variable. Consequently, the four options of the PMQ variable were coded as 1, 2, 4, and 5 respectively as described previously, so that the coding system can be consistent across variables.

In the MSQ, there was another variable also indicating students' attitudes towards the importance of research in their professional pursuit (MSQ_CAREER_RESEARCH; Figure 3-3), which asked for the extent to which students expected to be involved in research during their medical career. This question also had five options: "not involved", "involved in a limited way", "somewhat involved", "significantly involved", and "exclusively", coded as 1 through 5 respectively. Since both variables in the MSQ suggested the importance of research to students' medical careers, principal component analysis (PCA) was conducted to examine whether there was one variable that could be generated to best represent these two variables (Stevens, 2002).

In the GQ when students graduated from medical schools, they were asked the same question (GQ_CAREER_RESEARCH; Figure 3-4) as discussed about the second related variable in the MSQ data set—expected research involvement during a medical career. The five options were also the same as in the MSQ question and thus were assigned with the same values.

Although all the four variables elaborated above did not contain exactly the same content, all of these variables had some indication of students' reported attitudes towards importance of research in their medical academic pursuit and career path in different stages. When beginning to consider studying medicine, individuals were asked the importance of research interest to their decision to study medicine. When actually entering medical schools, students were asked the importance of research to their pursuit

of medicine related careers, and the amount of research involvement they expect in their medical careers. When graduating from medical schools and starting medical careers, students were asked again the amount of research involvement they expect in their medical careers. In addition, according to the SCCT (social cognitive career theory) as discussed in the previous chapter, it was assumed that individuals' interest in research before and during medical school study is consistent with their career aspirations of being involved in research. Therefore, the variables described above were considered as the same latent measure of students' reported research interest in this study.

In addition, it should be noted that the dependent variables all contained five-point Likert-type scales. There have been discussions about optimal number of response categories in rating scales, and researchers have argued that five response categories already have the ability to present good enough reliability, validity, and respondent preferences (Preston & Colman, 2000; Weng, 2004). Further, the dependent variables should theoretically be considered as ordinal variables. Therefore, the dependent variable cannot be treated as continuous variables and used in the models directly. However, research has investigated that the maximum likelihood (ML) method can be applied when the categorical variables are normally districted and thus considered as continuous variables in the models (Rhemtulla, Brosseau-Liard, & Savalei, 2012). As a result, before statistical analysis, the normality assumption was checked on these dependent variables. If the assumption is retained, then the dependent variables can be considered as continuous variables.

Independent Variables

Based on the research questions of this study, the independent variables in the analyses were participants' gender, race/ethnicity, previous research experiences, and matriculated program. As discussed thoroughly in the previous chapter, a large body of literature links doctoral students' persistence in the academic pursuit to their demographic background, educational preparation, and current program characteristics, which provides a strong support to the rationale for examining how these variables are associated with medical students' long-term research interest.

Both gender and race/ethnicity variables were from the SRS data set. The gender variable was dummy-coded and named as "Female"; female participants had a value of 1, while male participants, as the reference group, had a value of 0. In terms of race/ethnicity, the SRS data sets provided detailed information about an array of specific races/ethnicities for each individual. Based on these different specific races/ethnicities, seven mutually exclusive racial/ethnic categories were generated: White, Black, Hispanic, Asian, Pacific Islander, American Indian/Alaska Native American, and multiple races (NRC, 2004). Due to the small number of Pacific Islanders and reference to previous research (Andriole et al., 2008; Jeffe et al., 2012), Asian and Pacific Islander groups were combined to one group. The White group was considered as the reference group. In addition, students with multiple races were not analyzed in this study due to the mixed pattern within this group. As a result, there were totally four dummy-coded race/ethnicity variables: Asian/Pacific Islander (named as "Asian/PI"), Black (named as "Black"), Hispanic (named as "Hispanic"), and American Indian/Alaska Native American (named

as “Native”). Within each group, if a participant belonged to that racial/ethnic group, then he/she had a value of 1 for the corresponding variable; otherwise, he/she had a value of 0.

Regarding the educational experiences, there were three variables in the PMQ data set that were related: HS_LAB, HS_PROG, and COLL_LAB. All of these three variables were dummy-coded with the value of 1 indicating respectively that individuals participated in high school summer laboratory research apprenticeship; high school classroom-based summer, after-school, or Saturday program; and college laboratory research apprenticeship; and with the value of 0 indicating respectively that individuals did not participate in the corresponding activities. The full descriptions of these variables were presented in Figures 3-5.

In addition, it is also important to examine the research interest trajectory difference between MD students and MD/PhD students, since these two groups are different in many aspects as discussed in the previous chapters, but are both essential sources of perspective physician-scientists. Based on the information provided by the SRS data set, the participants were matriculated into various programs: MD, BA/MD, MS/MD, MD/JD, MD/PhD, MD/Other, BS/MD, MA/MD, MD/MBA, MD/MPH, and MD/Dental (OMS). Since the focus of this study was the trajectory of research interest before, during and after medical schools among individuals who pursued an advanced degree, the matriculated program variable was converted to a dummy-coded variable named as “PROGRAM”. For this variable, the value of 1 indicated students that were matriculated in the MD/PhD program, and the value of 0 indicated students that were matriculated in the MD program. All the other programs that combine MD program with other programs were not included in the analysis, since there was a mixture among those

programs which might influence the analyses and results of this study in an unpredictable way.

According to the characteristic of longitudinal analysis design in regression models, time was also considered as an independent variable. During the analysis, the longitudinal data set was transformed from a wide format to a long format. In this study, a wide format referred to a data set where each case represented an individual that had three dependent measures of research interest; while a long format referred to a data set where each case represented a measure of research interest within an individual at one time point. That is to say, in a long-format longitudinal data set, there was only one variable suggesting individuals' research interest. Meanwhile, a new variable—time was created to indicate a temporal order of the repeated measures. In the long-format longitudinal data set, the time variable was assigned to the value of 0 if the case indicated individuals' research interest obtained from the PMQ; 1 if the case indicates individuals' research interest obtained from the MSQ; and 2 if the case indicates individuals' research interest obtained from the GQ.

Control Variables

Although there were multiple independent variables of interest in this study, only one independent variable was the focus in the model at one time. In that case, all the other independent variables discussed above were treated as control variables to that particular independent variable with focus.

In addition, parental education and profession were included in the analyses as control variables. As reviewed previously, a body of literature has found a noticeable impact of parents on students' decision making in their academic and career pursuits

(Bleeker & Jacobs, 2004; Mullen et al., 2003). In the surveys, two aspects of parental background were considered: parental education and parental profession. Regarding parental education, questions in both PMQ and MSQ were asked about participants' father's and mother's highest educational levels respectively. In this study, a dummy variable was generated based on this information. Participants who had at least one parent holding bachelor's degree or above were coded as 1, otherwise coded as 0. Concerning parental profession, questions in both the PMQ and the MSQ asked about participants' father's and mother's occupations respectively. Another dummy variable was generated to group parental occupations: profession in health-related fields and profession in other fields.

Another control variable included in this study was students' age, which may also indicate their educational level to some extent. According to previous literature, age plays an important role in individuals' participation in science and persistence in post-secondary programs (Alexander, Johnson, & Kelley, 2012; Fisher & Engemann, 2009). In the PMQ, the individuals who registered the MCAT exam were asked about their age at that year, which was considered as a continuous variable included in the analytic model.

Analytic Approach

In order to address the research questions proposed in this study, the analytic approach included descriptive analyses and general linear regression models. Descriptive analyses were conducted to provide basic information about the variables discussed above, including demographics, parental background, previous research experiences, and

matriculated program. Below are the descriptions of development of general linear regression models to answer the research questions.

The research questions proposed in this study seek to compare the research interest among different groups of students, to examine the change of medical students' research interest over time, and to investigate whether the patterns of change in research interest over time vary across different groups of students. One of the most suitable approaches to conduct longitudinal data analysis is general linear regression model, since the regression paradigm is considered as a very flexible and versatile approach for analyzing longitudinal data (Fitzmaurice, Laird, & Ware, 2004). Regression models can provide a parsimonious and explicit description and explanation of two important aspects of information: first, how the mean response, in this case research interest, changes with time; second, how such changes are related to covariates of interest, in this case gender, race/ethnicity, previous research experiences, and matriculated program (Fitzmaurice et al., 2004). In each model, only one independent variable is focused and examined. Within each model, there are two general steps to conduct longitudinal data analysis: covariance modeling and mean modeling.

Before modeling the mean response (in this case research interest) over time, a best covariance structure is selected to represent the covariance among repeated measures obtained on the same individuals. An important feature of longitudinal data is that they are correlated, which should be captured by certain covariance structures. With an appropriate model for the covariance, the study can make valid inferences about the regression parameters. Two broad approaches to modeling the covariance are discussed and examined in this study. The first approach is to allow the covariance among repeated

measures to be in any arbitrary pattern, which is referred to as an "unstructured" covariance. The advantage of this approach is that there is no assumption made about the variances and covariances. However, if the correlation among repeated measures follows distinctive patterns, then the covariance structure may be built based on a covariance pattern mode. Covariance pattern models for longitudinal data examined in this study include: compound symmetry, Toeplitz, and autoregressive (Fitzmaurice et al., 2004). After all the covariance structures are modeled, the likelihood ratio test, Akaike information criterion (AIC), and Bayesian information criterion (BIC) are used to select a model that can provide the most adequate fit to the covariance in the data (Fitzmaurice et al., 2004).

After the covariance structure is modeled appropriately, the regression models for the mean response (in this case research interest) are selected and then interpreted to address the three research questions. The three alternative models to be tested and compared are: (1) linear trends, (2) quadratic trends, and (3) linear splines (Fitzmaurice et al., 2004). In the formulas illustrated below, the control variables are not explicitly listed because the formulas focus on the particular independent variable of interest. However, in the actual analyses, the control variables are all included.

(1) First, it is assumed that the curve for changes in the mean response (in this case research interest) over time is a straight line. Then,

$$E(Y_{ij}) = \beta_1 + \beta_2 \times Time_{ij} + \beta_3 \times Group_i + \beta_4 \times Time_{ij} \times Group_i \qquad (3.\ 1)$$

where

Y_{ij} denotes the response variable (in this case research interest) for the i^{th} individual at the j^{th} occasion (in this case PMQ, MSQ, or GQ);

$E(Y_{ij})$ denotes the expectation of Y_{ij};

$Time_{ij}$ denotes the value of the "time" variable (discussed above as an independent variable) for the j^{th} occasion on the i^{th} individual; and

$Group_i$ denotes the group membership of the i^{th} individual in terms of the particular independent variable. That is to say, $Group_i = 1$ if the i^{th} individual has a value of 1 in the particular independent variable, and $Group_i = 0$ otherwise. For example, if the model is to examine the association between medical students' gender and their long-term research interest, then the independent variable in this model is gender. In that case, female students ($Group_i = 1$) are the group with a value of 1, and male students ($Group_i = 0$, i.e., the reference group) are the group with a value of 0. Since all the independent variables in this study are dummy-coded, this model and all the models below apply to all the independent variables to be examined in this study.

Therefore, the model for the mean research interest for the individuals in the group with a value of 0 (i.e., reference group) is

$$E(Y_{ij}) = \beta_1 + \beta_2 \times Time_{ij}$$

The model for the individuals in the group with a value of 1 is

$$E(Y_{ij}) = (\beta_1+\beta_3) + (\beta_2 + \beta_4) \times Time_{ij}$$

To answer the first research question, β_3 is tested. The null hypothesis is that the medical students' research interest at the beginning (when completing the PMQ) is not different between the two groups; $\beta_3 = 0$. To answer the second research question, β_2 is tested. The null hypothesis is that the reference group students' research interest does not change over time; $\beta_2 = 0$. To answer the third research question, β_4 is tested. The null

hypothesis is that the changes in medical students' research interest over time do not differ between the two groups; $\beta_4 = 0$.

(2) Second, it is assumed that the changes in the mean response over time can be approximated by quadratic trends. Then,

$$E(Y_{ij}) = \beta_1 + \beta_2 \times Time_{ij} + \beta_3 \times {Time_{ij}}^2 + \beta_4 \times Group_i + \beta_5 \times Time_{ij} \times Group_i + \beta_6 \times {Time_{ij}}^2 \times Group_i \quad (3.\ 2)$$

Therefore, the model for the mean research interest for the individuals in the group with a value of 0 (i.e., reference group) is

$$E(Y_{ij}) = \beta_1 + \beta_2 \times Time_{ij} + \beta_3 \times {Time_{ij}}^2$$

The model for the individuals in the group with a value of 1 is

$$E(Y_{ij}) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) \times Time_{ij} + (\beta_3 + \beta_6) \times {Time_{ij}}^2$$

To answer the first research question, β_4 is tested. The null hypothesis is that the medical students' research interest at the beginning (when completing the PMQ) is not different between the two groups; $\beta_4 = 0$. To answer the second research question, β_2 and β_3 are tested. The null hypothesis is that the reference group students' research interest does not change over time; $\beta_2 = \beta_3 = 0$. In terms of the third research question, the rates of change between two groups are different and both depend on $Time_{ij}$. The rate of change in the reference group (Group = 0) is given by $\beta_2 + 2\beta_3 Time_{ij}$; while the rate of change in the focused group (Group = 1) is given by $(\beta_2 + \beta_5) + 2(\beta_3 + \beta_6)Time_{ij}$.

(3) Third, it is assumed that the changes in the mean response over time may follow a piecewise linear pattern, which is referred to as a linear spline model. Since

there are three time points in this study, the only possible time knot is at the middle time point—when students complete the MSQ (i.e., $Time_{ij} = 1$).

$$E(Y_{ij}) = \beta_1 + \beta_2 \times Time_{ij} + \beta_3 \times (Time_{ij} - t^*)_+ + \beta_4 \times Group_i + \beta_5 \times Time_{ij} \times Group_i + \beta_6 \times (Time_{ij} - t^*)_+ \times Group_i \quad (3.3)$$

where t^* is the time knot (which is 1 in this case), and $(Time_{ij} - t^*)_+ = \max(0, Time_{ij} - t^*)$.

Therefore, the model for the mean research interest for the individuals in the group with a value of 0 (i.e., reference group) is

$$E(Y_{ij}) = \beta_1 + \beta_2 \times Time_{ij} + \beta_3 \times (Time_{ij} - t^*)_+$$

The model for the individuals in the group with a value of 1 is

$$E(Y_{ij}) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) \times Time_{ij} + (\beta_3 + \beta_6) \times (Time_{ij} - t^*)_+$$

To answer the first research question, β_4 is tested. The null hypothesis is that the medical students' research interest at the beginning (when completing the PMQ) is not different between the two groups; $\beta_4 = 0$. To answer the second research question, it is complicated to examine whether the reference group students' research interest does not change over time. Interpretations of the results should be based on the change rate before and after the time knot respectively. As a result, β_2 *and* β_3 are tested respectively. The null hypothesis is that the reference group students' research interest does not change over time; $\beta_2 = \beta_3 = 0$. To answer the third research question, β_5 *and* β_6 are tested. The null hypothesis is that the changes in medical students' research interest over time do not differ between the two groups; $\beta_5 = \beta_6 = 0$.

After the three proposed models are evaluated, likelihood ratio test is conducted to compare the three models, and to select the model that best fit the longitudinal data in this study (Fitzmaurice et al., 2004). With the most appropriate covariance and mean model, the three research questions are addressed based on the interpretations of the estimated parameters from the regression models. All the longitudinal data analyses are completed with the use of SAS 9.3. Meanwhile, considering the large sample size in the analyses, effect sizes are calculated for each of the three research questions to provide important information from a practical perspective (Feingold, 2009).

Missing Data

Before all the inferential analyses, missing data were examined. Again, the sample of this study consists of 39,839 individuals who completed all the three questionnaires. However, there are cases where some questions were answered by those individuals while some were not. As a result, it is necessary and important to investigate whether the individuals did not answer certain questions intentionally. Correlation analyses were conducted to see whether missing data in the research interest variables were related to their group membership. Meanwhile, before completing the survey, individuals were notified that the questionnaires were not used for any evaluation. As a result, it was assumed that whether they answer the questions about their research interest was not related to their current particular research interest. If the data in this study were considered as missing completely at random (MCAR), and the analytic approach proposed above was able to yield valid inferences when missing data are MCAR (Fitzmaurice et al., 2004).

Table 3-1

MSQ and GQ Completion Status of 2001-2006 PMQ Respondents Who Entered and Graduated from Medical Schools (Total N=77,541)

		MSQ	
		Y	N
GQ	Y	39,839	14,164
	N	13,947	9,591

Note. Y = completed, N = not completed.

Table 3-2

Gender and Race/Ethnicity Composition within Each Subgroup

	PMQ ALL n=77,541		MSQ Y; GQ Y n=39,839		MSQ Y; GQ N n=13,947		MSQ N; GQ Y n=14,164		MSQ N; GQ N n=9,591	
	n	%	n	%	n	%	n	%	n	%
Gender										
Female	38,142	49.2	20,464	51.4	6,676	47.9	6,713	47.4	4,289	44.7
Male	39,399	50.8	19,375	48.6	7,271	52.1	7,451	52.6	5,302	55.3
Race/Ethnicity										
White	49,779	64.2	26,967	67.7	8,538	61.2	8,872	62.6	5,402	56.3
Black	4,957	6.4	2,246	5.6	992	7.1	945	6.7	774	8.1
Hispanic	5,747	7.4	2,632	6.6	1,142	8.2	1,148	8.1	825	8.6
Asian/ Pacific Islander	14,453	18.7	6,688	16.8	2,763	19.8	2,768	19.6	2,234	23.3
American Indian/ Alaska Native	245	0.3	126	0.3	56	0.4	33	0.2	30	0.3
American Multiple Races	2,043	2.6	1,011	2.5	412	3.0	342	2.4	278	2.9
No Response	317	0.4	169	0.4	44	0.3	56	0.4	48	0.5

Note. Y = completed, N = not completed.

Figure 3-1. Question from the PMQ on Interest in Research

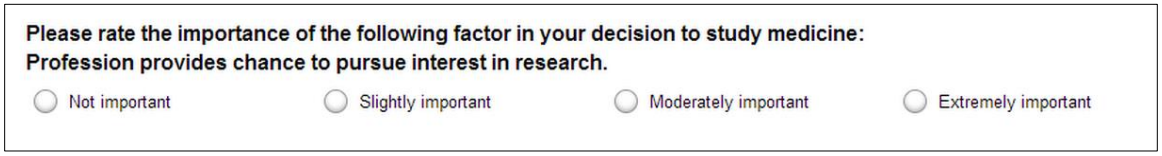


Figure 3-2. Question from the MSQ on Interest in Research

Figure 3-3. Question from the MSQ on Involvement in Research

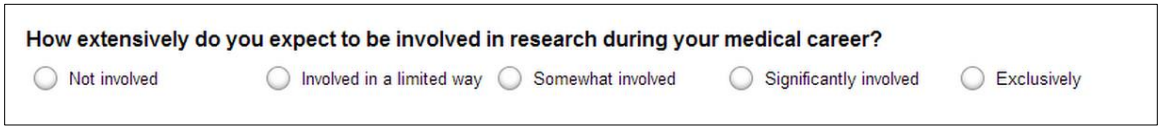


Figure 3-4. Question from the GQ on Involvement in Research

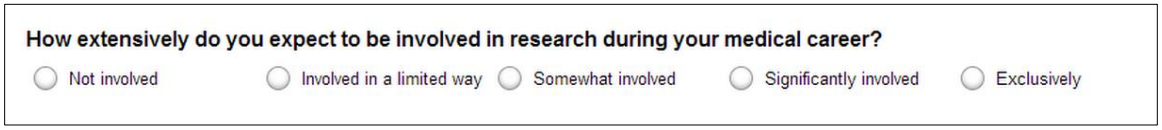


Figure 3-5. Question from the PMQ on Research Related Experiences

CHAPTER 4

RESULTS AND DISCUSSION

The results and discussion in this study are based on two main components of analysis: descriptive analysis and general linear regression model analysis. Descriptive analysis provides a description of the study sample, demographics, previous research experiences, matriculated program, as well as reported research interest levels over time. General linear regression model analysis provides evidence about the trajectory of medical students' reported research interest across time from when they registered the MCAT, to when they were matriculated in medical schools, and to when they graduated from medical schools; and whether patterns of change were significantly associated with students' characteristics of gender, race/ethnicity, previous research experiences, or matriculated program.

Descriptive Analysis

This section provides a review of the sample in this study and a series of descriptive analyses for all the related variables: demographics (including gender, race/ethnicity, age, as well as parental education and profession), previous research experiences, matriculated program, and reported research interest levels over time across different groups. The descriptive analysis results present an initial overview of the variables examined in this study, including distributions and descriptive statistics. It should be noted that the descriptive analysis is not for any inferential conclusions, which is reserved for the statistical analysis to be discussed later in this chapter.

Sample

The sample analyzed in this study was composed of the data from 39,839 medical school students who completed the PMQ between 2001 and 2006 when they registered the MCAT, then completed the MSQ between 2001 and 2007 when they were matriculated in medical schools, and finally completed the GQ between 2005 and 2011 when they graduated from medical schools. In the data obtained in Project TrEMUR, totally 262,113 individuals completed the PMQ (Premedical Student Questionnaire) between 2001 and 2006. In this population, 77,541 individuals were matriculated in medical schools between 2001 and 2007, and graduated from medical schools between 2005 and 2011. Other PMQ respondents either were not matriculated in medical schools between 2001 and 2007, or did not graduate from medical schools between 2005 and 2011. In order to appropriately and effectively address the research questions proposed in this longitudinal study analysis, only the students among the 77,541 individuals who completed both the MSQ (Matriculating Student Questionnaire) at matriculation and the GQ (Graduation Questionnaire) at graduation were included in the analysis. As a result, the sample size for this study was 39,839.

Demographics

Gender distribution is displayed in Table 4-1. Among the 39,839 medical students analyzed in this study, 51.4% (n = 20,464) were females while 48.6% (n = 19,375) were males. Female and male students were almost evenly distributed, with females accounting for only a little larger proportion of the sample. With regard to the race/ethnicity distribution presented in Table 4-2, the majority group, White students accounted for 67.7% (n = 26,967) of the sample. Among the other racial/ethnic groups,

Asians and Pacific Islanders were the largest, accounting for 16.8% (*n* = 6,688) of the sample. A relatively equal number of Black (5.6% or *n* = 2,246) and Hispanic (6.6% or *n* = 2,632) students were found in the sample. In addition, 126 students (0.3%) of the sample were American Indian and Alaska Native American; 1,011 students (2.5%) were identified with more than one race/ethnicity. There were missing data for 169 students (0.4%) in the sample.

Figure 4-1 and Table 4-3 display the distribution and descriptive statistics of the age at the MCAT test for the study sample. The age range was between 13 and 51. The mean age at test of the study sample was 21.56, with a standard deviation of 2.65. Additionally, it can be indicated that the majority of the students took the MCAT test around their college time, especially when they were 20 and 21 years old.

The sample students' parental background is also important demographic information. With regard to parental education level (Table 4-4), 32,646 (84.5%) students reported that at least one of their parents had college education or above; while 5,993 (15.5%) reported that neither of their parents had college education or above. For parental profession (Table 4-5), a little more than a third (36.9% or *n* = 13,981) of the sample students indicated that at least one of their parents worked in health related fields, while others (63.1% or *n* = 23,937) indicated that neither of their parents worked in health related fields.

Previous Research Experiences

As introduced in the previous chapter, three variables were focused in this study regarding students' previous research related experiences. First, students were asked in the PMQ whether they had participated in summer laboratory research apprenticeship in

high school. According to descriptive analysis results shown in Table 4-6, almost ten percent (9.2% or *n* = 3,666) of the sample students reported that they had such experience, while the others (90.8% or *n* = 36,173) did not. Second, students were asked in the same PMQ question whether they had participated in classroom-based summer, after-school, or Saturday program in high school (Table 4-7). Among the 39,839 students analyzed in this study, 8.9% (n = 3,536) reported that they had such experience, while the others (91.1% or *n* = 36,303) did not. Third, students were also asked at the same time whether they had participated in laboratory research apprenticeship in college. Through descriptive analysis (Table 4-8), a little more than a third (36.6% or *n* = 14,596) of the students reported that they had such experience, while others (63.4% or *n* = 25,243) did not. In this study sample, it seems that the number of students who had research related experience in college was larger than the number of students who had research related experience in high school.

Matriculated Program

As explained in the previous chapter, this study focused on the students at the graduate or professional level in the medical field. As a result, only the students who were matriculated into the MD/PhD program and those who were matriculated into the MD-only program were compared in this study. According to Table 4-9, among the 39,839 students analyzed in this study, the majority of the students were matriculated in the MD-only program (97.1% or *n* = 38,684); while 377 students (1.0%) were matriculated in the MD/PhD program. In addition, 778 (1.9%) students were matriculated in other programs, including BA/MD, MS/MD, MD/JD, MD/Other, BS/MD, MD/MBA, MD/MPH, and MD/Dental (OMS).

Research Interest

Before a descriptive analysis for the research interest levels over time, principal component analysis (PCA) was conducted to examine whether the two variables derived from the two questions in the MSQ discussed in the previous chapter (FAC_RESEARCH and MSQ_CAREER_RESEARCH) could be represented by one variable indicating students' reported research interest levels at the matriculation time point. The PCA result (Table 4-10) showed that the first component accounted for 85% of the variance of the two variables, while the other component accounted for only 15% of the variance. The result indicated that the two variables could be represented by one variable without losing much variance or information. In terms of conceptual meaning of this PCA result, students' reported levels of research interest were consistent with their expected levels of research involvement in their future medical careers, which verified the assumption based on the SCCT (social cognitive career theory). A new variable was then generated by averaging the two variables based on the two MSQ questions to indicate students' reported research interest levels when they were matriculated in medical schools. So far, there was only one variable that indicated students' reported research interest levels at each time point.

Descriptive analyses were performed on the research interest levels of the sample students across time from prior to their entry to medical schools, to when they were matriculated in medical schools, and to when they graduated from medical schools. As described earlier, the research interest level variables were based on the questions asking students how important they thought research interest or involvement was to their academic and career pursuits. The highest score was 5 representing "extremely

important", and the lowest was 1 representing "not important at all". Overall, the means and standard deviations for the research interest levels at the three time points (i.e., when students registered the MCAT, when students were matriculated in medical schools, and when students graduated from medical schools) were found to be 3.05 (*SD*[4] = 1.47), 2.65 (*SD* = 0.90), and 2.66 (*SD* = 0.83) respectively and shown in Table 4-11. Meanwhile, Table 4-11 also shows that the skewness and kurtosis statistics for the research interest levels at the three time points were all acceptable and did not violate the assumption of normality. Therefore, students' reported research interest levels at the three time points were considered as continuous variables in the statistical analyses to be illustrated later in this chapter. Through a cursory inspection of Figure 4-2 without considering any factors (e.g., gender, race/ethnicity, etc.), it appears that students' research interest levels decreased from when they registered the MCAT exam to when were matriculated in medical schools, but remained almost the same after they entered medical schools until they graduated from medical schools. Moreover, the research interest levels reported by the sample students were generally low, which indicated that the students did not consider research as a very important factor in their academic and career pursuits on average. As a repeated note, valid inferences are made later in this chapter. What follows are further descriptive analyses on the research interest levels across different subgroups of the sample medical school students (i.e., groups by gender, race/ethnicity, previous research experiences, and matriculated program).

When the reported research interest levels were separated by gender (Table 4-12), female students reported the mean scores of 3.04 (*SD* = 1.48), 2.61 (*SD* = 0.89), and 2.62 (*SD* = 0.81) respectively at the three time points; while male students reported 3.06 (*SD* =

[4] *SD* denotes standard deviation.

1.47), 2.69 (*SD* = 0.90), and 2.71 (*SD* = 0.85) respectively. Figure 4-3 displays the mean reported research interest levels for female and male students over the three time points. It appears that in general, male students reported relatively a little higher research interest level than female students. Inferential analysis and results are presented later in this chapter.

The reported research interest levels separated by race/ethnicity are presented in Table 4-13 and Figure 4-4. Based on the descriptive statistics, Asian/Pacific Islander, Black, and Hispanic students on average reported to consider research as a more important factor in their academic and career pursuits than their White peers; while American Indian and Alaska Native American students reported lower scores for the research interest than the White students. Again, statistical analysis and inferences are discussed later in this chapter.

With regard to previous research experiences, the mean reported research interest levels were compared three times based on three different previous research related activities. The first activity examined in this study was high school summer laboratory research apprenticeship (Table 4-14 and Figure 4-5). Mean research interest levels of students who reported having participated in high school summer laboratory research apprenticeship were 3.46 (*SD* = 1.43), 2.95 (*SD* = 0.90), and 2.88 (*SD* = 0.81) respectively at the three time points. Mean research interest levels of students who did not have such experience were 3.00 (*SD* = 1.47), 2.62 (*SD* = 0.89), and 2.64 (*SD* = 0.83) respectively. Second, Table 4-15 and Figure 4-6 show the comparison between the mean research interest levels over time of students who participated in classroom-based summer, after-school, or Saturday programs in high school and those of students who did

not participated in such programs. Descriptive statistics indicated that the mean research interest levels of students who participated were 3.16 (*SD* = 1.46), 2.69 (*SD* = 0.90), and 2.76 (*SD* = 0.83) respectively; while the mean research interest levels of students who did not participate in such programs were 3.04 (*SD* = 1.48), 2.65 (*SD* = 0.90), and 2.65 (*SD* = 0.83) respectively. Third, with regard to the college laboratory research apprenticeship experience (Table 4-16 and Figure 4-7), the mean research interest levels of students who participated in laboratory research apprenticeship during college were 3.41 (*SD* = 1.44), 2.87 (*SD* = 0.90), and 2.79 (*SD* = 0.81) respectively; while the mean research interest levels of students who did not were 2.79 (*SD* = 1.45), 2.52 (*SD* = 0.87), and 2.58 (*SD* = 0.83) respectively. In summary, it seems that students who participated in high school and college laboratory research apprenticeship, as well as high school classroom-based summer/after-school/Saturday program reported higher research interest than those who did not over time in general, though such difference varied across different comparisons, which is described in more details in the longitudinal data analysis results.

In addition, when separated by matriculated program (Table 4-17 and Figure 4-8), the mean reported research interest levels of students from the MD/PhD programs were 4.66 (*SD* = 0.77), 4.33 (*SD* = 0.43), and 3.67 (*SD* = 0.67) respectively at the three time points; while the mean research interest levels of students from the MD-only programs were 3.04 (*SD* = 1.47), 2.63 (*SD* = 0.88), and 2.65 (*SD* = 0.83) respectively. The descriptive analysis results indicated that the students from the MD/PhD programs had much higher level of research interest than the students from the MD-only programs. This observation is conceptually understandable, because the MD/PhD program typically trains perspective physician scientists who pursue biomedical research as their career

focus in the medical field. On the other hand, not all the students from the MD-only programs were interested in the research area of the medical field. Again, details about the statistical analysis of the comparison between those two groups of students will be illustrated later in this chapter.

General Linear Regression Model

Longitudinal data analyses through general linear regression model were conducted to address the three research questions proposed in this study by providing inferential evidence. This section contains the results and discussions based on a series of statistical analyses. First, missing data were evaluated before all the further statistical analyses were conducted. Second, appropriate covariance structures and mean models were selected for each model with only one focus independent variable as discussed in the previous chapter. With the selected covariance structures and mean models applied, general linear regression models were conducted for each focus independent variable and the results were discussed to address the research questions.

Missing Data

Missing data analysis was conducted to examine the missing data mechanism of the data set used in this study. Based on the results shown in Table 4-18, the probability that students' reported research interest scores at the three time points were missing was unrelated to the set of all the variables included in the models. All the correlations were equal to or smaller than 0.12, except that the correlation between the probability of missing data in the PMQ research interest variable and whether students participated in college laboratory research apprenticeship was 0.27, which could still be considered as small. In addition, due to the characteristics of the survey administration, the probability

of missing data was also unrelated to the research interest score that should have been obtained. To conclude, the missing data in this data set was considered as missing completely at random (MCAR), and the general linear regression models for longitudinal data analyses proposed in the previous chapter are able to yield valid inferences when missing data are MCAR.

Covariance Structure and Mean Model Selection

As discussed in the methodology chapter, only one independent variable was focused and investigated in each model. In that particular model, covariance structures (i.e., unstructured covariance, compound symmetry, Toeplitz, and autoregressive) were compared in each of the three mean models (i.e., linear, quadratic, and spline models). After appropriate covariance structures were selected respectively, the three mean models were compared and one mean model was selected to represent the data in this study and to address the research questions proposed previously.

Model with gender as the focus independent variable. In this model, gender was treated as the focus independent variable in the series of analyses. Meanwhile, all the other variables were included in the model. What follows is a description of the covariance structure selection and mean model selection for the gender model.

Covariance structure selection. First, it was assumed that the mean model followed a linear trend. The four covariance structures proposed in the previous chapter were applied respectively in the model. For each covariance structure, negative two log likelihood (-2LL), Akaike information criterion (AIC), and Bayesian information criterion (BIC) were obtained for comparison ("Linear Model" in Table 4-19a). Since compound symmetry, Toeplitz, and autoregressive models could be considered as nested

models of unstructured covariance structure, these three structures could be compared to the unstructured covariance structure directly by using likelihood ratio test. According to the likelihood ratio test, compound symmetry, Toeplitz, and autoregressive covariance patterns did not provide adequate fits to the covariance of the data used in this study when compared to the unstructured covariance (all p's < 0.001). As a result, unstructured covariance model was selected to represent the covariance pattern of this data set assuming that the gender model followed a linear trend.

Second, it was assumed that the mean model presented a quadratic model. The four covariance structures proposed were tested respectively in the model. For each covariance structure, -2LL, AIC, and BIC were obtained for comparison ("Quadratic Model" in Table 4-19a). Compound symmetry, Toeplitz, and autoregressive covariance structures were compared to the unstructured covariance structure directly based on likelihood ratio test. According to the likelihood ratio test, compound symmetry, Toeplitz, and autoregressive structures did not provide adequate fits to the covariance of the data used in this study when compared to the unstructured covariance (all p's < 0.001). As a result, unstructured covariance model was selected to represent the covariance pattern of this data set assuming that the gender model followed a quadratic trend.

Lastly, it was assumed that the mean model followed a piecewise linear trend (i.e., spline model). The four covariance structures were examined respectively in the model. For each covariance structure, -2LL, AIC, and BIC were obtained for comparison ("Spline Model" in Table 4-19a). Compound symmetry, Toeplitz, and autoregressive covariance structures were compared to the unstructured covariance structure directly. According to the likelihood ratio test, compound symmetry, Toeplitz, and autoregressive

structures did not provide adequate fits to the covariance of the data when compared to the unstructured covariance (all p's < 0.001). As a result, unstructured covariance model was selected to represent the covariance pattern of this data set assuming that the gender model was a spline model.

Mean model selection. With the results obtained above, one appropriate covariance structure was selected in each mean model. What follows was the comparison of the three mean models (linear, quadratic, and spline models) with their respective appropriate covariance structure (i.e., unstructured covariance pattern, which was found to be appropriate for all the three mean models). For each mean model, the -2LL, AIC and BIC were obtained (Table 4-19b). In the comparison between mean models, the linear model could both be considered as a nested model of the quadratic and spline models respectively. As a result, the he quadratic and spline models could be compared to the linear model directly based on the likelihood ratio test. Results showed that only a linear trend over time could not adequately account for the pattern of change in research interest levels by gender. The quadratic and the spline model were better fits. In addition, since quadratic and spline models had the same number of estimated parameters, so the two models could also be compared directly through likelihood ratio test. Based on the -2LL, AIC and BIC statistics, it could be indicated that the quadratic and spline model were equally better model to represent the data compared to the linear model. Therefore, in the following data analyses and result interpretation, only one model was selected to address the research questions proposed in this study. Since there were only three time points and the trend in research interest over time shown in Figure 4-3 presented a

piecewise linear trend and did not present a whole quadratic picture, therefore spline model was selected for the gender model to represent the data in this study.

In short, for the model with gender as the focus independent variable, a spline model with unstructured covariance matrix was selected for longitudinal data analysis to address the research questions.

Model with one race/ethnicity as the focus independent variable. It should be noted again that in this analysis, the White group was considered as the reference group. For each of the other racial/ethnic groups, there was a dummy coded variable, aiming to compare that particular racial/ethnic group to the White group. In this study, the other races/ethnicities included Asian/Pacific Islander, Black, Hispanic, and American Indian/Alaska Native American groups. As a result, in each model, only one race/ethnicity was the focus independent variable. Meanwhile, in that particular model, the covariance structure selection and mean model selection were conducted. That is to say, the covariance structure selection and mean model selection were conducted respectively for totally four times (Asian/Pacific Islander vs. White, Black vs. White, Hispanic vs. White, and American Indian/Alaska Native American vs. White). For each time, the covariance structure comparison and selection as well as model comparison and selection followed the same procedures and rules as illustrated above in the model with gender as the focus independent variable.

Covariance structure selection. The covariance structure comparison process was the same as described in the model with gender as the focus independent variable. For each of the four racial/ethnic group comparison models, the -2LL, AIC and BIC were obtained and compared respectively in the linear model, the quadratic model, and the

spline model (Asian/Pacific Islander vs. White: Table 4-20a; Black vs. White: Table 4-21a; Hispanic vs. White: Table 4-22a; and American Indian/Alaska Native American vs. White: Table 4-23a). Results showed that for all the four the racial/ethnic group comparison models, unstructured covariance was the most appropriate covariance structure among the proposed covariance structures and therefore was selected respectively to represent to covariance structure of the data in this study.

Mean model selection. With the unstructured covariance selected in each of the mean model (i.e., linear, quadratic, and spline models), the same analyses as illustrated in the model with gender as the focus independent variable were conducted to compare linear, quadratic, and spline models in each of the four racial/ethnic group comparison models (Asian/Pacific Islander vs. White: Table 4-20b; Black vs. White: Table 4-21b; Hispanic vs. White: Table 4-22b; and American Indian/Alaska Native American vs. White: Table 4-23b). Results also indicated that quadratic and spline model could equally better represent the data in this study compared to linear model. For the same reasons discussed previously, the spline model was selected respectively to represent the data in the analyses of each comparison among racial/ethnic groups in terms of the trajectory of students' research interest over time.

In short, for the model with each of the races/ethnicities as the focus independent variable, a spline model with unstructured covariance matrix was selected respectively for longitudinal data analysis to address the research questions.

Model with one previous research experience as the focus independent variable. Regarding students' previous research experience, three variables were examined in this study: whether students participated in high school laboratory research

apprenticeship, whether students participated in high school classroom-based programs, and whether students participated in college laboratory research apprenticeship. Each variable was considered as the focus independent variable at a time. As a result, covariance structure selection and mean model selection were conducted respectively in the three models.

Covariance structure selection. For all the three related focus independent variables, the covariance structure comparison process was the same as illustrated previously. For each model with one focus independent variable, -2LL, AIC, and BIC were obtained and compared respectively in the linear model, the quadratic model, and the spline model (high school laboratory research participation: Table 4-24a; high school classroom-based program participation: Table 4-25a; and college laboratory research participation: Table 4-26a). Results indicated that for each model the unstructured covariance was the most appropriate covariance structure among the proposed covariance structures according to likelihood ratio test and thus selected respectively to represent the covariance structure of the data in this study.

Mean model selection. With the unstructured covariance selected in each mean model (i.e., linear, quadratic, and spline models), the same analyses as described previously were conducted to compare linear, quadratic, and spline models in each of the three models (high school laboratory research participation: Table 4-24b; high school classroom-based program participation: Table 4-25b; and college laboratory research participation: Table 4-26b). Results indicated that the spline model was selected to represent the data in the study to compare respectively the research interest change over time of students who participated in high school laboratory research apprenticeship, high

school classroom-based programs and college laboratory research apprenticeship to the research interest change over time of students who did not.

In short, for the model with each of the previous research experiences as the focus independent variable, a spline model with unstructured covariance matrix was selected for longitudinal data analysis to address the research questions.

Model with matriculated program as the focus independent variable. In this model, the degree program at medical school matriculation was treated as the focus independent variable. At the same time, all the other variables were included in the model as control variables. Covariance structure selection and mean model selection for the matriculated program model were discussed below.

Covariance structure selection. The same as all the covariance structure selection analyses above, four covariance structures proposed were examined respectively in each of the three mean models (Table 4-27a). Based on likelihood ratio test, unstructured covariance was selected respectively in the three mean models to represent the covariance structure of the data in this study.

Mean model selection. After the same mean model selection analyses as stated previously (Table 4-27b), the spline model was selected for further data analysis of comparison between students who were matriculated in MD/PhD programs and those who were matriculated in MD-only programs to address the three research questions proposed.

In short, for the model with matriculated program as the focus independent variable, a spline model with unstructured covariance matrix was selected for longitudinal data analysis to address the research questions.

General Linear Regression Models

Through covariance structure and mean model comparison, the selected covariance structure and mean model were used in each corresponding model to address the three research questions proposed in this study. As a review, the research questions to be addressed are:

(1) Does medical students' reported research interest differ among students with different characteristics of gender, race/ethnicity, previous research experiences, or matriculated program prior to their entry to medical schools?

(2) Does medical students' reported research interest change in general across time from prior to their entry to medical schools, to when they are matriculated in medical schools, and to when they graduate from medical schools?

(3) Are patterns of change in medical students' reported research interest across time associated with gender, race/ethnicity, previous research experiences, or matriculated program?

As proposed in the methodology chapter, the answers to the three proposed research questions will be addressed within each comparison (i.e., gender, race/ethnicity, educational experience, and matriculated program), and finally summarized as a whole picture in the end of this chapter.

Gender. Through general linear regression model with unstructured covariance and spline model used, the results are summarized in Table 4-28. The estimated model is shown below. It should be noted again that the all the formulas presented in this chapter only contains the focus independent variable; however, in the actual analyses, the control variables were all included.

$$E(Y_{ij}) = 2.73 - 0.38 \times Time_{ij} + 0.41 \times (Time_{ij} - t^*)_+ - 0.04 \times Female_i - 0.05 \times Time_{ij} \times Female_i + 0.02 \times (Time_{ij} - t^*)_+ \times Female_i \quad (4.1)$$

Where

Y_{ij} denotes the response variable (in this case research interest) for the i^{th} individual at the j^{th} occasion (in this case PMQ, MSQ, or GQ);

$E(Y_{ij})$ denotes the expectation of Y_{ij};

$Time_{ij}$ denotes the value of the "time" variable (discussed above as an independent variable) for the j^{th} occasion on the i^{th} individual; and

$Female_i$ denotes the value of the "Female" variable of the i^{th} individual; and

t^* denotes the time knot (which is 1 in this case), and $(Time_{ij} - t^*)_+ = \max(0, Time_{ij} - t^*)$.

Longitudinal data analysis results were interpreted in line with the three research questions. First, female students reported significantly higher research interest levels than male students prior to their entry to medical schools, though such statistical difference was not quite practically significant ($t(1) = -2.46$; $p = 0.014$; d[5] $= 0.02$). Second, after considering gender and all the other variables included in the model, students' reported research interest levels significantly decreased from prior to their entry to medical schools, to when they were matriculated in medical schools ($t(1) = -36.89$; $p < 0.001$; $d = 0.31$); such decrease was significantly offset from when they were matriculated in medical schools to when they graduated from medical schools ($t(1) = 29.85$; $p < 0.001$; $d = 0.25$). Third, the decrease in research interest from prior to entry to medical schools to matriculation among female students was significantly larger than that among male

[5] d denotes the effect size discussed in Chapter 3.

students, though such difference was not quite practically significant ($t(1) = -3.12$; $p = 0.002$; $d = 0.03$); the change (i.e., offset) in research interest levels from before matriculation to after matriculation among female students was not significantly different from that among male students ($t(1) = 1.27$; $p = 0.203$).

Race/Ethnicity. Again, since the Asian/Pacific Islander, Black, Hispanic, and American Indian/Alaska Native American students were compared to the White students (reference group) respectively, the results were also illustrated separately corresponding to each comparison.

Asian/Pacific Islander versus White. In the comparison between Asian/Pacific Islander and White students, the estimated model is shown as below based on the data analysis results (Table 4-29).

$$E(Y_{ij}) = 2.74 - 0.39 \times Time_{ij} + 0.42 \times (Time_{ij} - t^*)_+ + 0.33 \times AsianPI_i - 0.10 \times Time_{ij} \times AsianPI_i + 0.03 \times (Time_{ij} - t^*)_+ \times AsianPI_i \quad (4.2)$$

Where $AsianPI_i$ denotes the value of the "AsianPI" variable of the i^{th} individual.

Null hypotheses were tested in order to address the three research questions. First, Asian/Pacific Islander students reported significantly higher research interest levels than White students prior to their entry to medical schools ($t(1) = 15.18$; $p < 0.001$; $d = 0.27$). Second, after considering race/ethnicity and all the other variables included in the model, students' reported research interest significantly decreased from prior to their entry to medical schools, to when they were matriculated in medical schools ($t(1) = -49.86$; $p < 0.001$; $d = 0.31$); such decrease was significantly offset from when they were matriculated in medical schools to when they graduated from medical schools ($t(1) = 40.21$; $p < 0.001$; $d = 0.25$). Third, the decrease in research interest from prior to entry to

medical schools to matriculation among Asian/Pacific Islander students was significantly larger than that among White students ($t(1) = -5.16$; $p < 0.001$; $d = 0.11$); the change (i.e., offset) in research interest from before matriculation to after matriculation among Asian/Pacific Islander students was not significantly different from that among White students ($t(1) = 1.26$; $p = 0.207$).

Black versus White. In the model of comparison between Black and White students, the estimated model is shown as below based on the data analysis results (Table 4-30).

$$E(Y_{ij}) = 2.75 - 0.40 \times Time_{ij} + 0.42 \times (Time_{ij} - t^*)_+ + 0.36 \times Black_i - 0.21 \times Time_{ij} \times Black_i + 0.18 \times (Time_{ij} - t^*)_+ \times Black_i \quad (4.3)$$

Where $Black_i$ denotes the value of the "Black" variable of the i^{th} individual.

Interpretations of the results were described corresponding to the three research questions. First, Black students reported significantly higher research interest levels than White students prior to their entry to medical schools ($t(1) = 10.37$; $p < 0.001$; $d = 0.27$). Second, after considering race/ethnicity and all the other variables included in the model, students' reported research interest significantly decreased from prior to their entry to medical schools, to when they were matriculated in medical schools ($t(1) = -53.20$; $p < 0.001$; $d = 0.31$); such decrease was significantly offset from when they were matriculated in medical schools to when they graduated from medical schools ($t(1) = 41.95$; $p < 0.001$; $d = 0.25$). Third, the decrease in research interest from prior to entry to medical schools to matriculation among Black students was significantly larger than that among White students ($t(1) = -6.53$; $p < 0.001$; $d = 0.20$); the change (i.e., offset) in research interest from before matriculation to after matriculation among Black students

was significantly different from that among White students ($t(1) = 4.24$; $p < 0.001$; $d = 0.13$).

Hispanic versus White. The comparison results between Hispanic and White students are presented in Table 4-31. The corresponding estimated model is shown as below.

$$E(Y_{ij}) = 2.75 - 0.40 \times Time_{ij} + 0.42 \times (Time_{ij} - t^*)_+ + 0.33 \times Hispanic_i - 0.16 \times Time_{ij} \times Hispanic_i + 0.14 \times (Time_{ij} - t^*)_+ \times Hispanic_i \quad (4.4)$$

Where $Hispanic_i$ denotes the value of the "Hispanic" variable of the i^{th} individual.

The three research questions were addressed based on the model and results shown above. First, Hispanic students reported significantly higher research interest levels than White students prior to their entry to medical schools ($t(1) = 10.47$; $p < 0.001$; $d = 0.26$). Second, after considering race/ethnicity and all the other variables included in the model, students' reported research interest significantly decreased from prior to their entry to medical schools, to when they were matriculated in medical schools ($t(1) = -52.93$; $p < 0.001$; $d = 0.31$); such decrease was significantly offset from when they were matriculated in medical schools to when they graduated from medical schools ($t(1) = 41.76$; $p < 0.001$; $d = 0.25$). Third, the decrease in research interest from prior to entry to medical schools to matriculation among Hispanic students was significantly larger than the that among White students ($t(1) = -5.54$; $p < 0.001$; $d = 0.18$); the change (i.e., offset) in research interest from before matriculation to after matriculation among Hispanic students was significantly different from that among White students ($t(1) = 3.53$; $p < 0.001$; $d = 0.11$).

American Indian/Alaska Native American versus White. The estimated model of comparison between American Indian/Alaska Native American and White groups is shown below based on the results presented in Table 4-32.

$$E(Y_{ij}) = 2.76 - 0.41 \times Time_{ij} + 0.43 \times (Time_{ij} - t^*)_+ - 0.25 \times Native_i + 0.20 \times Time_{ij} \times Native_i - 0.39 \times (Time_{ij} - t^*)_+ \times Native_i \quad (4.5)$$

Where $Native_i$ denotes the value of the "Native" variable of the i^{th} individual.

The three research questions were addressed respectively. First, American Indian/Alaska Native American students' reported research interest levels were not significantly different from those reported by White students prior to their entry to medical schools ($t(1) = -1.79$; $p = 0.074$). Second, after considering race/ethnicity and all the other variables included in the model, students' reported research interest significantly decreased from prior to their entry to medical schools, to when they were matriculated in medical schools ($t(1) = -56.30$; $p < 0.001$; $d = 0.31$); such decrease was significantly offset from when they were matriculated in medical schools to when they graduated from medical schools ($t(1) = 44.27$; $p < 0.001$; $d = 0.25$). Third, the decrease in research interest from prior to entry to medical schools to matriculation among American Indian/Alaska Native American students was not significantly different from the that among White students ($t(1) = 1.60$; $p = 0.109$); however, the change (i.e., offset) in research interest from before matriculation to after matriculation among American Indian/Alaska Native American students was significantly different from that among White students ($t(1) = -2.35$; $p = 0.019$; $d = 0.26$).

Previous research experiences. As discussed previously, three models were examined respectively concerning students' previous research experiences in order to

address the research questions. Correspondingly, results were explained separately for each model.

High school laboratory research apprenticeship. The estimated model of comparison between students who participated in laboratory research apprenticeship in high school and students who did not is presented below based on the results from Table 4-33.

$$E(Y_{ij}) = 2.74 - 0.39 \times Time_{ij} + 0.42 \times (Time_{ij} - t^*)_+ + 0.33 \times HS\_LAB_i - 0.13 \times Time_{ij} \times HS\_LAB_i + 0.05 \times (Time_{ij} - t^*)_+ \times HS\_LAB_i \quad (4.6)$$

Where HS_LAB_i denotes the value of the "HS_LAB" variable of the i^{th} individual.

According to the estimated model and the results, the three research questions were addressed. First, students with high school laboratory research experiences reported significantly higher research interest levels than students without such experiences prior to their entry to medical schools ($t(1) = 12.53$; $p < 0.001$; $d = 0.31$). Second, after considering high school laboratory research participation and all the other variables included in the model, students' reported research interest significantly decreased from prior to their entry to medical schools, to when they were matriculated in medical schools ($t(1) = -51.61$; $p < 0.001$; $d = 0.31$); such decrease was significantly offset from when they were matriculated in medical schools to when they graduated from medical schools ($t(1) = 41.38$; $p < 0.001$; $d = 0.25$). Third, the decrease in research interest from prior to entry to medical schools to matriculation among students with high school laboratory research experiences was significantly larger than that among other students ($t(1) = -5.33$; $p < 0.001$; $d = 0.11$); however, the change (i.e., offset) in research interest from before

matriculation to after matriculation among students with high school lab experiences was not significantly different from the change among other students ($t(1) = 1.45$; $p = 0.148$).

High school program. A comparison between students who participated in classroom-based program during high school and students who did not was conducted in terms of the trend of research interest levels over time (Table 4-34). The estimated model is shown below.

$$E(Y_{ij}) = 2.76 - 0.40 \times Time_{ij} + 0.41 \times (Time_{ij} - t^*)_+ + 0.03 \times HS\_PROG_i - 0.06 \times Time_{ij} \times HS\_PROG_i + 0.13 \times (Time_{ij} - t^*)_+ \times HS\_PROG_i \quad (4.7)$$

Where HS_PROG_i denotes the value of the "HS_PROG" variable of the i^{th} individual.

The three research questions were discussed based on the estimated model and the results. First, the research interest levels reported by students with high school classroom-based program experiences prior to their entry to medical schools was not significantly different from those reported by the students without such experiences ($t(1) = 1.14$; $p = 0.254$). Second, after considering high school classroom-based program participation and all the other variables included in the model, students' reported research interest significantly decreased from prior to their entry to medical schools, to when they were matriculated in medical schools ($t(1) = -52.64$; $p < 0.001$; $d = 0.31$); such decrease was significantly offset from when they were matriculated in medical schools to when they graduated from medical schools ($t(1) = 40.85$; $p < 0.001$; $d = 0.25$). Third, the decrease in the research interest from prior to entry to medical schools to matriculation among students with high school classroom-based program experiences is significantly larger than that among other students ($t(1) = -2.55$; $p = 0.011$; $d = 0.05$); the change in research

interest from before matriculation to after matriculation among students with high school program experiences was significantly different from the change among other students ($t(1) = 3.85$; $p < 0.001$; $d = 0.07$).

College laboratory research apprenticeship. With regard to the laboratory research apprenticeship participation in college, the estimated model based on the results (Table 4-35) is presented below.

$$\boldsymbol{E}(Y_{ij}) = 2.63 - 0.30 \times Time_{ij} + 0.38 \times (Time_{ij} - t^*)_+ + 0.55 \times COLL\_LAB_i - 0.24 \times Time_{ij} \times COLL\_LAB_i + 0.10 \times (Time_{ij} - t^*)_+ \times COLL\_LAB_i \quad (4.8)$$

Where $COLL\_LAB_i$ denotes the value of the "COLL_LAB" variable of the i^{th} individual.

What follows is the discussion about the three research questions respectively. First, students with college laboratory research experiences reported significantly higher research interest levels than other students prior to their entry to medical schools ($t(1) = 34.79$; $p < 0.001$; $d = 0.43$). Second, after considering college laboratory research participation and all the other variables included in the model, students' reported research interest significantly decreased from prior to their entry to medical schools, to when they were matriculated in medical schools ($t(1) = -32.34$; $p < 0.001$; $d = 0.31$); such decrease was significantly offset from when they were matriculated in medical schools to when they graduated from medical schools ($t(1) = 30.33$; $p < 0.001$; $d = 0.25$). Third, the decrease in research interest from prior to entry to medical schools to matriculation among students with college laboratory research experiences was significantly larger than the that among other students ($t(1) = -16.51$; $p < 0.001$; $d = 0.19$); the change (i.e., offset) in research interest from before matriculation to after matriculation among students with

high school lab experiences was significantly different from that among other students, though such difference was not quite practically significant ($t(1) = 5.09$; $p < 0.001$; $d = 0.06$).

Comparison between high school research experience and college research experience. Based on the results discussed previously, it could be indicated that high school research experience and college research experience seemed to have similar results in terms of being positively associated with students' long-term research interest. As discussed in the descriptive analyses earlier in this chapter, less than ten percent of the sample students reported high school research experiences, while more than one third reported college research experiences. Further analyses were conducted (based on the variable "LAB_TIME") to compare the students with high school research experience (coded as 1) against the students with no high school research experience but who reported college research experience (coded as 0). After covariance structure comparisons for the linear model, the quadratic model, and the spline model respectively (Table 4-36a), unstructured covariance was selected in each of the three mean models to represent the covariance of the data set based on the likelihood ratio test. The same as other mean model comparisons discussed previously, the spline model was selected according to the likelihood ratio test to estimate the parameters to examine the comparison between the two groups of interest (Table 4-36b). With the unstructured covariance and the spline model applied, the estimated model based on the general linear regression model results (Table 4-37 and Figure 4-9) is presented below.

$$E(Y_{ij}) = 3.20 - 0.54 \times Time_{ij} + 0.47 \times (Time_{ij} - t^*)_+ + 0.08 \times LAB\_TIME_i + 0.01 \times Time_{ij} \times LAB\_TIME_i - 0.01 \times (Time_{ij} - t^*)_+ \times LAB\_TIME_i \quad (4.9)$$

Where LAB_TIME_i denotes the value of the "LAB_TIME" variable of the i^{th} individual.

Results were interpreted in order to provide evidence about the comparison between students who participated in high school research apprenticeship ("early research participants" for short in this paragraph) and students who did not participated in high school research apprenticeship but who participated in college research apprenticeship ("later research participants" for short in this paragraph). When registering the MCAT, the early research participants reported significantly higher research interest than later research participants, though such difference was not quite practical significant ($t(1) = 2.68$; $p = 0.007$; $d = 0.07$). As time went, there was no significant interaction between the reported research interest levels and students' group membership ($t(1)_{before\ MSQ} = 0.58$, $p_{before\ MSQ} = 0.565$; $t(1)_{after\ MSQ} = -0.19$, $p_{after\ MSQ} = 0.853$). That is to say, the patterns of change in the reported research interest levels over time were the same for the two groups of students (i.e., early research participants and later research participants). In other words, the difference in research interest levels between the two groups was consistent over time until when they graduated from medical schools.

Matriculated program. In the model related to the comparison between students who were matriculated in the MD/PhD programs and students who were matriculated in the MD-only programs, the estimated model based on the results (Table 4-38) is shown below.

$$E(Y_{ij}) = 2.76 - 0.41 \times Time_{ij} + 0.43 \times (Time_{ij} - t^*)_+ + 1.51 \times PROGRAM_i + 0.09 \times Time_{ij} \times PROGRAM_i - 0.76 \times (Time_{ij} - t^*)_+ \times PROGRAM_i \quad (4.\ 10)$$

Where $PROGRAM_i$ denotes the value of the "PROGRAM" variable of the i^{th} individual.

Regarding this comparison, the three research questions were discussed respectively. First, MD/PhD students reported significantly higher research interest levels than MD students prior to their entry to medical schools, and meanwhile the effect size is noticeably large ($t(1) = 19.27$; $p < 0.001$; $d = 1.11$). Second, after considering matriculated program and all the other variables included in the model, students' reported research interest significantly decreased from prior to their entry to medical schools, to when they were matriculated in medical schools ($t(1) = -56.22$; $p < 0.001$; $d = 0.31$); such decrease was significantly offset from when they were matriculated in medical schools to when they graduated from medical schools ($t(1) = 44.94$; $p < 0.001$; $d = 0.25$). Third, the decrease in research interest from prior to entry to medical schools to matriculation among MD/PhD students was not significantly different from the that among MD students ($t(1) = 1.25$; $p = 0.212$); however, the change (i.e., offset) in research interest from before matriculation to after matriculation among MD/PhD students was significantly different from the change among MD students, which also presented a large effect size ($t(1) = -8.42$; $p < 0.001$; $d = 0.46$).

Summary of Findings

This chapter consists of descriptive analysis results and longitudinal data analysis results. Before the research questions were addressed, descriptive analyses were conducted on the independent, dependent, and control variables. Descriptive results showed that female and male students were almost evenly distributed in the study sample. White students were the majority group, followed by Asian/Pacific Islander, Hispanic, Black, and American Indian/Alaska Native American groups. The percentage of the students who reported college research related experiences was larger than that of the students who reported high school research related experiences. The majority of the sample students were matriculated in the MD-only programs; while still a group of students were matriculated in the MD/PhD programs, but accounted for only a small portion of the sample. These variables were found to be associated with students' reported research interest levels over time as the research questions were discussed.

After the unstructured covariance pattern and the spline model were selected respectively to represent the data for each model with one focus independent variable, general linear regression models were conducted to address the research questions. To summarize, the first research question examined whether the reported research interest differed among students with different characteristics prior to their entry to medical schools. Results suggested that female students reported significantly lower research interest than male students in this study. Asian/Pacific Islander, Black and Hispanic students indicated higher research interest than White students respectively. Students who participated in laboratory research apprenticeship (in high school and in college respectively) showed higher interest in research than students who did not. Students who

were matriculated in the MD/PhD programs were more interested in research than students who were matriculated in the MD programs.

The second research question in this study sought to explore the general trend of medical students' reported research interest levels over time. The sample students' reported research interest levels changed in the similar pattern across all the models after controlling for all the related variables discussed in the methodology chapter. In general, the students' reported research interest decreased from when they registered the MCAT to when they were matriculated in medical schools; and then such decrease was relieved after the matriculation until when they graduated from medical schools.

The third research question explored the association between the patterns of change over time in students' reported research interest levels and their different characteristics. Results indicated that the change across time in medical students' research interest was significantly associated with students' gender, race/ethnicity, whether students' reported previous research experiences, and students' matriculated program. Overall, this longitudinal study found that medical students' reported research interest waned over time in general, and also detected the differences in the long-term research interest among students from different groups.

Table 4-1

Gender Distribution

Gender	n	%
Female	20,464	51.4
Male	19,375	48.6
Total	39,839	100.0

Table 4-2

Race/Ethnicity Distribution

Race/Ethnicity	n	%
White	26,967	67.7
Black	2,246	5.6
Hispanic	2,632	6.6
Asian/Pacific Islander	6,688	16.8
American Indian/Alaska Native American	126	0.3
Multiple Races/Ethnicities	1,011	2.5
No Response	169	0.4
Total	39,839	100.0

Table 4-3

Descriptive Statistics of Age at the MCAT

	Mean	Standard Deviation	Minimum	Maximum
Age	21.56	2.65	13	51

Table 4-4

Distribution of Students' Parental Education Level

Parental Education Level	n	%
At Least One Parent Has College or Above Education	32,646	84.5
Neither Parent Has College or Above Education	5,993	15.5
Total	38,639	100.0

Table 4-5

Distribution of Students' Parental Profession

Parental Profession	n	%
At Least One Parent Works in Health Related Fields	13,981	36.9
Neither Parent Works in Health Related Fields	23,937	63.1
Total	37,918	100.0

Table 4-6

Distribution of Students' High School Laboratory Research Apprenticeship Participation

High School Lab Research	n	%
Participated	3,666	9.2
Not Participated	36,173	90.8
Total	39,839	100.0

Table 4-7

Distribution of Students' High School Classroom-Based Summer, After-School, or Saturday Program Participation

High School Program	n	%
Participated	3,536	8.9
Not Participated	36,303	91.1
Total	39,839	100.0

Table 4-8

Distribution of Students' College Laboratory Research Apprenticeship Participation

College Lab Research	n	%
Participated	14,596	36.6
Not Participated	25,243	63.4
Total	39,839	100.0

Table 4-9

Distribution of Students' Matriculated Program

Matriculated Program	n	%
MD/PhD Program	377	1.0
MD-Only Program	38,684	97.1
MD/Other Programs	778	1.9
Total	39,839	100.0

Note. MD/Other Programs include: BA/MD, MS/MD, MD/JD, MD/Other, BS/MD, MD/MBA, MD/MPH, and MD/Dental (OMS).

Table 4-10

Principal Component Analysis Results for the MSQ Research Interest Levels

	First Principal Component		Second Principal Component	
	Eigenvalue	% of Variance	Eigenvalue	% of Variance
MSQ Research Interest Level	1.70	85.00	0.30	15.00

Table 4-11

Mean Research Interest Levels Over Time

	Mean Research Interest Level					
	Mean	Standard Deviation	Minimum	Maximum	Skewness	Kurtosis
PMQ	3.05	1.47	1	5	0.004	-1.551
MSQ	2.65	0.90	1	5	0.234	-0.682
GQ	2.66	0.83	1	5	-0.161	-0.413

Table 4-12

Mean Research Interest Levels over Time by Gender

	Mean Research Interest Level by Gender			
	Female		Male	
	Mean	Standard Deviation	Mean	Standard Deviation
PMQ	3.04	1.48	3.06	1.47
MSQ	2.61	0.89	2.69	0.90
GQ	2.62	0.81	2.71	0.85

Table 4-13

Mean Research Interest Levels over Time by Race/Ethnicity

	Mean Research Interest Level by Race/Ethnicity									
	White		Black		Hispanic		Asian/Pacific Islander		Native	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PMQ	2.93	1.47	3.33	1.45	3.31	1.47	3.33	1.45	2.78	1.42
MSQ	2.58	0.89	2.74	0.91	2.76	0.90	2.83	0.88	2.52	0.89
GQ	2.62	0.83	2.71	0.83	2.74	0.84	2.79	0.82	2.36	0.85

Note. SD = Standard Deviation; Native = American Indian/Alaska Native American.

Table 4-14

Mean Research Interest Levels over Time by High School Laboratory Research Apprenticeship Participation

	Mean Research Interest Level by High School Lab Participation			
	Participated		Not Participated	
	Mean	Standard Deviation	Mean	Standard Deviation
PMQ	3.46	1.43	3.00	1.47
MSQ	2.95	0.90	2.62	0.89
GQ	2.88	0.81	2.64	0.83

Table 4-15

Mean Research Interest Levels over Time by High School Classroom-Based Summer, After-School, or Saturday Program Participation

	Mean Research Interest Level by High School Program Participation			
	Participated		Not Participated	
	Mean	Standard Deviation	Mean	Standard Deviation
PMQ	3.16	1.46	3.04	1.48
MSQ	2.69	0.90	2.65	0.90
GQ	2.76	0.83	2.65	0.83

Table 4-16

Mean Research Interest Levels over Time by College Laboratory Research Apprenticeship Participation

	Mean Research Interest Level by College Lab Participation			
	Participated		Not Participated	
	Mean	Standard Deviation	Mean	Standard Deviation
PMQ	3.41	1.44	2.79	1.45
MSQ	2.87	0.90	2.52	0.87
GQ	2.79	0.81	2.58	0.83

Table 4-17

Mean Research Interest Levels over Time by Matriculated Program

	Mean Research Interest Level by Matriculated Program			
	MD/PhD		MD-Only	
	Mean	Standard Deviation	Mean	Standard Deviation
PMQ	4.66	0.77	3.04	1.47
MSQ	4.33	0.43	2.63	0.88
GQ	3.67	0.67	2.65	0.83

Table 4-18

Correlation Analysis for Missing Data Evaluation

	Missing Data for Research Interest Levels Over Time			
	PMQ P_INTEREST_RSC	MSQ FAC_RESEARCH	MSQ MSQ_CAREER_RESEARCH	GQ GQ_CAREER_RESEARCH
P_INTEREST_RSC	N/A	0.0157	0.0161	0.0246
FAC_RESEARCH	0.0256	N/A	0.0028	0.0312
MSQ_CAREER_RESEARCH	0.0267	-0.0095	N/A	0.0224
GQ_CAREER_RESEARCH	0.0184	-0.0294	-0.0287	N/A
Female	-0.0280	-0.0055	-0.0053	-0.0115
White	-0.0449	-0.0063	-0.0064	-0.0311
Black	-0.0113	-0.0024	-0.0035	0.0143
Hispanic	-0.0164	-0.0078	-0.0074	0.0056
AsianPI	0.0754	0.0187	0.0185	0.0289
Native	-0.0061	-0.0073	-0.0074	-0.0051
Multiple	-0.0016	-0.0073	-0.0056	-0.0046
HS_LAB	-0.1124	-0.0151	-0.0155	0.0011
HS_PROG	-0.1094	-0.0197	-0.0204	-0.0044
COLL_LAB	-0.2701	-0.0168	-0.0179	-0.0137
PROGRAM	-0.0060	-0.0466	-0.0461	0.0059
AGE_AT_TEST	-0.0112	-0.0035	-0.0029	-0.0023
PARENTAL_EDUCATION	0.0365	0.0056	0.0055	0.0019
PARENTAL_PROFESSION	0.0215	0.0212	0.0204	0.0097

Table 4-19a

Covariance Structure Comparison in the Linear Model, the Quadratic Model, and the Spline Model for the Comparison between Female and Male Students

Covariance Structure	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear Model						
Unstructured	6	271350.4	271401.4	271338.4	N/A	N/A
CS	2	288499.0	288516.0	288495.0	4	17156.6***
Toeplitz	3	286233.0	286258.5	286227.0	3	14888.6***
AR	2	286243.6	286260.6	286239.6	4	14901.2***
Quadratic Model						
Unstructured	6	269472.1	269523.1	269460.1	N/A	N/A
CS	2	287418.2	287435.2	287414.2	4	17954.1***
Toeplitz	3	284776.0	284801.5	284770.0	3	15309.9***
AR	2	284775.5	284792.5	284771.5	4	15311.4***
Spline Model						
Unstructured	6	269469.3	269520.4	269457.3	N/A	N/A
CS	2	287415.4	287432.4	287411.4	4	17954.1***
Toeplitz	3	284773.2	284798.7	284767.2	3	15309.9***
AR	2	284772.8	284789.8	284768.8	4	15311.5***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood; CS = compound symmetry; AR = autoregressive.
[a] Compared to the unstructured covariance pattern.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-19b

Mean Model Comparison for the Comparison between Female and Male Students

Mean Model	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear	15	271265.4	271452.6	271221.4	N/A	N/A
Quadratic	17	269374.8	269579.0	269326.8	2	-1894.6***
Spline	17	269374.8	269579.0	269326.8	2	-1894.6***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood.
[a] Compared to the linear mean model.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-20a

Covariance Structure Comparison in the Linear Model, the Quadratic Model, and the Spline Model for the Comparison between Asian/Pacific Islander and White Students

Covariance Structure	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear Model						
Unstructured	6	271311.9	271362.9	271299.9	N/A	N/A
CS	2	288448.0	288465.0	288444.0	4	17144.1***
Toeplitz	3	286190.8	286216.3	286184.8	3	14884.9***
AR	2	286202.3	286219.3	286198.3	4	14898.4***
Quadratic Model						
Unstructured	6	269425.7	269476.8	269413.7	N/A	N/A
CS	2	287361.4	287378.4	287357.4	4	17943.7***
Toeplitz	3	284727.9	284753.4	284721.9	3	15308.2***
AR	2	284727.7	284744.8	284723.7	4	15310.0***
Spline Model						
Unstructured	6	269423.0	269474.0	269411.0	N/A	N/A
CS	2	287358.6	287375.6	287354.6	4	17943.6***
Toeplitz	3	284719.1	284750.6	284719.1	3	15308.1***
AR	2	284725.0	284742.0	284721.0	4	15310.0***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood; CS = compound symmetry; AR = autoregressive.
[a] Compared to the unstructured covariance pattern.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-20b

Mean Model Comparison for the Comparison between Asian/Pacific Islander and White Students

Mean Model	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear	15	271227.5	271414.7	271183.5	N/A	N/A
Quadratic	17	269329.7	269533.9	269281.7	2	-1901.8***
Spline	17	269329.7	269533.9	269281.7	2	-1901.8***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood.
[a] Compared to the linear mean model.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-21a

Covariance Structure Comparison in the Linear Model, the Quadratic Model, and the Spline Model for the Comparison between Black and White Students

Covariance Structure	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear Model						
Unstructured	6	271334.2	271385.3	271322.2	N/A	N/A
CS	2	288464.1	288481.1	288460.1	4	17137.9***
Toeplitz	3	286202.0	286227.5	286196.0	3	14873.8***
AR	2	286213.1	286230.1	286209.1	4	14886.9***
Quadratic Model						
Unstructured	6	269437.2	269488.2	269425.2	N/A	N/A
CS	2	287372.1	287389.1	287368.1	4	17942.9***
Toeplitz	3	284730.1	284755.7	284724.1	3	15298.9***
AR	2	284729.8	284746.8	284725.8	4	15300.6***
Spline Model						
Unstructured	6	269434.4	269485.4	269422.4	N/A	N/A
CS	2	287369.3	287386.3	287365.3	4	17942.9***
Toeplitz	3	284727.4	284752.9	284721.4	3	15299.0***
AR	2	284727.0	284744.0	284723.0	4	15300.6***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood; CS = compound symmetry; AR = autoregressive.
[a] Compared to the unstructured covariance pattern.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-21b

Mean Model Comparison for the Comparison between Black and White Students

Mean Model	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear	15	271250.8	271437.9	271206.8	N/A	N/A
Quadratic	17	269343.0	269547.2	269295.0	2	-1911.8***
Spline	17	269343.0	269547.2	269295.0	2	-1911.8***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood.
[a] Compared to the linear mean model.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-22a

Covariance Structure Comparison in the Linear Model, the Quadratic Model, and the Spline Model for the Comparison between Hispanic and White Students

Covariance Structure	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear Model						
Unstructured	6	271341.5	271392.5	271329.5	N/A	N/A
CS	2	288476.7	288493.7	288472.7	4	17143.2***
Toeplitz	3	286212.6	286238.2	286206.6	3	14877.1***
AR	2	286223.4	286240.5	286219.4	4	14889.9***
Quadratic Model						
Unstructured	6	269450.4	269501.5	269438.4	N/A	N/A
CS	2	287388.0	287405.0	287384.0	4	17945.6***
Toeplitz	3	284742.9	284768.4	284736.9	3	15298.5***
AR	2	284742.4	284759.4	284738.4	4	15300.0***
Spline Model						
Unstructured	6	269447.6	269498.7	269435.6	N/A	N/A
CS	2	287385.2	287402.2	287381.2	4	17945.6***
Toeplitz	3	284740.1	284765.6	284734.1	3	15298.5***
AR	2	284739.6	284756.6	284735.6	4	15300.0***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood; CS = compound symmetry; AR = autoregressive.
[a] Compared to the unstructured covariance pattern.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-22b

Mean Model Comparison for the Comparison between Hispanic and White Students

Mean Model	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear	15	271257.8	271445.0	271213.8	N/A	N/A
Quadratic	17	269355.9	269560.1	269307.9	2	-1905.9***
Spline	17	269355.9	269560.1	269307.9	2	-1905.9***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood.
[a] Compared to the linear mean model.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-23a

Covariance Structure Comparison in the Linear Model, the Quadratic Model, and the Spline Model for the Comparison between American Indian/Alaska Native American and White Students

Covariance Structure	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear Model						
Unstructured	6	271361.4	271412.4	271349.4	N/A	N/A
CS	2	288518.7	288535.7	288514.7	4	17165.3***
Toeplitz	3	286248.7	286274.2	286242.7	3	14893.3***
AR	2	286258.9	286275.9	286254.9	4	14905.5***
Quadratic Model						
Unstructured	6	269473.6	269524.7	269461.6	N/A	N/A
CS	2	287431.8	287448.8	287427.8	4	17966.2***
Toeplitz	3	284783.8	284809.3	284777.8	3	15316.2***
AR	2	284783.1	284800.2	284779.1	4	15317.5***
Spline Model						
Unstructured	6	269470.9	269521.9	269458.9	N/A	N/A
CS	2	287429.1	287446.1	287425.1	4	17966.2***
Toeplitz	3	284781.0	284806.5	284775.0	3	15316.1***
AR	2	284780.4	284797.4	284776.4	4	15317.5***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood; CS = compound symmetry; AR = autoregressive.
[a] Compared to the unstructured covariance pattern.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-23b

Mean Model Comparison for the Comparison between American Indian/Alaska Native American and White Students

Mean Model	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear	15	271280.7	271467.9	271236.7	N/A	N/A
Quadratic	17	269385.0	269589.2	269337.0	2	-1899.7***
Spline	17	269385.0	269589.2	269337.0	2	-1899.7***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood.
[a] Compared to the linear mean model.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-24a

Covariance Structure Comparison in the Linear Model, the Quadratic Model, and the Spline Model for the Comparison between Students Who Participated in High School Laboratory Research Apprenticeship and Students Who Did Not

Covariance Structure	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear Model						
Unstructured	6	271298.0	271349.1	271286.0	N/A	N/A
CS	2	288428.4	288445.4	288424.4	4	17138.4***
Toeplitz	3	286169.7	286195.3	286163.7	3	14877.7***
AR	2	286181.2	286198.2	286177.2	4	14891.2***
Quadratic Model						
Unstructured	6	269427.6	269478.6	269415.6	N/A	N/A
CS	2	287351.4	287368.4	287347.4	4	17931.8***
Toeplitz	3	284717.3	284742.8	284711.3	3	15295.7***
AR	2	284717.1	284734.1	284713.1	4	15297.5***
Spline Model						
Unstructured	6	269424.8	269475.9	269412.8	N/A	N/A
CS	2	287348.7	287365.7	287344.7	4	17931.9***
Toeplitz	3	284714.5	284740.1	284708.5	3	15295.7***
AR	2	284714.3	284731.4	284710.3	4	15297.5***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood; CS = compound symmetry; AR = autoregressive.
[a] Compared to the unstructured covariance pattern.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-24b

Mean Model Comparison for the Comparison between Students Who Participated in High School Laboratory Research Apprenticeship and Students Who Did Not

Mean Model	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear	15	271214.1	271401.2	271170.1	N/A	N/A
Quadratic	17	269332.4	269536.6	269284.4	2	-1885.7***
Spline	17	269332.4	269536.6	269284.4	2	-1885.7***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood.
[a] Compared to the linear mean model.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-25a

Covariance Structure Comparison in the Linear Model, the Quadratic Model, and the Spline Model for the Comparison between Students Who Participated in High School Classroom-Based Programs and Students Who Did Not

Covariance Structure	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear Model						
Unstructured	6	271364.7	271415.8	271352.7	N/A	N/A
CS	2	288521.7	288538.7	288517.7	4	17165.0***
Toeplitz	3	286251.5	286277.0	286245.5	3	14892.8***
AR	2	286261.7	286278.7	286257.7	4	14905.0***
Quadratic Model						
Unstructured	6	269469.8	269520.8	269457.8	N/A	N/A
CS	2	287431.2	287448.2	287427.2	4	17969.4***
Toeplitz	3	284780.8	284806.4	284774.8	3	15317.0***
AR	2	284780.2	284797.2	284776.2	4	15318.4***
Spline Model						
Unstructured	6	269467.0	269518.1	269455.0	N/A	N/A
CS	2	287428.5	287445.5	287424.5	4	17969.5***
Toeplitz	3	284778.1	284803.6	284772.1	3	15317.1***
AR	2	284777.4	284794.4	284773.4	4	15318.4***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood; CS = compound symmetry; AR = autoregressive.
[a] Compared to the unstructured covariance pattern.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-25b

Mean Model Comparison for the Comparison between Students Who Participated in High School Classroom-Based Programs and Students Who Did Not

Mean Model	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear	15	271280.8	271468.0	271236.8	N/A	N/A
Quadratic	17	269374.6	269578.8	269326.6	2	-1910.2***
Spline	17	269374.6	269578.8	269326.6	2	-1910.2***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood.
[a] Compared to the linear mean model.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-26a

Covariance Structure Comparison in the Linear Model, the Quadratic Model, and the Spline Model for the Comparison between Students Who Participated in College Laboratory Research Apprenticeship and Students Who Did Not

Covariance Structure	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear Model						
Unstructured	6	270770.7	270821.7	270758.7	N/A	N/A
CS	2	287694.5	287711.5	287690.5	4	16931.8***
Toeplitz	3	285530.7	285556.2	285524.7	3	14766.0***
AR	2	285553.8	285570.8	285549.8	4	14791.1***
Quadratic Model						
Unstructured	6	268950.0	269001.0	268938.0	N/A	N/A
CS	2	286633.8	286650.8	286629.8	4	17691.8***
Toeplitz	3	284100.5	284126.0	284094.5	3	15156.5***
AR	2	284105.3	284122.3	284101.3	4	15163.3***
Spline Model						
Unstructured	6	268947.2	268998.3	268935.2	N/A	N/A
CS	2	286631.0	286648.1	286627.0	4	17691.8***
Toeplitz	3	284097.7	284123.3	284091.7	3	15156.5***
AR	2	284102.5	284119.5	284098.5	4	15163.3***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood; CS = compound symmetry; AR = autoregressive.
[a] Compared to the unstructured covariance pattern.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-26b

Mean Model Comparison for the Comparison between Students Who Participated in College Laboratory Research Apprenticeship and Students Who Did Not

Mean Model	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear	15	270685.7	270872.9	270641.7	N/A	N/A
Quadratic	17	268852.8	269056.9	268804.8	2	-1836.9***
Spline	17	268852.8	269056.9	268804.8	2	-1836.9***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood.
[a] Compared to the linear mean model.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-27a

Covariance Structure Comparison in the Linear Model, the Quadratic Model, and the Spline Model for the Comparison between MD/PhD Program Enrollees and MD Program Enrollees

Covariance Structure	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear Model						
Unstructured	6	271255.6	271306.7	271243.6	N/A	N/A
CS	2	288447.2	288464.2	288443.2	4	17199.6***
Toeplitz	3	286186.0	286211.5	286180.0	3	14936.4***
AR	2	286197.2	286214.2	286193.2	4	14949.6***
Quadratic Model						
Unstructured	6	269300.0	269351.0	269288.0	N/A	N/A
CS	2	287320.9	287337.9	287316.9	4	18028.9***
Toeplitz	3	284667.5	284693.0	284661.5	3	15373.5***
AR	2	284667.0	284684.0	284663.0	4	15375.0***
Spline Model						
Unstructured	6	269297.2	269348.2	269285.2	N/A	N/A
CS	2	287318.1	287335.1	287314.1	4	18028.9***
Toeplitz	3	284664.7	284690.2	284658.7	3	15373.5***
AR	2	284664.2	284681.2	284660.2	4	15375.0***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood; CS = compound symmetry; AR = autoregressive.
[a] Compared to the unstructured covariance pattern.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-27b

Mean Model Comparison for the Comparison between MD/PhD Program Enrollees and MD Program Enrollees

Mean Model	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear	15	271173.8	271361.0	271129.8	N/A	N/A
Quadratic	17	269209.0	269413.1	269161.0	2	-1968.8***
Spline	17	269209.0	269413.1	269161.0	2	-1968.8***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood.
[a] Compared to the linear mean model.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-28

General Linear Regression Model with Gender ("Female") as the Focus Independent Variable

	Estimate	Standard Error	*t* Statistic	*p* Value
Intercept	2.7333	0.03564	76.70	<.0001
Time	-0.3837	0.01040	-36.89	<.0001
Timeknot	0.4129	0.01383	29.85	<.0001
Female	-0.03858	0.01571	-2.46	0.0140
Time*Female	-0.04508	0.01446	-3.12	0.0018
Timeknot*Female	0.02451	0.01927	1.27	0.2034
Control Variables		Included		

Note. Control variables include: races/ethnicities, previous research experiences, matriculated program, age, parental education level, parental profession.

Table 4-29

General Linear Regression Model with Asian/Pacific Islander ("AsianPI") as the Focus Independent Variable

	Estimate	Standard Error	*t* Statistic	*p* Value
Intercept	2.7402	0.03502	78.24	<.0001
Time	-0.3913	0.007849	-49.86	<.0001
Timeknot	0.4211	0.01047	40.21	<.0001
AsianPI	0.3317	0.02185	15.18	<.0001
Time*AsianPI	-0.1037	0.02011	-5.16	<.0001
Timeknot*AsianPI	0.03360	0.02664	1.26	0.2072
Control Variables		Included		

Note. Control variables include: gender, other races/ethnicities, previous research experiences, matriculated program, age, parental education level, parental profession.

Table 4-30

General Linear Regression Model with Black ("Black") as the Focus Independent Variable

	Estimate	Standard Error	*t* Statistic	*p* Value
Intercept	2.7496	0.03494	78.69	<.0001
Time	-0.3956	0.007436	-53.20	<.0001
Timeknot	0.4157	0.009909	41.95	<.0001
Black	0.3561	0.03433	10.37	<.0001
Time*Black	-0.2051	0.03144	-6.53	<.0001
Timeknot*Black	0.1778	0.04190	4.24	<.0001
Control Variables		Included		

Note. Control variables include: gender, other races/ethnicities, previous research experiences, matriculated program, age, parental education level, parental profession.

Table 4-31

General Linear Regression Model with Hispanic ("Hispanic") as the Focus Independent Variable

	Estimate	Standard Error	*t* Statistic	*p* Value
Intercept	2.7503	0.03495	78.69	<.0001
Time	-0.3962	0.007485	-52.93	<.0001
Timeknot	0.4164	0.009972	41.76	<.0001
Hispanic	0.3276	0.03130	10.47	<.0001
Time*Hispanic	-0.1591	0.02872	-5.54	<.0001
Timeknot*Hispanic	0.1354	0.03830	3.53	0.0004
Control Variables		Included		

Note. Control variables include: gender, other races/ethnicities, previous research experiences, matriculated program, age, parental education level, parental profession.

Table 4-32

General Linear Regression Model with American Indian/Alaska Native American ("Native") as the Focus Independent Variable

	Estimate	Standard Error	*t* Statistic	*p* Value
Intercept	2.7617	0.03491	79.12	<.0001
Time	-0.4077	0.007241	-56.30	<.0001
Timeknot	0.4270	0.009646	44.27	<.0001
Native	-0.2459	0.1376	-1.79	0.0738
Time*Native	0.2007	0.1251	1.60	0.1088
Timeknot*Native	-0.3884	0.1651	-2.35	0.0186
Control Variables	Included			

Note. Control variables include: gender, other races/ethnicities, previous research experiences, matriculated program, age, parental education level, parental profession.

Table 4-33

General Linear Regression Model with High School Laboratory Research Experience ("HS_LAB") as the Focus Independent Variable

	Estimate	Standard Error	*t* Statistic	*p* Value
Intercept	2.7445	0.03498	78.46	<.0001
Time	-0.3938	0.007630	-51.61	<.0001
Timeknot	0.4199	0.01015	41.38	<.0001
HS_LAB	0.3285	0.02622	12.53	<.0001
Time*HS_LAB	-0.1268	0.02379	-5.33	<.0001
Timeknot*HS_LAB	0.04669	0.03225	1.45	0.1477
Control Variables		Included		

Note. Control variables include: gender, races/ethnicities, other previous research experiences, matriculated program, age, parental education level, parental profession.

Table 4-34

General Linear Regression Model with High School Classroom-Based Program Experience ("HS_PROG") as the Focus Independent Variable

	Estimate	Standard Error	*t* Statistic	*p* Value
Intercept	2.7589	0.03498	78.88	<.0001
Time	-0.4010	0.007618	-52.64	<.0001
Timeknot	0.4137	0.01013	40.85	<.0001
HS_PROG	0.03027	0.02651	1.14	0.2536
Time*HS_PROG	-0.06158	0.02415	-2.55	0.0108
Timeknot*HS_PROG	0.1259	0.03269	3.85	0.0001
Control Variables		Included		

Note. Control variables include: gender, races/ethnicities, other previous research experiences, matriculated program, age, parental education level, parental profession.

Table 4-35

General Linear Regression Model with College Laboratory Research Experience ("COLL_LAB") as the Focus Independent Variable

	Estimate	Standard Error	*t* Statistic	*p* Value
Intercept	2.6342	0.03537	74.48	<.0001
Time	-0.3048	0.009425	-32.34	<.0001
Timeknot	0.3768	0.01242	30.33	<.0001
COLL_LAB	0.5543	0.01593	34.79	<.0001
Time*COLL_LAB	-0.2412	0.01461	-16.51	<.0001
Timeknot*COLL_LAB	0.1001	0.01969	5.09	<.0001
Control Variables		Included		

Note. Control variables include: gender, races/ethnicities, other previous research experiences, matriculated program, age, parental education level, parental profession.

Table 4-36a

Covariance Structure Comparison in the Linear Model, the Quadratic Model, and the Spline Model for the Comparison between Students Who Participated in High School Laboratory Research Apprenticeship and Students Who Only Participated in College Laboratory Research Apprenticeship

Covariance Structure	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear Model						
Unstructured	6	115266.0	115311.7	115254.0	N/A	N/A
CS	2	122653.5	122668.7	122649.5	4	7395.5***
Toeplitz	3	121673.2	121696.1	121667.2	3	6413.2***
AR	2	121682.1	121697.3	121678.1	4	6424.1***
Quadratic Model						
Unstructured	6	114272.2	114317.9	114260.2	N/A	N/A
CS	2	122084.0	122099.2	122080.0	4	7819.8***
Toeplitz	3	120886.1	120909.0	120880.1	3	6619.9***
AR	2	120885.3	120900.5	120881.3	4	6621.1***
Spline Model						
Unstructured	6	114269.5	114315.2	114257.5	N/A	N/A
CS	2	122081.2	122096.4	122077.2	4	7811.7***
Toeplitz	3	120883.4	120906.2	120877.4	3	6613.9***
AR	2	120882.5	120897.8	120878.5	4	6613.0***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood; CS = compound symmetry; AR = autoregressive.
[a] Compared to the unstructured covariance pattern.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-36b

Mean Model Comparison for the Comparison between Students Who Participated in High School Laboratory Research Apprenticeship and Students Who Only Participated in College Laboratory Research Apprenticeship

Mean Model	NEP	AIC	BIC	-2LL	Difference in NEP[a]	Difference in -2LL[a]
Linear	15	115199.3	115359.3	115157.3	N/A	N/A
Quadratic	17	114195.3	114370.5	114149.3	2	-1008.0***
Spline	17	114195.3	114370.5	114149.3	2	-1008.0***

Note. NEP = number of estimated parameters; -2LL = negative two log-likelihood.
[a] Compared to the linear mean model.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4-37

General Linear Regression Model of Comparison between Students with High School Laboratory Research Experience and Students with Only College Laboratory Research Experience ("LAB_TIME") as the Focus Independent Variable

	Estimate	Standard Error	*t* Statistic	*p* Value
Intercept	3.2038	0.06805	47.08	<.0001
Time	-0.5360	0.01217	-44.05	<.0001
Timeknot	0.4740	0.01664	28.49	<.0001
LAB_TIME	0.07523	0.02804	2.68	0.0073
Time*LAB_TIME	0.01463	0.02544	0.58	0.5652
Timeknot*LAB_TIME	-0.00645	0.03468	-0.19	0.8525
Control Variables		Included		

Note. Control variables include: gender, races/ethnicities, high school classroom-based program, matriculated program, age, parental education level, parental profession.

Table 4-38

General Linear Regression Model with Matriculated Program ("PROGRAM") as the Focus Independent Variable

	Estimate	Standard Error	*t* Statistic	*p* Value
Intercept	2.7585	0.03491	79.02	<.0001
Time	-0.4087	0.007269	-56.22	<.0001
Timeknot	0.4348	0.009675	44.94	<.0001
PROGRAM	1.5087	0.07828	19.27	<.0001
Time*PROGRAM	0.08761	0.07018	1.25	0.2119
Timeknot*PROGRAM	-0.7616	0.09044	-8.42	<.0001
Control Variables		Included		

Note. Control variables include: gender, races/ethnicities, previous research experiences, age, parental education level, parental profession.

Figure 4-1. Percentage of Age When Registering the MCAT

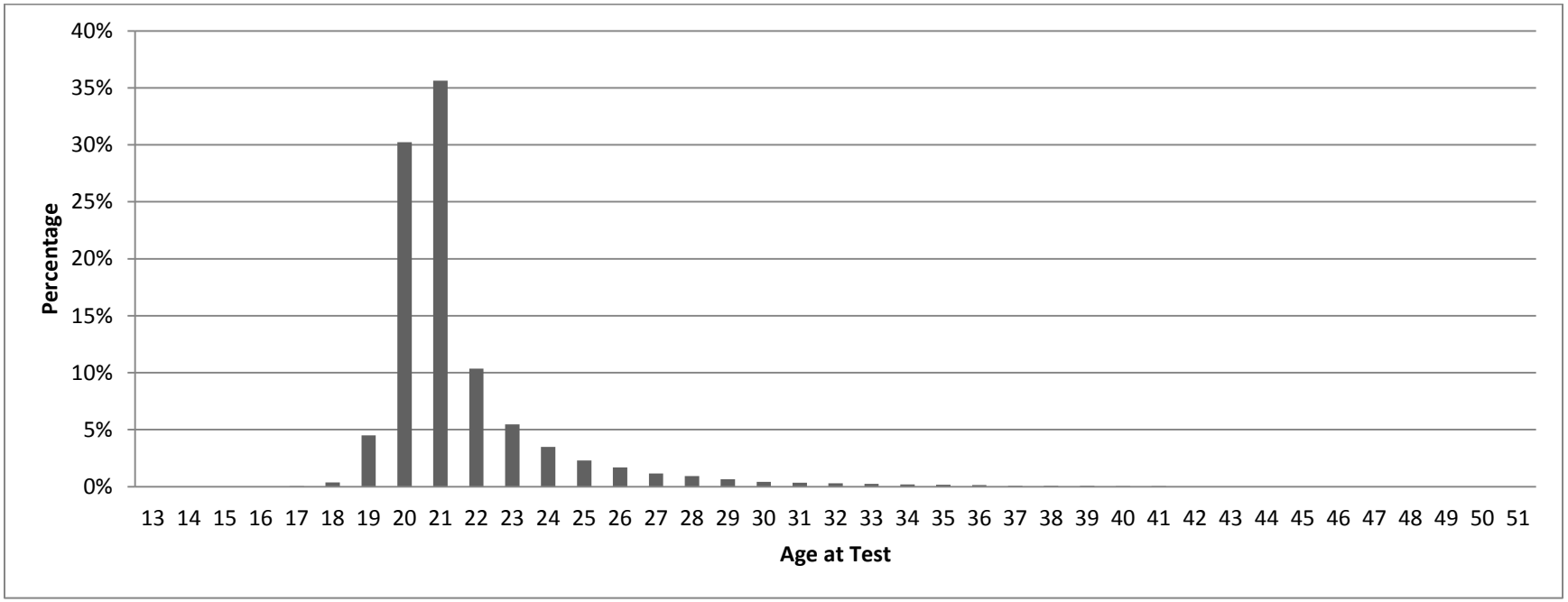


Figure 4-2. Mean Research Interest Levels in General

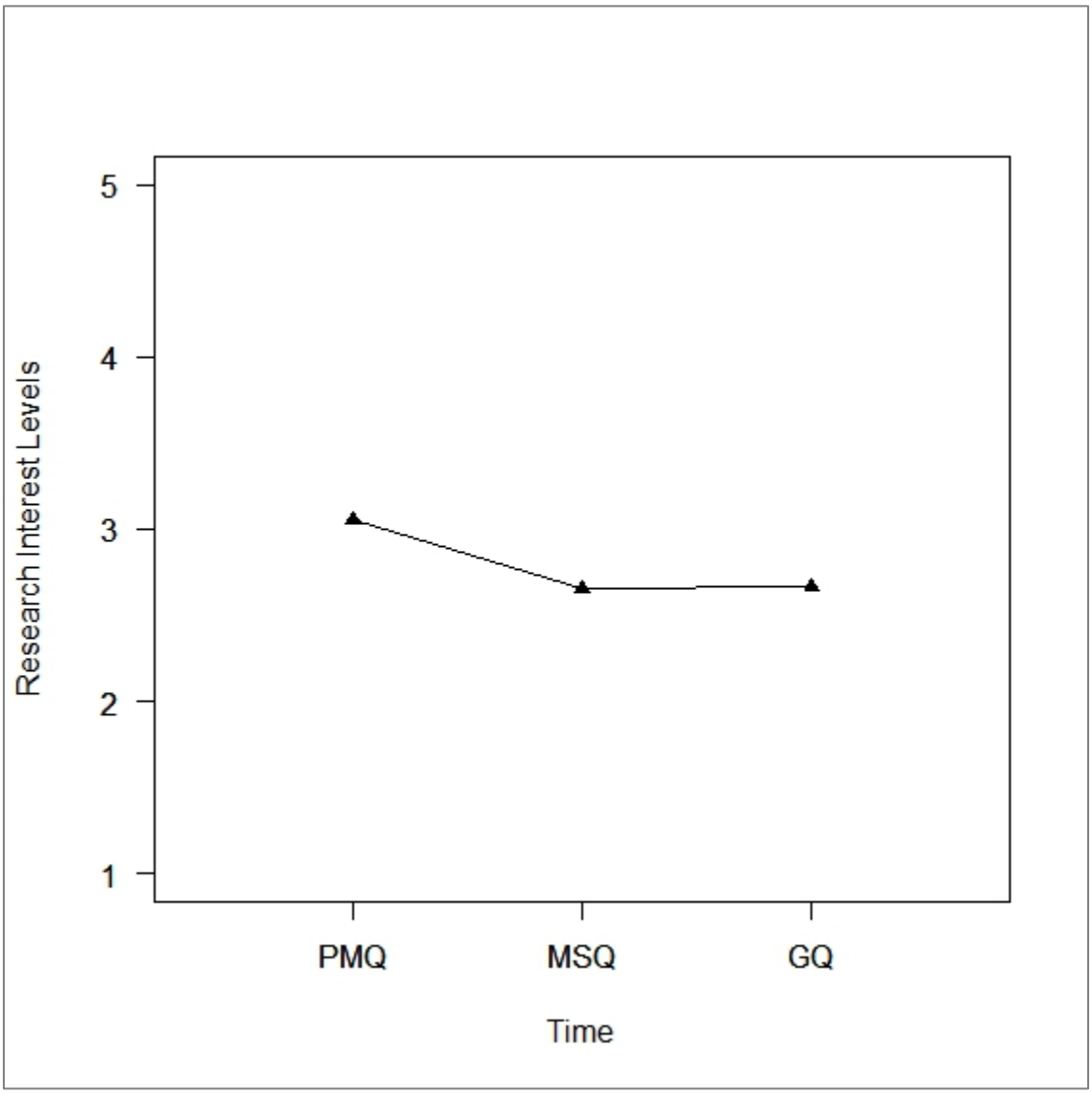


Figure 4-3. Mean Research Interest Levels by Gender

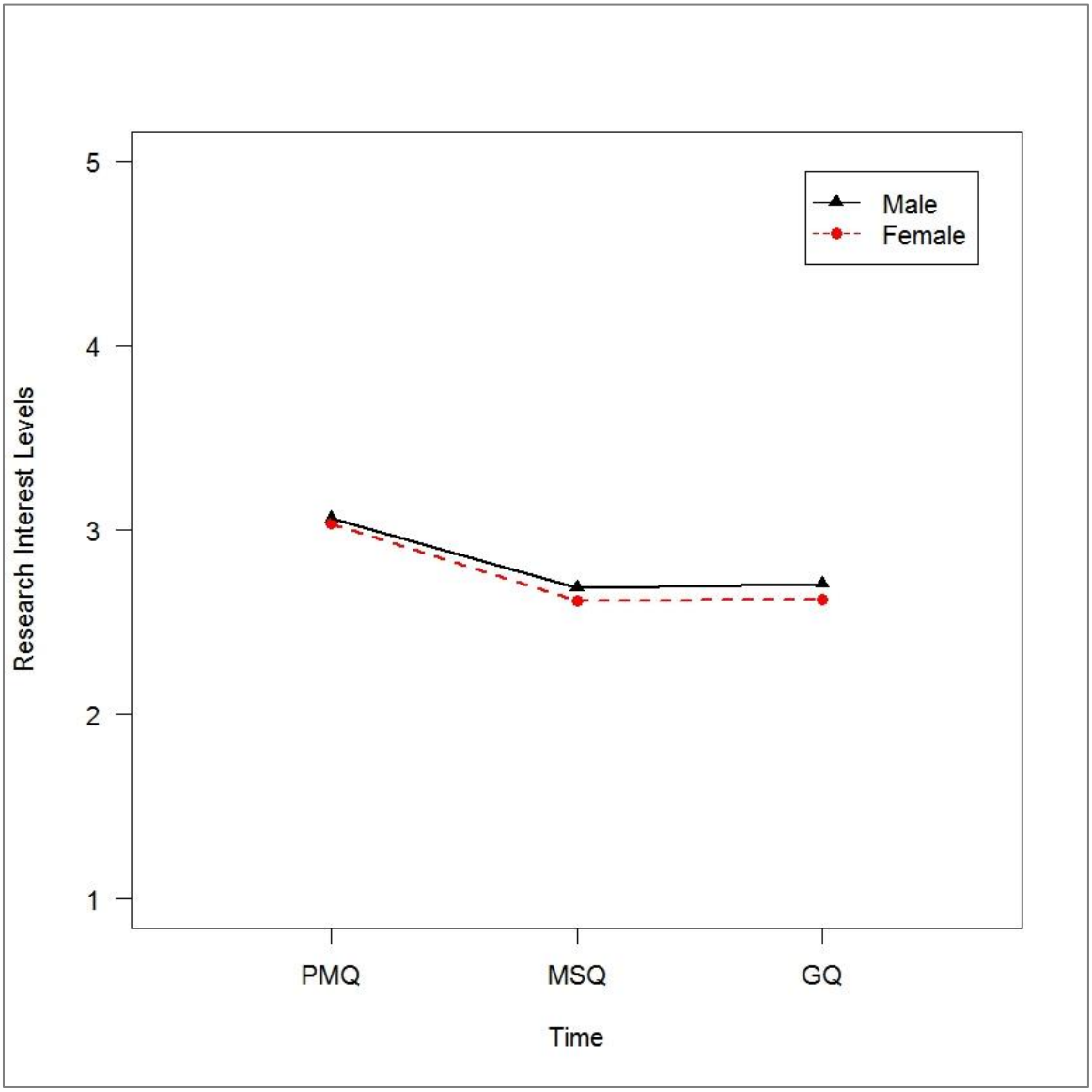


Figure 4-4. Mean Research Interest Levels by Race/Ethnicity

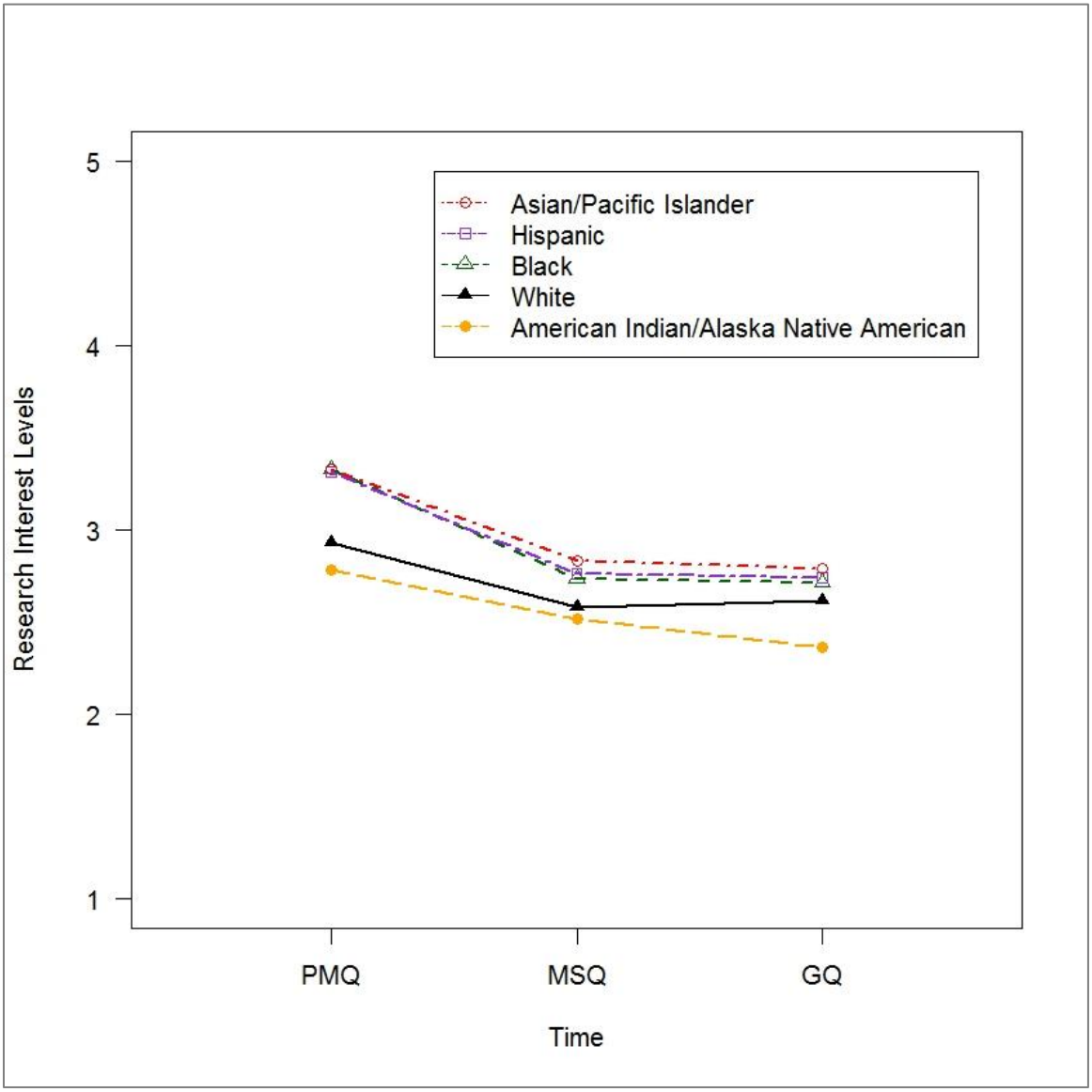


Figure 4-5. Mean Research Interest Levels by High School Laboratory Research Apprenticeship Participation

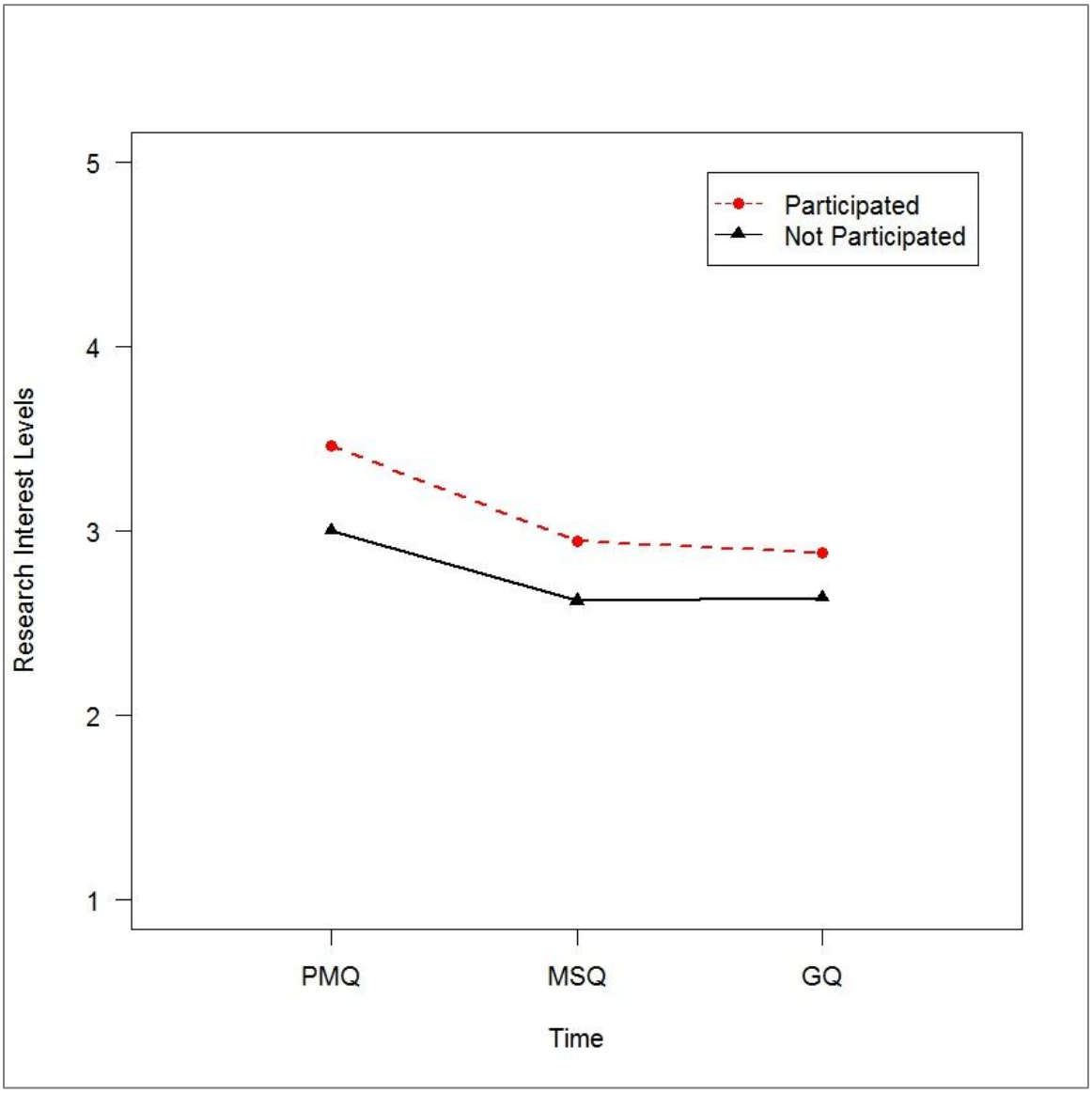


Figure 4-6. Mean Research Interest Levels by High School Classroom-Based Summer, After-School, or Saturday Program Participation

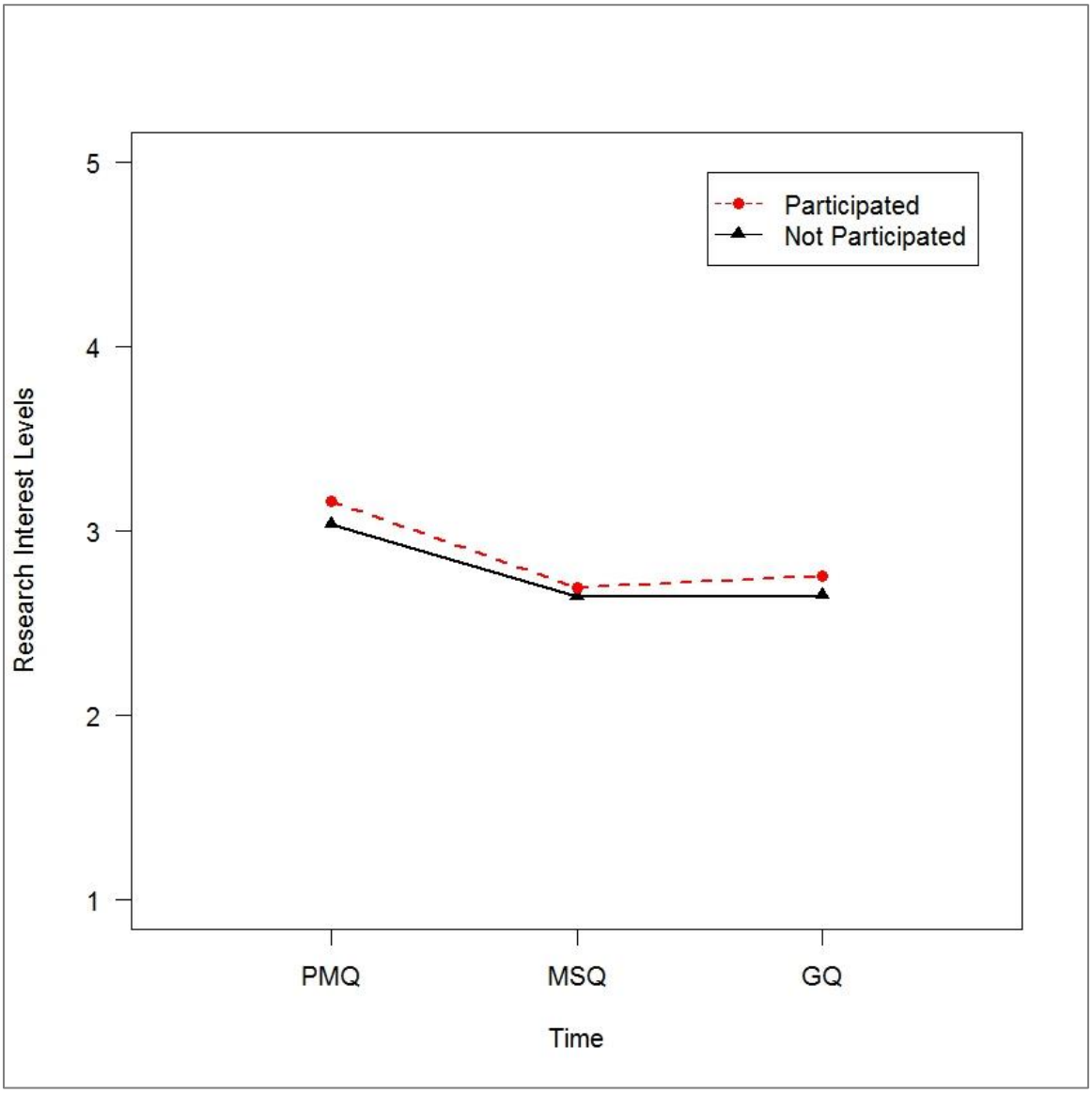


Figure 4-7. Mean Research Interest Levels by College Laboratory Research Apprenticeship Participation

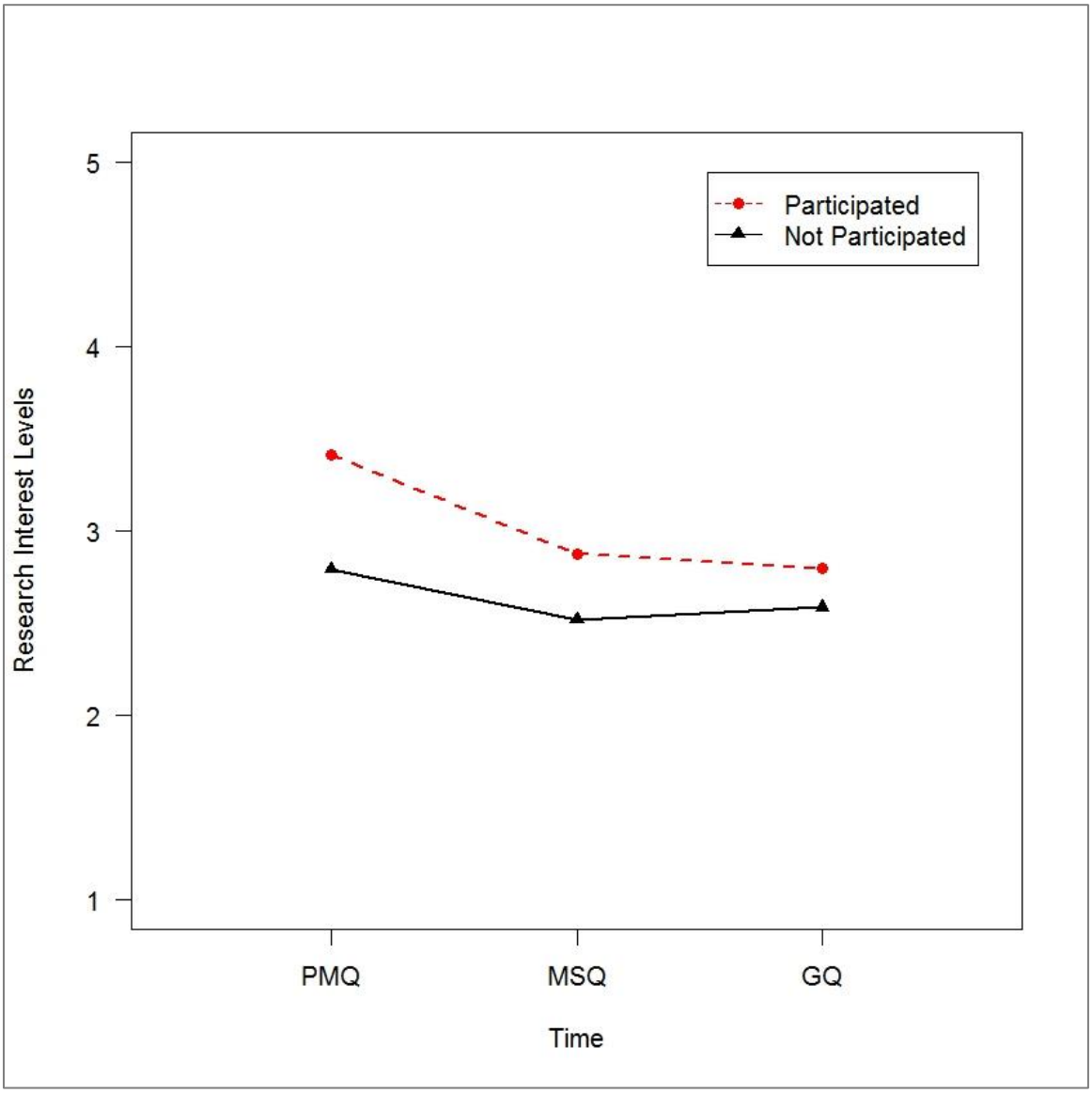


Figure 4-8. Mean Research Interest Levels by Matriculated Program

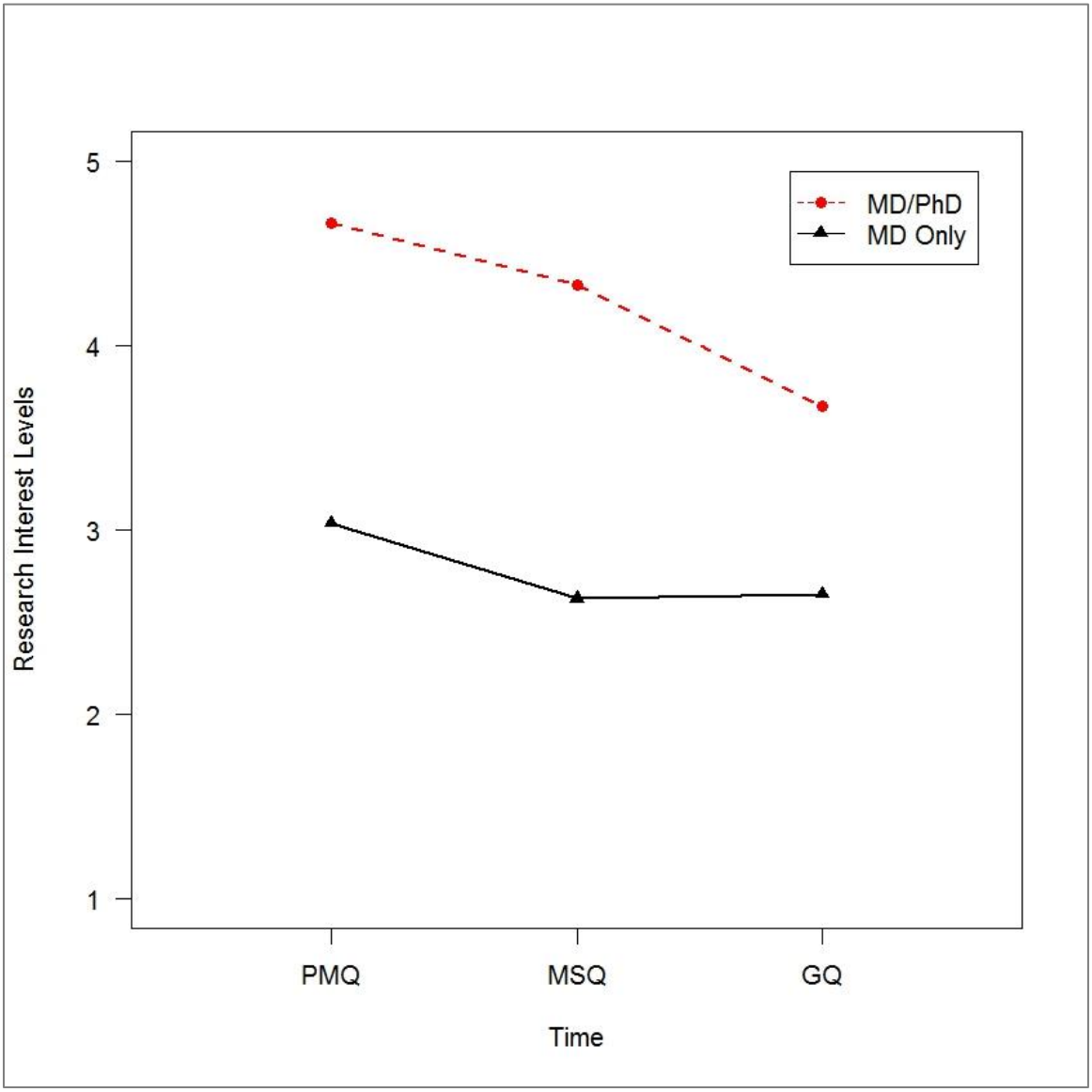
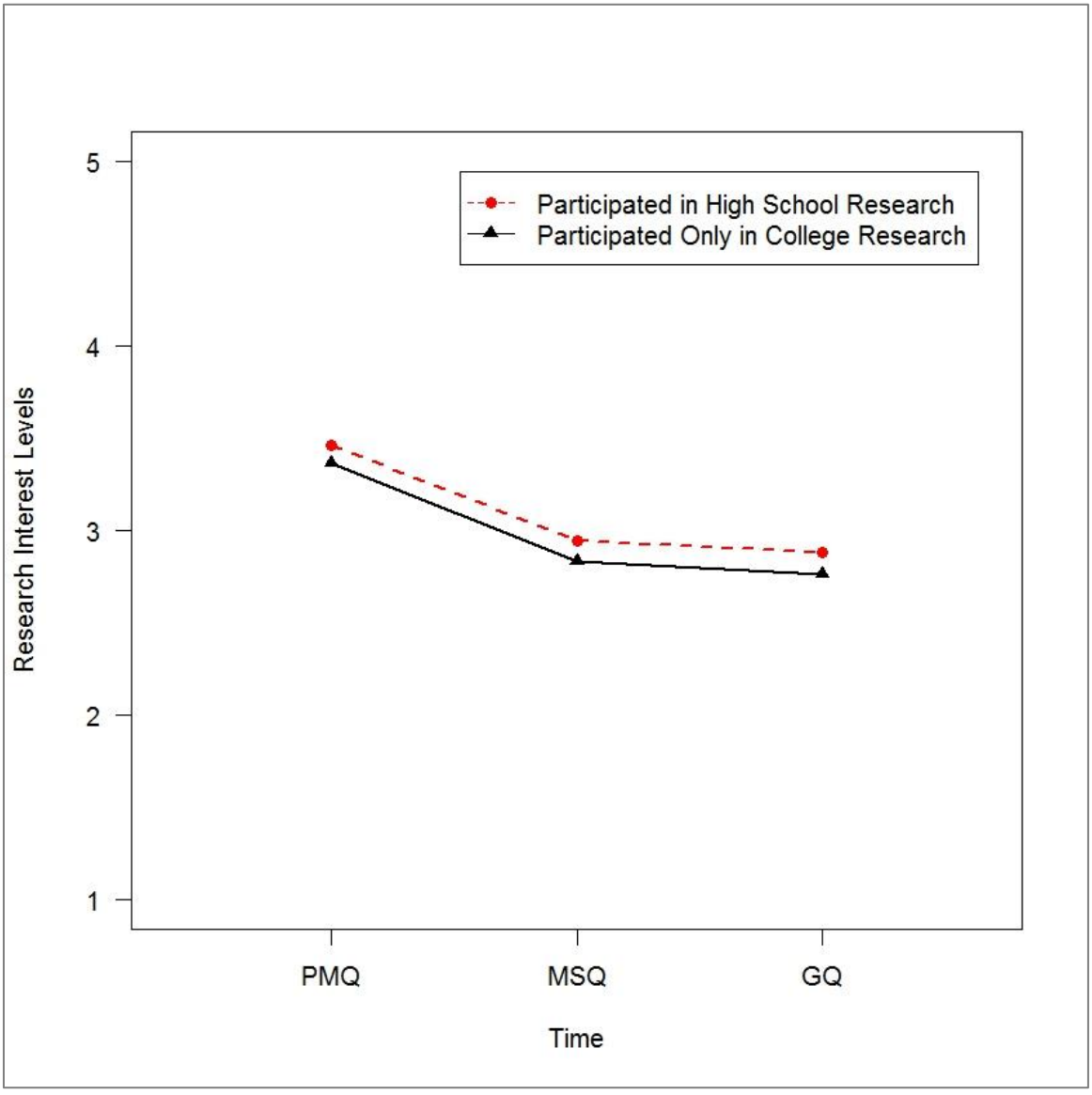


Figure 4-9. Mean Research Interest Levels by Whether Students Participated in High School Research Apprenticeship or Only Participated in College Research Apprenticeship



CHAPTER 5

CONCLUSIONS AND IMPLICATION

STEM education has become a national critical focus with the increasing demand for the STEM workforce in support of innovation and competitiveness (NRC, 2011b; NSB, 2014). Medical education is an important component in STEM education that is considered a valuable source for the biomedical research workforce—physician-scientists (NRC, 2011c). In the biomedical research field, the attrition problem has been a big concern especially among the female and underrepresented racial/ethnic minority groups (Jagsi et al., 2011; Dyrbye et al., 2010; Schafer, 2010). However, many researchers have been examining completion of the degree among students in the general STEM related fields in cross-sectional studies. There is a paucity of research on exploring medical students' persistence in their research interest from a longitudinal perspective.

The overall focus of the study was to examine the trajectory of medical students' research interest across time from prior to their entry to medical schools, to their matriculation in medical schools, and to their graduation from medical schools. To be more specific, the particular research questions addressed in this study sought to investigate: (1) whether medical students' research interest differed among students with different characteristics of gender, race/ethnicity, previous research experiences, or matriculated program prior to their entry to medical schools; (2) whether medical students' research interest changed over time; and (3) whether such patterns of change in

research interest were associated with students' different characteristics. To address the research questions, descriptive analyses and longitudinal data analyses based on general linear regression models were conducted. Descriptive analyses provided an overview of independent, dependent, and control variables, as well as medical students' mean research interest levels by subgroups across time. After appropriate covariance structures and mean models were selected, longitudinal data analyses were conducted so as to address the research questions. Results indicated that (1) students' research interest levels differed among students with different characteristics of gender, race/ethnicity, previous research experiences, or matriculated program prior to their entry to medical schools; that (2) in general, students' research interest levels decreased before matriculation, and then such decrease was offset after matriculation until graduation; and that (3) the patterns of change in the reported research interest over time were different among students with different characteristics.

This study has the potential to fill the gap in the related research pool from four perspectives. First, this study provided a detailed examination on the specific characteristics of the unique population—medical school student population in the United States. Second, this study focused on students' persistence in their research interest from prior to their entry to medical schools to their graduation from medical schools instead of the completion of the program considered as the persistence in that field in most previous literature (e.g., Maher et al., 2004). Third, the data analyzed in this study were from a nationwide database which guaranteed the large sample size of this study. Fourth, and the most importantly, this study shed light on the trajectory of medical students' research interest in a long term. This chapter discusses the findings based on the results from

descriptive analyses and longitudinal data analyses to provide evidence about the overall trend of medical student's research interest, and the factors that were associated with the patterns of change in research interest. In addition, this chapter also illustrates the recommendations based on the research findings of the study followed by limitations of the study and final thoughts.

Descriptive Analysis

The descriptive analyses provided a general picture of the basic information of the medical student sample in this study. The sample information included medical students' demographics, previous research experiences, matriculated program, and their research interest levels over time by different subgroups. It should be noted again that the descriptive analyses did not conclude any inferential evidence, but these analyses did provide an overall description of the variables analyzed in the study.

The sample of the study consisted of 39,839 medical school graduates who completed all the three questionnaires: the PMQ (2001-2006), the MSQ (2001-2007), and the GQ (2005-2011). Although this sample could not exactly correspond with a particular group of medical student population in the United States, the patterns of gender and racial/ethnic composition in the sample of this study were similar to those in the American medical school applicant, matriculant, and graduate population within the corresponding year ranges (AAMC, 2013c). The race/ethnicity distribution in the sample provides additional evidence that Black, Hispanic, and American Indian/Alaska Native American students were underrepresented in medical schools compared to the percentage of those groups in the U.S. population, as discussed and focused in many previous studies on medical school students' race/ethnicity (e.g., Cooper, 2003; Fang et al., 2000). In

terms of previous research experience, fewer students in this sample reported high school research experiences compared to the students who reported college research experiences. The students in this study appeared to have less access or more limited opportunities to be exposed to the research environment in high school than in college. Similarly, there are fewer studies on students' research activities in high school than the studies on students' post-secondary research activities as reviewed in Chapter 2. Another independent variable focused in this study was the degree program at matriculation. Descriptive analysis results indicated that MD students were the majority in medical schools, while MD/PhD students accounted for a very small portion of the medical school students in the sample. Considering the substantial contribution of the MD/PhD graduates to the biomedical research field (Dickler, Fang, Heinig, Johnson, & Korn, 2007; Ley & Rosenberg, 2005), it is necessary and important to retain the research interest of this small but significant group of students.

With regard to the research interest levels over time, it appeared that the research interest reported by the sample in this study decreased from when they registered the MCAT to when they were matriculated in medical schools, but then was retrained during medical schools until they graduated. In addition, it also appeared that different subgroups might present different patterns of change in students' reported research interest over time. For example, Asian/Pacific Islander, Black, and Hispanic groups appeared to differ from the White group, students with previous research experiences appeared to differ from students with no such experiences, and students in MD/PhD programs appeared to differ from students in MD programs. It was also interesting to find

visually that some differences were consistent over time, while some were not (e.g., narrowing).

According to descriptive analyses, the basic information of the sample and the overview of the trajectory of research interest reported by the sample by different subgroups were discussed. The descriptive analyses only provided an initial understanding about the sample and the variables of interest in the study. These analyses were not translated into inferential conclusions, which were reserved for the discussion of the longitudinal data analyses on the reported research interest over time.

Longitudinal Data Analysis

Based on the likelihood ratio test, the spline model was selected in each model with the respective focus independent variable. The result that the spline model was selected implied that medical students' reported research interest levels did not change in a linear pattern over time. Instead, in general, medical students' reported research interest levels decreased significantly from when they registered the MCAT to when they were matriculated in medical schools. Such decrease was then significantly offset after their matriculation in medical schools until their graduation from medical schools. The decrease in students' research interest existed before matriculation, which verifies the implication of the findings from the study by Guelich et al. (2002). However, such decrease was eased to some extent during medical schools. To be more specific, the results showed that students' reported research interest levels did not continuously decrease but instead remained flat during medical schools. It appears that students' experiences during medical schools did not contribute to students' research interest.

Female students' reported research interest was lower at the beginning when students registered the MCAT than male students. In addition, although both female and male students' reported research interest decreased from prior to entry to medical schools to matriculation in medical schools, female students' reported research interest decreased more significantly compared to male students. After matriculation, the research interest levels reported by the two groups were offset in the same degree. It was interesting to observe that male students consistently reported higher research interest than female students over time. Similarly, the previous study conducted by Guelich et al. (2002) indicated that a consistently smaller percentage of women than men reported strong research intentions among the matriculating and graduating medical students between 1987 and 1997. Therefore, this study adds such consistent difference to the previous literature showing that the gender gap in terms of biomedical research interest has still existed over decades. Another study by Ley and Hamilton (2008) suggested that the gender gap became even larger after graduation through post-doctoral training and independent career. While women account for an increasing fraction of medical students who train to become physician-scientists, their attrition rate is disproportionately larger than men. If such trend continues, there will be a serious shortage in the physician-scientists. Regarding the consistent difference, continuous attention should be paid to retaining female students' research interest in medical schools, or even as early as prior to their entry to medical schools.

At the point of registering the MCAT (i.e. prior to entry to medical school), the research interest levels reported by medical students from different racial/ethnic groups were already different. Asian/Pacific Islander, Black, and Hispanic students reported

higher research interest levels than their White peers as early as when they started considering to pursue an advanced degree in medical schools. However, the difference in the research interest levels between Black and White students, and between Hispanic and White students, became smaller as students entered medical schools and graduated from medical schools. The good news is that although the difference became smaller as time went, students from the Black and Hispanic groups still consistently reported higher research interest than White students. However, among the physician-scientist population in the biomedical research field, the Black and Hispanic population is still underrepresented (AAMC, 2013c). Assuming students who reported intending to involve in research in their medical careers finally entered the biomedical research career pathway, then such underrepresentation phenomenon in the biomedical research field might be explained by the underrepresentation of the Black and Hispanic students in the very beginning—at the matriculation in medical schools. Therefore, medical schools are suggested to continue to diversify their enrollees in the admission stage. From another perspective, the result that Black and Hispanic students reported consistently higher research interest than White students until graduation may imply that the lack of representation for Black and Hispanic physician-scientists in the biomedical research field does not appear to be associated with the lack of reported research interest. In this case, cumulative advantage (e.g., systematic barrier) might be involved (DiPrete & Eirich, 2006; Ginther et al., 2011). Future research on why Black and Hispanic groups with higher research interest are less represented throughout the biomedical research pipeline is warranted. In addition, for American Indian/Alaska Native American students, their research interest levels were consistently lower than the White students, and decreased

significantly across time. Although the sample size for this group was relatively small, it could still be indicated at least in the sample of this study that more attention should be paid to this particular underrepresented minority group during medical school.

In this study, three pieces of information related to students' previous research experiences were examined: high school laboratory research apprenticeship, high school classroom-based summer, after-school, or Saturday program participation, as well as college laboratory research apprenticeship. Results indicated that different activities had unique and interesting patterns respectively. Students who participated in high school laboratory research apprenticeship reported higher research interest than students who did not when they started to consider pursuing an advanced degree in medical schools. When it comes to the matriculation time point, although the research interest of students who participated in high school laboratory research apprenticeship decreased more significantly than that of students who did not, the difference in the reported research interest between the two groups were still noticeably large. Then, from matriculation to graduation, the changes of the change rates were not significantly different between the two groups, indicating that the difference was consistent. That is to say, students who participated in high school laboratory research apprenticeship reported consistently higher research interest than students who did not. In other words, participating in high school laboratory research apprenticeship was longitudinally associated with higher reported research interest. Therefore, the results in this study suggested that students should be exposed to research related environment as early as in high school, since such experience is found to have a long-term association with students' consistently higher interest in the research activities.

Though still belonging to high school experiences, it appeared that high school various classroom-based programs did not have much significance in terms of maintaining or increasing students' research interest. The research interest of students who participated in high school classroom-based program decreased more before matriculation, but increased more after matriculation, compared to that of students who did not. Meanwhile, the research interest levels of the two groups were not different prior to entry to medical schools. It should be noticed that the survey questions did not specify the format or content of the high school programs, and thus the results in the comparison between the two groups may come as no surprise. Since there was no specific activities or programs mentioned in the questionnaire, the patterns of change in students' research interest might not be explained in a detailed way.

Similar to the pattern of change in high school laboratory research apprenticeship participation results, the college laboratory research apprenticeship participation results indicated that students who participated in college laboratory research apprenticeship reported significantly higher research interest than students who did not at the beginning when they registered the MCAT. However, unlike the patterns in high school lab results, the difference in the research interest levels between students who participated in in college laboratory research apprenticeship and students who did not became smaller over time from prior to entry to medical schools, to matriculation in medical schools, and graduation from medical schools. It could be indicated that although participating in college laboratory research apprenticeship was associated with higher research interest, such relationship might not be as strong as participating in high school laboratory research apprenticeship in the long term. In addition, previous research indicated that

exposure to more research in well-designed programs can also help motivate current medical school students (Solomon et al., 2003).

Further analyses were conducted on the comparison between students with high school research experiences and students with no high school research experiences but with college research experiences. Results showed that students with high school research experiences reported significantly higher research interest prior to their entry to medical schools than students with only college research experiences. In addition, such research interest difference between the two groups was very consistent across time until their graduation from medical schools. It should still be encouraged to get students involved in research during college, since research experience has a long-term positive association with students' interest in research in medical schools. Meanwhile, it can also be implied that the earlier students had research experiences, the higher long-term research interest they might maintain.

Through a comparison between students from the MD/PhD programs and students from the MD programs, results indicated that MD/PhD students reported significantly much higher research interest than MD students when they started considering to pursue a degree in medical schools, which was in line with the findings from the study conducted by McGee and Keller (2007). The results match the original purpose of founding the MD/PhD program, which was to particularly train potential physician scientists for the biomedical research field (Rosenberg, 2008). The difference remained the same until students' were matriculated in medical schools. However, the change in the reported research interest from matriculation in medical schools to graduation from medical schools among MD/PhD students was significantly different from that among

MD students. To be more specific, MD/PhD students' reported research interest significantly dropped in an alarming rate, while MD-only students' reported research interest was retained. At the graduation point, though, MD/PhD students still indicated higher research interest than MD students. The results about the change in students' reported research interest during medical schools suggested an urgent improvement especially in MD/PhD programs. As discussed previously, the MD/PhD programs were developed with the purpose to deliver more experienced and expert physician-scientists to the biomedical research field. However, the results from this study showed that through training in the MD/PhD program for years, students seemed to become much less interested in doing biomedical research than they did when they just entered the program. In other words, the MD/PhD programs appeared to turn their students who initially chose to do research away from the biomedical research pipeline. This argument might partially explain the phenomenon that some MD/PhD students at the time of matriculation later left the PhD program in the middle of their academic pursuit (Andriole et al., 2008; McGee & Keller, 2007). Therefore, MD/PhD programs in medical schools should put more focus on how to retain their program enrollees' interest in the biomedical research. Otherwise, even if some students in the MD/PhD programs stayed and completed their degrees, these graduates might not stay in the biomedical research pipeline after graduation, which is still a big loss for the biomedical research workforce.

Recommendations from the Study

Based on the descriptive analyses and longitudinal data analyses, this study provides insights about the trajectory of medical students' reported research interest over time among students with different characteristics of gender, race/ethnicity, previous

research experiences, and matriculated program. Several recommendations for medical education are developed from these insights.

Recommendation 1: Continued focus on retaining the biomedical research interest among the female population during medical school is necessary and important.

Recommendation 2: It is essential to investigate, understand, and improve other factors that are associated with the underrepresentation of the Blacks and Hispanics throughout the biomedical research pipeline (i.e. from starting to consider medicine as a career field to finally working as a biomedical scientist) in order to diversify the biomedical research workforce. It seems that the underrepresentation is not related to lack of interest in biomedical research.

Recommendation 3: Increased attention and resources should be devoted to laboratory research exposure and experiences in high school and college. The earlier students are exposed to the research environment, the higher their interest in research may be retained in the long term.

Recommendation 4: Effective and appropriate interventions in the MD/PhD programs are in urgent demand to make MD/PhD enrollees retain their initial interest in the biomedical research over time.

These recommendations stem from the discussions of a series of descriptive and statistical analyses in this study. The recommendations provide informative suggestions about when and how to retain medical students' research interest, and meanwhile point out certain perspectives that are worth considering and examining in future studies. Especially, more longitudinal analyses on medical students' retrospective opinions and

experiences are encouraged to continuously explore the trajectory of physician-scientist academic and career paths in the biomedical research field.

Limitations of the Study

This study has several limitations which can be viewed from two perspectives: study sample and survey instrument. Concerning the study sample, first, it should be noted again that the sample students analyzed in this study registered the MCAT exam between 2001 and 2006, were matriculated in medical schools between 2001 and 2007, and finally graduated from medical schools between 2005 and 2011. Meanwhile, the sample students were also those who completed all the three questionnaires (PMQ, MSQ, and GQ) along the time. Though not nationally representative, this large group of students contains all the individuals in the record system in terms of providing their retrospective experiences and opinions within certain year ranges, based on which this study was developed. Second, due to the limited participation in the three questionnaire completion, so far there is no way to track the dropouts—individuals who left the medical field in the middle of their academic pursuits. It is evident that these individuals did not complete the programs, but it is still unclear whether and how these individuals' research interest changed thereafter. Therefore, this limitation may lead to further study about the long-term experiences of the dropout group. Third, the students analyzed in this study all pursued either an MD degree or combination of MD degree and another degree. The students who obtained a PhD degree separately from their MD programs (either before or after the MD programs) were not captured in this analysis.

In addition, there are limitations related to the survey instrument in this study. First, there was no exactly the same question asking for individuals' research interest

over time in the three questionnaires. In a longitudinal study, it would be ideal to have the same measure over time. But unfortunately, based on the data analyzed in this study, the phrasing about research interest changed from questionnaire to questionnaire. Although it was understandable that individuals were asked about their opinions and expectations in customized and different ways along the time, it was still considered a limitation for a longitudinal data analysis design. Second, the questionnaires were designed to ask for as much information as possible from the individuals, but there are always more factors related than collected in social science. In this study, although many factors were included in the models, there might be more related factors that were not captured in the questionnaires. Third, survey instrument always collects reported data. Therefore, it should always be noted that the data and the results may only indicate the students' reported opinions and experiences. There is no guarantee that the survey instrument was able to collect the actual opinions and experiences from the individuals who completed the survey (Rosenberg, 2008).

Final Thoughts

Overall, longitudinal data analysis results indicate that medical students' reported research interest levels change significantly over time following a non-linear trend. The patterns of change in students' reported research interest are significantly associated with students' characteristics of gender, race/ethnicity, previous research experiences, and matriculated program. This study focuses on the "whether" questions, which may lead to further examination of the "why" questions in the future. Why do the students' reported research interest levels significantly decrease in general from when they register the MCAT to when they are matriculated in medical schools? Why do females consistently

report lower research interest than males? Why are Black and Hispanic students who have higher reported research interest still underrepresented throughout the biomedical research pipeline? Why do MD/PhD students' reported research interest levels decrease dramatically during medical schools? These essential questions are worth further investigating in future studies, since a better understanding of these questions may help retain and increase students' research interest over time even when the students start considering to pursue an advanced degree in medical schools. Additionally, this study provides important evidence on the trajectory of medical students' research interest in a longitudinal perspective. In the future, more longitudinal studies on medical school students are encouraged to investigate the trends in their records, opinions and experiences over time so that corresponding strategies can be recommended to improve such trends.

REFERENCES

Ahn, J., Watt, C. D., Greeley, S. A., & Bernstein, J. (2004). MD-PhD students in a major training program show strong interest in becoming surgeon-scientists. *Clinical Orthopaedics and Related Research, 425*, 258-263.

Ahn, J., Watt, C. D., Man, L. X., Greeley, S. A., & Shea, J. A. (2007). Educating future leaders of medical research: Analysis of student opinions and goals from the MD-PhD SAGE (students' attitudes, goals, and education) survey. *Academic Medicine, 82*(7), 633-645.

Alexander, J. M., Johnson, K. E., & Kelley, K. (2012). Longitudinal analysis of the relations between opportunities to learn about science and the development of interests Related to science. *Science Education, 96*(5), 763-786.

Anderson, E., & Kim, D. (2006). *Increasing the success of minority students in science and technology*. Washington, DC: American Council on Education.

Andrews, N. C. (2002). The other physician-scientist problem: Where have all the young girls gone? *Nature Medicine, 8*(5), 439-441.

Andriole, D. A., Jeffe, D. B. (2012). The road to an academic medicine career: A national cohort study of male and female U.S. medical graduates. *Academic Medicine, 87*(12), 1722-1733.

Andriole, D. A., Whelan, A. J., & Jeffe, D. B. (2008). Characteristics and career intentions of the emerging MD-PhD workforce. *Journal of the American Medical Association (JAMA), 300(10),* 1165-1173.

Armstrong, P., & Vogel, D. L. (2009). Interpreting the interest-efficacy association from a RIASEC perspective. *Journal of Counseling Psychology*, *56*(3), 392-407.

Ashworth, J., & Evans, L. J. (2001). Modeling student subject choice at secondary and tertiary level: A cross-section study. *Journal of Economic Education*, *32*(4), 311–322.

Association of American Medical Colleges. (2010). *Matriculating student questionnaire: 2010 all schools summary report.* Washington, DC: Author.

Association of American Medical Colleges. (2011a). *Matriculating student questionnaire: 2011 all schools summary report.* Washington, DC: Author.

Association of American Medical Colleges. (2011b). *Medical school graduation questionnaire: 2011 all schools summary report.* Washington, DC: Author.

Association of American Medical Colleges. (2012a). *Matriculating student questionnaire: 2012 all schools summary report.* Washington, DC: Author.

Association of American Medical Colleges. (2012b). *Medical school graduation questionnaire: 2012 all schools summary report.* Washington, DC: Author.

Association of American Medical Colleges. (2013a). *Unpublished report.* Washington, DC: Author.

Association of American Medical Colleges. (2013b). *Medical school graduation questionnaire: 2013 all schools summary report.* Washington, DC: Author.

Association of American Medical Colleges. (2013c). *AAMC Data Book. (Unpublished report).* Washington, DC: Author.

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

Berkes, E. (2008). Undergraduate research participation at the University of California, Berkeley. Research and Occasional Paper Series CSHE.17.08. Center for Studies in Higher Education, University of California, Berkeley.

Bleeker, M., & Jacobs, J. (2004). Achievement in math and science: Do mothers' beliefs matter 12 years later? *Journal of Educational Psychology, 96,* 97-109.

Byars-Winston, A. M., & Fouad, N. A. (2008). Math and science social cognitive variables in college students: Contributions of contextual factors in predicting goals. *Journal of Career Assessment, 16*(4), 425-440.

Canes, B., & Rosen, H. (1995). Following in her footsteps? Faculty gender composition and women's choices of college majors. *Industrial and Labor Relations Review, 48*(3), 486-504.

Carell, S. M., Page, M., & West, P. J. (2010). Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics, 125*(3), 1101-1144.

Carter, F. D., Mandell, M., & Maton K. I. (2009). The influence of on-campus, academic year undergraduate research on STEM Ph.D. outcomes: Evidence from the Meyerhoff Scholarship Program. *Educational Evaluation and Policy Analysis, 31*(4), 441-462.

Cech, T. R., Egan, L. W., Doyle, C., Gallin, E., Lichtman, M. A., Queenan, C. J., III, & Sung, N. S. (2001). The biomedical research bottleneck. *Science, 293*(5530), 573.

Clark, J. M., & Hanel, D. P. (2001). The contribution of MD-PhD training to academic orthopaedic faculties. *Journal of Orthopaedic Research, 19,* 505-510.k

Cole, D., & Espinoza, A. (2008). Examining the academic success of Latino students in science technology engineering and mathematics (STEM) majors. *Journal of College Student Development, 49*(4), 285-300.

Cooper, R. A. (2003). Impact of trends in primary, secondary, and postsecondary education on application to medical school. II: Considerations of race, ethnicity, and income. *Academic Medicine, 78,* 864-876.

Congressional Research Service. (2012). Science, technology, engineering, and mathematics (STEM) education: A primer. *CRS Report for Congress*.

Cregler L. L. (1993). Enrichment programs to create a pipeline to biomedical science careers. *Journal of the Association for Academic Minority, 4*(4), 127-131.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation in education: The self-determination perspective. *Educational Psychology, 26,* 325–46.

DeValero, Y.F. (2001). Departmental factors affecting time-to-degree and completion rates of doctoral students at one land-grant research institution. *The Journal of Higher Education, 72*(3), 341-367.

Dickler, H. B., Fang, D., Heinig, S. J., Johnson, E., & Korn, D. (2007). New physician-investigators receiving National Institutes of Health research project grants: A historical perspective on the "endangered species". *Journal of the American Medical Association (JAMA), 297*(22), 2496-2501.

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology, 32*, 271-297.

Dyrbye, L. N., Thomas, M. R., Power, D. V., Durning, S., Moutier, C., Massie Jr., S., … Shanafelt, T. D. (2010). Burnout and serious thoughts of dropping out of medical school: A multi-institutional study. *Academic Medicine, 85*(1), 94-102.

Ethington, C. A., & Smart, J. C. (1986). Persistence to graduate education. *Research in Higher Education, 24*, 287-303.

Fang, D., Moy, E., Colburn, L., & Hurley, J. (2000). Racial and ethnic disparities in faculty promotion in academic medicine. *Journal of the American Medical Association (JAMA), 284*, 1085-1092.

Featherman, D. L., and R. M. Hauser. (1978). *Opportunity and change.* New York: Academic Press.

Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychol Methods, 14*(1), 43-53.

Fisher, R., & Engemann, J. (2009). *Factors affecting attrition at a Canadian college*. Vancouver, BC: Canadian Council on Learning.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Fouad, N. A., Smith, P. L., & Zao, K. E. (2002). Across academic domains: Extensions of the social-cognitive career model. *Journal of Counseling Psychology, 49*(2), 164-171.

Fried, T., & MacCleave, A. (2009). Influence of role models and mentors on female graduate students' choice of science as a career. *Alberta Journal of Educational Research, 55*(4), 482-496.

Gallin, E. K., Le Blancq, S. M., & Clinical Research Fellowship Program Leaders. (2005). Launching a new fellowship for medical students: The first years of the Doris Duke Clinical Research Fellowship Program. *Journal of Investigative Medicine, 53,* 73-81.

Gardner, S. K. (2009). Student and faculty attributions of attrition in high and low-completing doctoral programs in the United States. *Higher Education, 58,* 97-112.

Garrison, G., Mikesell, C., & Matthew, D. (2007). Medical school graduation and attrition rates. *Analysis in Brief, 7*(2). Retrieved from https://www.aamc.org/download/102346/data/aibvol7no2.pdf

Garrison, H. H., & Deschamps, A. M. (2013). NIH research funding and early career physician scientists: Continuing challenges in the 21st century. *The Journal of Federation of American Societies for Experimental Biology*. doi:10.1096/fj.13-241687

George, R., & Kaplan, D. (1998). A structural model of parent and teacher influences on science attitudes of eighth graders: Evidence from NELS: 88. *Science Education, 82*(1), 93-109.

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, ethnicity, and NIH research awards. *Science, 333*(6045), 1015-1019.

Golde, C. M. (2002). Beginning graduate school: Explaining first-year doctoral attrition. *New Directions for Higher Education, 1998*(101), 55-64.

Grandy, J. (1998). Persistence in science of high-ability minority students: Results of a longitudinal study. *The Journal of Higher Education, 69*(6), 589-620.

Griffith, A. L. (2010). Persistence of women and minorities in STEM field majors: Is it the school that matters? *Economics of Education Review, 29*, 911-922.

Guelich, J. M., Singer, B. H., Castro, M. C., & Rosenberg, L. E. (2002). A gender gap in the next generation of physician-scientists: Medical student interest and participation in research. *Journal of Investigative Medicine, 50*(6), 412-418.

Hathaway, R., Nagda, B., & Gregerman, S. (2002). The relationship of undergraduate research participation to graduate and professional educational pursuit: An empirical study. *Journal of College Student Development, 43*(5), 614 – 631.

Herrera, F. A., & Hurtado, S. (2011, April). *Maintaining initial interests: Developing science, technology, engineering, and mathematics (STEM) career aspirations among underrepresented racial minority students.* Paper presented at the Association for Educational Research annual meeting, New Orleans, LA. Retrieved from http://www.heri.ucla.edu/nih/downloads/AERA%202011%20-%20Herrera%20and%20Hurtado%20-%20Maintaining%20Initial%20Interests.pdf

Hiatt, H., & Sutton, J. (2000). The nation's changing needs for biomedical and behavioral scientists. *Academic Medicine, 75*, 778-779.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the twenty-first century. *Review of Educational Research, 70*(2), 151-179.

Higher Education Research Institute. (2010). *Degrees of success: Bachelor's degree completion rates among initial STEM majors.* Higher Education Research Institute at UCLA, Los Angeles.

Hoffer, T. B., Welsh, V., Webber, K., Williams, K., Lisek, B., Hess, M., … Guzman-Barron, I. (2006). *Doctorate recipients from United States universities: Summary report 2005.* Chicago: National Opinion Research Center.

Holland, J. H. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6*(1), 35-45.

Holland, J. H. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Hossler, D., Schmit, J. L., & Vesper, N. (1999). *Going to college: How social, economic, and educational factors influence the decisions students make.* Baltimore, MD: Johns Hopkins University Press.

Huang, G., Taddese, N., & Walter, E. (2000). *Entry and persistence of women and minorities in college science and engineering education.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Hunter, A., Laursen, S. L., & Seymour, E. (2006). Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. *Science Education, 91*(1), 36-74.

Hurtado, S., Han, J. C., Saenz, V. B., Espinosa, L. L., Cabrera, N. L., & Cerna, O. S. (2007). Predicting transition and adjustment to college: Biomedical and behavioral science aspirants' and minority students' first year of college. *Reviews of Higher Education, 48,* 841-887.

Jagsi, R., DeCastro, R., Griffith, K. A., Rangarajan, S., Churchill, C., Stewart, A., & Ubel, P. A. (2011). Similarities and differences in the career trajectories of male and female career development award recipients. *Academic Medicine, 86,* 1415-1421.

Jeffe, D. B., Yan, Y., Andriole, D. A. (2012). Do Research Activities During College, Medical School, and Residency Mediate Racial/Ethnic Disparities in Full-Time Faculty Appointments at U.S. Medical Schools? *Academic Medicine, 87*(11),1582-1593.

Kaushansky, K. (2003). Physician-scientists: Preparation, opportunities, and national need. *Experimental Biology and Medicine, 228,* 1258-1260.

Kotchen, T. A., Lindquist, T., Malik, K., & Ehrenfeld, E. (2004). NIH peer review of grant applications for clinical research. *Journal of the American Medical Association, 291*(7), 836-843.

Lent, R. W., Brown, S. D., & Gore, P. A. (1997). Discriminant and predictive validity of academic self-concept, academic self-efficacy, and mathematics-specific self-efficacy. *Journal of Counseling Psychology, 44,* 307-315.

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.

Lent, R. W., Brown, S. D., Schmidt, J., Brenner, B., Lyons, H., & Treistman, D. (2003). Relation of contextual supports and barriers to choice behavior in engineering majors: Test of alternative social cognitive models. *Journal of Counseling Psychology, 50*(4), 458-465.

Lent, R. W., Brown, S. D., Sheu, H., Schmidt, J., Brenner, B., Gloster, C. S.,… Treistman, D. (2005). Social cognitive predictors of academic interests and goals in engineering: Utility for women and students at historically Black universities. *Journal of Counseling Psychology, 52*(1), 84-92.

Ley, T. J., & Hamilton, B. H. (2008). The gender gap in NIH grant applications. *Science, 322,* 1472-1474.

Ley, T. J., & Rosenberg, L., E. (2005). The physician-scientist career pipeline in 2005: Build it, and they will come. *Journal of American Medical Association, 294*(11), 1343-1351.

Lloyd, T., Phillips, B. R., & Aber, R. C. (2004). Factors that influence doctors' participation in clinical research. *Medical Education, 38*, 848-851.

Lopatto, D. (2004). Survey of undergraduate research experiences (SURE): First findings. *Cell Biology Education, 3*, 270-277.

Lopatto, D. (2007). Undergraduate research experiences support science career decisions and active learning. *CBE—Life Science Education, 6*, 297-306.

Lovitts, B. E., Nelson, C. (2000). The Hidden Crisis in Graduate Education: Attrition from Ph.D. Programs. *Academe, 86*(6), 44-50.

Luzzo, D., Hasper, P., Albert, K. A., Bibby, M. A., & Martinelli, E. A. (1999). Effects of self-efficacy-enhancing interventions on the math/science self-efficacy and career interests, goals, and actions of career undecided college students. *Journal of Counseling Psychology*, *46*(2), 233-43.

Maher, M. A., Ford, M. E., & Thompson, C. M. (2004). Degree progress of women doctoral students: Factors that constrain, facilitate, and differentiate. *The Review of Higher Education, 27*(3), 385-408.

Manson, S. M. (2009). Personal journeys, professional paths: Persistence in navigating the crossroads of a research career. *American Journal of Public Health, 99*(S1), S20-S25.

Mare, R. D. (1980). Social background and school continuation decisions. *Journal of the American Statistical Association, 75*(370), 295-305.

Maton, K. I., & Hrabowski, F. A. (2004). Increasing the number of African American PhDs in the sciences and engineering. *American Psychologist, 59*(6), 547-556.

Maton, K. I., Sto Domingo, M. R., Stolle-McAllister, K. E., Zimmerman, J. L., & Hrabowski, F. A. (2009). Enhancing the number of African Americans who pursue STEM PhDs: Meyerhoff Scholarship Program outcomes, processes, and individual predictors. *Journal of Women and Minorities in Science and Engineering, 15*(1), 15-37.

McGee, R., & Keller, J. L. (2007). Identifying future scientists: Predicting persistence into research training. *CBE—Life Sciences Education, 6*, 316-331.

McManus, I.C., Livingston, G., & Katona, C. (2006). The attractions of medicine: The generic motivations of medical school applicants in relation to demography, personality and achievement. *BMC Medical Education*, 6-11.

Moen, P. (1989). Women as a human resource in science and engineering. Paper prepared for the National Science Foundation, Task Force on Women in Science and Engineering.

Mullen, A. L., Goyette, K. A., & Soares, J. A. (2003). Who goes to graduate school? Social and academic correlates of educational continuation after college. *Sociology of Education, 76*(2), 143-169.

Museus, S. D., Palmer, R. T., Davis, R. J., & Maramba, D. C. (2011). *Racial and ethnic minority students' success in STEM education*. San Francisco, CA: Wiley.

Nathan, D. G., & Wilson, J. D. (2003). Clinical research and the NIH—A report card. *New England Journal of Medicine, 349,* 1860-1865.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter future.* Washington, DC: The National Academies Press.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2010). *Rising above the gathering storm, revisited: Rapidly approaching category 5.* Washington, DC: The National Academies Press.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2011). *Expanding underrepresented minority participation: America's science and technology talent at the crossroads.* Washington, DC: The National Academies Press.

National Research Council. (2004). *Critical perspectives on racial and ethnic differences in health in late life.* Washington, D.C.: The National Academies Press.

National Research Council. (2011a). *Research training in the biomedical, behavioral, and clinical research sciences*. Washington, D.C.: The National Academies Press.

National Research Council. (2011b). *Successful K-12 STEM education*. Washington, D.C.: The National Academies Press.

National Research Council. (2011c). *Research-doctorate programs in the biomedical sciences: Selected findings from the NRC assessment*. Washington, D.C.: The National Academies Press.

National Science Board. (2010). *Science and Engineering Indicators 2010.* Arlington VA: National Science Foundation (NSB 10-01). Retrieved from http://www.nsf.gov/statistics/seind10/pdf/seind10.pdf

National Science Board. (2012). *Science and Engineering Indicators 2012.* Arlington VA: National Science Foundation (NSB 12-01). Retrieved from http://www.nsf.gov/statistics/seind12/pdf/seind12.pdf

National Science Board. (2014). *Science and Engineering Indicators 2014*. Arlington VA: National Science Foundation (NSB 14-01). Retrieved from http://www.nsf.gov/statistics/seind14/content/etc/nsb1401.pdf

National Science Foundation. (2012). *Doctorate recipients from U.S. universities: 2011*. Arlington, VA: Author. Retrieved from http://www.nsf.gov/statistics/sed/2011/

National Science Foundation, & Division of Science Resources Statistics. (2011). Women, Minorities, and Persons with Disabilities in Science and Engineering: 2011. (Special Report NSF 11-309). Arlington, VA: Author.

Nauta, M. M., & Epperson, D. L. (2003). A Longitudinal Examination of the Social-Cognitive Model Applied to High School Girls' Choices of Nontraditional College Majors and Aspirations. *Journal of Counseling Psychology, 50*(4), 448-457.

Neilson, E. G. (2003). The role of medical school admissions committees in the decline of physician-scientists. *Journal of Clinical Investigation, 111*, 765-767.

Newton, D. A., & Grayson, M. S. (2003). Trends in career choice by US medical school graduates. *Journal of the American Medical Association (JAMA), 290*(9), 1179-1182.

Odom, K. L., Roberts, L. M., Johnson, R. L., Cooper, L. A. (2007). Exploring obstacles to and opportunities for professional success among ethnic minority medical students. *Academic Medicine, 82*(2), 146-153.

Olson, S., & Fagen, A. (2007). *Understanding interventions that encourage minorities to pursue research careers: Summary of a workshop.* Washington, DC: National Academies Press.

Osborn, J. and Karukstis, K. K. (2009). The benefits of undergraduate research, scholarship, and creative activity. In M. Boyd and J. Wesemann (Eds.), *Broadening participation in undergraduate research: Fostering excellence and enhancing the impact* (pp. 41-53). Washington, DC: Council on Undergraduate Research.

Osborne, J., Simon, S., & Collin, S. (2003). Attitudes towards science: a review of the literature and its implications. *International Journal of Science Education, 25*(9), 1049-1079.

Perna, L., Lundy-Wagner, V., Drezner, N. D., Gasman, M., Yoon, S., Bose, E., & Gary, S. (2009). The contribution of HBCUS to the preparation of African American women for STEM Careers: A case study. *Reviews of Higher Education, 50*, 1-23.

Pheley, A. M., Lois, H., & Strobl, J. (2006). Interests in research electives among osteopathic medical students. *Journal of the American Osteopathic Association, 106*(11), 667-670.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1-15.

Rask, K. N., & Bailey, E. M. (2002). Are faculty role models? Evidence from major choice in an undergraduate institution. *Journal of Economic Education, 33*(2), 99–124.

Rayman, P., & Brett, B. (1995). Women science majors: What makes a difference in persistence after graduation? *The Journal of Higher Education, 66*(4), 388-414.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354-373.

Rohrbaugh, M. C., & Corces, V. G. (2011). Opening pathways for underrepresented high school students to biomedical research careers: The Emory University RISE program. *Genetics, 189*, 1135-1143.

Rosenberg, L. E. (1999). The physician-scientists: An essential—and fragile—link in the medical research chain. *The Journal of Clinical Investigation, 103*(12), 1621-1626.

Rosenberg, L. E. (2002). Exceptional economic returns on investments in medical research. *MJA, 177*, 368-371.

Rosenberg, L. E. (2008). MD/PhD programs—A call for an accounting. *Journal of American Medical Association, 300*(10), 1208-1209.

Russell, S. H., Hancock, M. P., & McCullogh, J. (2007). The pipeline: Benefits of undergraduate research experiences. *Science, 316*, 548–549.

Sax, L. J. (1994). Retaining tomorrow's scientists: Exploring the factors that keep male and female students interested in science careers. *Journal of Women and Minorities in Science and Engineering, 1*(1), 45-61.

Sax, L. J. (2001). Undergraduate science majors: Gender differences in who goes to graduate school. *Review of Higher Education*, *24*(2), 153-172.

Schafer, A. I. (2010). The vanishing physician-scientist? *Translational Research, 155*(1), 1-3.

Schultz, P. W., Estrada-Hollenbeck, M., & Wood, A. (2008, May) The benefits of being in a minority training program: Preliminary evidence from a national longitudinal study. Presented at the 2nd annual conference for Understanding Interventions, Atlanta, GA.

Seymour, E., Hunter, A., Laursen, S. L., & Deantoni, T. (2004). Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. *Science Education, 88*(4), 493-534.

Simpson, R. D., Koballa, Jr., T. R., Oliver, J. S., Crawley, F. E. (1994). Research on the affective dimension of science learning. In D. Gabel (Ed.) *Handbook of research on science teaching and learning* (pp. 211–234). New York: Macmillan.

Sobral, D. T. (2004). What kind of motivation drives medical students' learning quests? *Medical Education, 38*, 950–957.

Solomon, S. S., Tom, S. C., Pichert, J., Wasserman, D., & Powers, A. C. (2003). Impact of medical student research in the development of physician-scientists. *Journal of Investigative Medicine, 51*, 149-156.

Stevens, J. (2009). *Applied Multivariate Statistics for the Social Sciences* (5th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Stolzenberg, R. M. (1994). Educational continuation by college graduates. *American Journal of Sociology, 99*, 1042-1077.

Summers, M. F., & Hrabowski, F. A. (2006). Preparing minority scientists and engineers. *Science, 311*, 1870–1871.

Sung, N. S., Crowley, W. F., Jr., Genel, M., Salber, P., Sandy, L., Sherwood, L. M., … Rimoin, D. (2003). *Journal of American Medical Association, 289*(10), 1278-1287.

Szelenyi, K., & Inkelas, K. K. (2011). The role of living-learning programs in women's plans to attend graduate school in STEM Fields. *Research in Higher Education, 52*, 349-369.

Tanaka, M., Mizuno, K., Fukuda, S., Tajima, S., & Watanabe, Y. (2009). Personality traits associated with intrinsic academic motivation in medical students. *Medical Education, 43*(4), 384-387. doi:10.1111/j.1365-2923.2008.03279.x

Tang, M., Pan, W., & Newmeyer, M. D. (2008). Factors influencing high school students' career aspirations. *Professional School Counseling*, *11*(5), 285-295.

Thier, S. O., Challoner, D. R., Cockerham, J., Johns, T. R., Mann, M., Skinner, D., ... & MORGAN, T. (1980). Proposals addressing the decline in the training of physician investigators: Report of the ad hoc committee of the AAMC. *Clinical Research*, *28*(2), 85-93.

Tsui, L. (2007). Effective strategies to increase diversity in STEM fields: A review of the research literature. *Journal of Negro Education, 76*,555-581.

U.S. Department of Commerce. (2011, August). *Women in STEM: A gender gap to innovation* (ESA Issue Brief No. 04-11). Washington, DC: Author. Retrieved from http://www.esa.doc.gov/Reports/women-stem-gender-gap-innovation

U.S. Department of Labor. (2007). *The STEM workforce challenge: The role of the public workforce system in a national solution for a competitive science, technology, engineering, and mathematics (STEM) workforce*. Washington, DC: Author.

Varki, A., & Rosenberg, L. E. (2002). Emerging opportunities and career paths for the young physician-scientist. *Nature Medicine, 8*(5), 437-439.

Watt, C. D., Greeley, S. A., Shea, J. A., & Ahn, J. (2005). Educational views and attitudes, and career goals of MD-PhD students at the University of Pennsylvania School of Medicine. *Academic Medicine, 80*, 193-198.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*, 956-972.

Wenzel T. J. (1997). What is undergraduate research? *Council on Undergraduate Research Quarterly, 17*, 163.

Winkleby, M. A. (2007). The Stanford Medical Youth Science Program: 18 years of a biomedical program for low-income high school students. *Academic Medicine, 82*(2), 139-145.

Wyer, M. (2003). Intending to stay: Images of scientists, attitudes toward women, and gender as influences on persistence among science and engineering majors. *Journal of Women and Minorities in Science and Engineering, 9*, 1-16.

Zaikowski, L., Lichtman, P., Quarless, D. (2007). Scientific discovery for all: Keys to developing and sustaining a successful research program. *The Science Teacher, 74*(3), 28-33.

Appendix A

Stata Code for Data Management and Analysis

```
set mem 500m
use "\tremur_data_working_file_20130128.dta", clear

*drop duplicates
drop if R_ID=="262D84E99F25D09E" | R_ID=="CEEC1B3755EDE3D7"

*drop if not matriculated
drop if MAT_AYEAR==" "

*keep if matriculation year is the same or after MCAT administration
year
drop if MAT_AYEAR=="1995-1996" | MAT_AYEAR=="1996-1997" |
MAT_AYEAR=="1997-1998" | MAT_AYEAR=="1998-1999" | MAT_AYEAR=="1999-
2000" | MAT_AYEAR=="2000-2001"

drop if MAT_AYEAR=="2001-2002" & MCAT_ADMIN_YEAR=="2002"
drop if MAT_AYEAR=="2001-2002" & MCAT_ADMIN_YEAR=="2003"
drop if MAT_AYEAR=="2002-2003" & MCAT_ADMIN_YEAR=="2003"
drop if MAT_AYEAR=="2001-2002" & MCAT_ADMIN_YEAR=="2004"
drop if MAT_AYEAR=="2002-2003" & MCAT_ADMIN_YEAR=="2004"
drop if MAT_AYEAR=="2001-2002" & MCAT_ADMIN_YEAR=="2006"

*drop if not graduated
drop if GRAD_AYEAR==" "
count

*totally 87,281 PMQ respondents who entered and graduated from medical
schools within appropriate range

*timeline for MSQ and GQ completion for these 87,281 PMQ respondents

*timeline for PMQ
tab MCAT_ADMIN_YEAR

*timeline for MSQ
tab MSQ_YEAR
tab MSQ_YEAR MSQ_dummy_XK
tab MAT_AYEAR MSQ_YEAR
tab MCAT_ADMIN_YEAR MSQ_YEAR

*drop if no MSQ data (in the corresponding year range) available at all
drop if MAT_AYEAR=="2009-2010" | MAT_AYEAR=="2010-2011"

*timeline for GQ
tab GQ_year_r
tab GQ_year_r GQ_dummy_XK
tab GRAD_AYEAR GQ_year_r if GQ_year_r>0

*drop if no GQ data (in the corresponding year range) available at all
drop if GRAD_AYEAR=="2003-2004" | GRAD_AYEAR=="2011-2012"
```

```
*after the last drop step, I found that no MSQ_YEAR==2008 available,
drop if mat_ayear=2008-2009
drop if MAT_AYEAR=="2008-2009"
count

*totally 77,541 PMQ respondents who entered and graduated from medical
schools within appropriate range, which is the same as the
corresponding year range of MSQ and GQ completion
tab MAT_AYEAR /* matriculation year range: 2001-2002 ~ 2007-2008 */
tab GRAD_AYEAR /* graduation year range: 2004-2005 ~ 2010-2011 */
tab MSQ_YEAR /* MSQ survey completion year: 2001 ~ 2007 */
tab GQ_year_r if GQ_year_r!=0 /* GQ survey completion year: 2005 ~ 2011
*/

*Table 3-1
tab MSQ_dummy_XK GQ_dummy_XK

*Table 3-2
*total (77541)
tab PMQ_SRS_SEX_R
tab PMQ_SRS_race_ethnicity_r

*PMQ respondents who completed both MSQ and GQ (39839)
tab PMQ_SRS_SEX_R if MSQ_dummy_XK==1 & GQ_dummy_XK==1
tab PMQ_SRS_race_ethnicity_r if MSQ_dummy_XK==1 & GQ_dummy_XK==1

*PMQ respondents who completed MSQ, but not completed GQ (13947)
tab PMQ_SRS_SEX_R if MSQ_dummy_XK==1 & GQ_dummy_XK==0
tab PMQ_SRS_race_ethnicity_r if MSQ_dummy_XK==1 & GQ_dummy_XK==0

*PMQ respondents who completed GQ, but not completed MSQ (14164)
tab PMQ_SRS_SEX_R if MSQ_dummy_XK==0 & GQ_dummy_XK==1
tab PMQ_SRS_race_ethnicity_r if MSQ_dummy_XK==0 & GQ_dummy_XK==1

*PMQ respondents who did not complete MSQ, and did not complete GQ
(9591)
tab PMQ_SRS_SEX_R if MSQ_dummy_XK==0 & GQ_dummy_XK==0
tab PMQ_SRS_race_ethnicity_r if MSQ_dummy_XK==0 & GQ_dummy_XK==0

*so far, this data set has cleaned target population data
*then, to obtain the data set of the sample who completed all the three
questionnaires
keep if MSQ_dummy_XK==1 & GQ_dummy_XK==1

*four research interest variables
gen P_INTEREST_RSC_XK=1 if P_INTEREST_RSC=="1"
replace P_INTEREST_RSC_XK=2 if P_INTEREST_RSC=="2"
replace P_INTEREST_RSC_XK=4 if P_INTEREST_RSC=="3"
replace P_INTEREST_RSC_XK=5 if P_INTEREST_RSC=="4"

gen FAC_RESEARCH_XK=FAC_RESEARCH+1

tab MSQ_career_research_rev, nolabel

tab GQ_CAREER_RESEARCH_r, nolabel
```

```
*generate the mean score of the two MSQ variables to get the score that
can represent the MSQ research interest level
egen msq_research_mean= rowmean (FAC_RESEARCH_XK
MSQ_career_research_rev)

*Table 4-11
sum P_INTEREST_RSC_XK, detail
sum msq_research_mean, detail
sum GQ_CAREER_RESEARCH_r, detail

*gender
gen female=1 if PMQ_SRS_SEX_R==1
replace female=0 if PMQ_SRS_SEX_R==2

*Table 4-1
tab female

*race/ethnicity
gen white=0 if PMQ_SRS_race_ethnicity_r!=8
replace white=1 if PMQ_SRS_race_ethnicity_r==1

gen black=0 if PMQ_SRS_race_ethnicity_r!=8
replace black=1 if PMQ_SRS_race_ethnicity_r==2

gen hispanic=0 if PMQ_SRS_race_ethnicity_r!=8
replace hispanic=1 if PMQ_SRS_race_ethnicity_r==3

gen asianpi=0 if PMQ_SRS_race_ethnicity_r!=8
replace asianpi=1 if PMQ_SRS_race_ethnicity_r==4 |
PMQ_SRS_race_ethnicity_r==5

gen native=0 if PMQ_SRS_race_ethnicity_r!=8
replace native=1 if PMQ_SRS_race_ethnicity_r==6

gen othermultiple=0 if PMQ_SRS_race_ethnicity_r!=8
replace othermultiple=1 if PMQ_SRS_race_ethnicity_r==7

*Table 4-2
tab PMQ_SRS_race_ethnicity_r

*three previous research experience variables
*Table 4-6
tab hs_lab_xk

*Table 4-7
tab hs_prog_xk

*Table 4-8
tab coll_lab_xk

*generate two groups: high school lab yes; high school lab no, but
college lab yes
gen hscol=1 if hs_lab_xk==1
replace hscol=0 if hs_lab_xk==0 & coll_lab_xk==1

*matriculating program variable
gen research_prog=1 if MAT_PROG_CD==5
```

```
replace research_prog=0 if MAT_PROG_CD==1

*Table 4-9
tab research_prog

*age
tab AGE_AT_TEST

*Table 4-3
sum AGE_AT_TEST

*parental education
gen pmq_fa_ed=1
replace pmq_fa_ed=. if PMQ_FATHER_EDUC==" "
replace pmq_fa_ed=0 if PMQ_FATHER_EDUC=="A" | PMQ_FATHER_EDUC=="B" |
PMQ_FATHER_EDUC=="C" | PMQ_FATHER_EDUC=="D" | PMQ_FATHER_EDUC=="E"

gen pmq_mo_ed=1
replace pmq_mo_ed=. if PMQ_MOTHER_EDUC==" "
replace pmq_mo_ed=0 if PMQ_MOTHER_EDUC=="A" | PMQ_MOTHER_EDUC=="B" |
PMQ_MOTHER_EDUC=="C" | PMQ_MOTHER_EDUC=="D" | PMQ_MOTHER_EDUC=="E"

gen msq_fa_ed=0 if MSQ_FATHER_EDUC>0 & MSQ_FATHER_EDUC<6
replace msq_fa_ed=1 if MSQ_FATHER_EDUC>5 & MSQ_FATHER_EDUC<14

gen msq_mo_ed=0 if MSQ_MOTHER_EDUC>0 & MSQ_MOTHER_EDUC<6
replace msq_mo_ed=1 if MSQ_MOTHER_EDUC>5 & MSQ_MOTHER_EDUC<14

gen par_ed=0 if pmq_fa_ed==0 & pmq_mo_ed==0
replace par_ed=0 if msq_fa_ed==0 & msq_mo_ed==0
replace par_ed=1 if pmq_fa_ed==1 | pmq_mo_ed==1 | msq_fa_ed==1 |
msq_mo_ed==1
replace par_ed=99 if pmq_fa_ed>1 & pmq_mo_ed>1 & msq_fa_ed>1 &
msq_mo_ed>1
replace par_ed=. if par_ed==99

*Table 4-4
tab par_ed

*parental profession
gen pmq_fa_occ=0
replace pmq_fa_occ=. if NEW_FATHER_OCC==" "
replace pmq_fa_occ=1 if NEW_FATHER_OCC=="01" | NEW_FATHER_OCC=="02" |
NEW_FATHER_OCC=="03" | NEW_FATHER_OCC=="04" | NEW_FATHER_OCC=="05" |
NEW_FATHER_OCC=="06" | NEW_FATHER_OCC=="07" | NEW_FATHER_OCC=="08" |
NEW_FATHER_OCC=="09" | NEW_FATHER_OCC=="10" | NEW_FATHER_OCC=="39" |
NEW_FATHER_OCC=="40" | NEW_FATHER_OCC=="42"

gen pmq_mo_occ=0
replace pmq_mo_occ=. if NEW_MOTHER_OCC==" "
replace pmq_mo_occ=1 if NEW_MOTHER_OCC=="01" | NEW_MOTHER_OCC=="02" |
NEW_MOTHER_OCC=="03" | NEW_MOTHER_OCC=="04" | NEW_MOTHER_OCC=="05" |
NEW_MOTHER_OCC=="06" | NEW_MOTHER_OCC=="07" | NEW_MOTHER_OCC=="08" |
NEW_MOTHER_OCC=="09" | NEW_MOTHER_OCC=="10" | NEW_MOTHER_OCC=="39" |
NEW_MOTHER_OCC=="40" | NEW_MOTHER_OCC=="42"

gen msq_fa_occ=0
```

```
replace msq_fa_occ=1 if FATHER_OCC>0 & FATHER_OCC<5
replace msq_fa_occ=. if FATHER_OCC>23

gen msq_mo_occ=0
replace msq_mo_occ=1 if MOTHER_OCC>0 & MOTHER_OCC<5
replace msq_mo_occ=. if MOTHER_OCC>23

gen par_occ=0 if pmq_fa_occ==0 & pmq_mo_occ==0
replace par_occ=0 if msq_fa_occ==0 & msq_mo_occ==0
replace par_occ=1 if pmq_fa_occ==1 | pmq_mo_occ==1 | msq_fa_occ==1 |
msq_mo_occ==1

*Table 4-5
tab par_occ

*missing data anlaysis
*Table 4-18
tab P_INTEREST_RSC_XK
gen pmq_miss=0
replace pmq_miss=1 if P_INTEREST_RSC_XK>100
pwcorr pmq_miss FAC_RESEARCH_XK MSQ_career_research_rev
GQ_CAREER_RESEARCH_r female white black hispanic asianpi native
othermultiple hs_lab_xk hs_prog_xk coll_lab_xk research_prog
AGE_AT_TEST par_ed par_occ, sig

tab FAC_RESEARCH_XK
gen msq_a_miss=0
replace msq_a_miss=1 if FAC_RESEARCH_XK>100
pwcorr msq_a_miss P_INTEREST_RSC_XK MSQ_career_research_rev
GQ_CAREER_RESEARCH_r female white black hispanic asianpi native
othermultiple hs_lab_xk hs_prog_xk coll_lab_xk research_prog
AGE_AT_TEST par_ed par_occ, sig

tab MSQ_career_research_rev
gen msq_b_miss=0
replace msq_b_miss=1 if MSQ_career_research_rev>100
pwcorr msq_b_miss P_INTEREST_RSC_XK FAC_RESEARCH_XK
GQ_CAREER_RESEARCH_r female white black hispanic asianpi native
othermultiple hs_lab_xk hs_prog_xk coll_lab_xk research_prog
AGE_AT_TEST par_ed par_occ, sig

tab GQ_CAREER_RESEARCH_r
gen gq_miss=0
replace gq_miss=1 if GQ_CAREER_RESEARCH_r>100
pwcorr gq_miss P_INTEREST_RSC_XK FAC_RESEARCH_XK
MSQ_career_research_rev female white black hispanic asianpi native
othermultiple hs_lab_xk hs_prog_xk coll_lab_xk research_prog
AGE_AT_TEST par_ed par_occ, sig
```

Appendix B

SPSS Code for Data Analysis

```
*PCA for the two MSQ research interest variables
*Table 4-10
FACTOR
  /VARIABLES FAC_RESEARCH_XK MSQ_career_research_rev
  /MISSING LISTWISE
  /ANALYSIS FAC_RESEARCH_XK MSQ_career_research_rev
  /PRINT INITIAL EXTRACTION
  /FORMAT SORT
  /PLOT EIGEN
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /ROTATION NOROTATE
  /METHOD=CORRELATION.
```

Appendix C

SAS Code for Data Analysis

```
proc import datafile="\test4_20130908_sample_2_fromR.dta" out=mydata
dbms = dta replace;
run;

data mydata_long;
set mydata;
y=P_INTEREST_RSC_XK;time=0;t=time;output;
y=msq_research_mean;time=1;t=time;output;
y=GQ_CAREER_RESEARCH_r;time=2;t=time;output;
drop P_INTEREST_RSC_XK msq_research_mean GQ_CAREER_RESEARCH_r;
run;

data mydata_long;
set mydata_long;
timesqr=time*time;
run;

data mydata_long;
set mydata_long;
timeknot=max(time-1,0);
run;

*gender;

*Table 4-12;
proc means data=mydata;
var P_INTEREST_RSC_XK msq_research_mean GQ_CAREER_RESEARCH_r;
class female;
run;

*Table 4-19a;
*linear model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*female female research_prog hs_lab_xk hs_prog_xk
coll_lab_xk black hispanic asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*linear model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*female female research_prog hs_lab_xk hs_prog_xk
coll_lab_xk black hispanic asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*linear model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
```

```
model y=time time*female female research_prog hs_lab_xk hs_prog_xk
coll_lab_xk black hispanic asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*linear model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*female female research_prog hs_lab_xk hs_prog_xk
coll_lab_xk black hispanic asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*quadratic model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr female time*female timesqr*female research_prog
hs_lab_xk hs_prog_xk coll_lab_xk black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*quadratic model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr female time*female timesqr*female research_prog
hs_lab_xk hs_prog_xk coll_lab_xk black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*quadratic model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr female time*female timesqr*female research_prog
hs_lab_xk hs_prog_xk coll_lab_xk black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*quadratic model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr female time*female timesqr*female research_prog
hs_lab_xk hs_prog_xk coll_lab_xk black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*spline model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot female time*female timeknot*female research_prog
hs_lab_xk hs_prog_xk coll_lab_xk black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
```

```
repeated t / subject=R_ID type=un r rcorr;
run;

*spline model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot female time*female timeknot*female research_prog
hs_lab_xk hs_prog_xk coll_lab_xk black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*spline model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot female time*female timeknot*female research_prog
hs_lab_xk hs_prog_xk coll_lab_xk black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*spline model--Autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot female time*female timeknot*female research_prog
hs_lab_xk hs_prog_xk coll_lab_xk black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*after covariance structures are selected in each of the three mean
models using REML, ML is used to compare the mean models;
*Table 4-19b;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time time*female female research_prog hs_lab_xk hs_prog_xk
coll_lab_xk black hispanic asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timesqr female time*female timesqr*female research_prog
hs_lab_xk hs_prog_xk coll_lab_xk black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Table 4-28;
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timeknot female time*female timeknot*female research_prog
hs_lab_xk hs_prog_xk coll_lab_xk black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
```

```
run;

*Race/Ethnicity;

*Table 4-13;
proc means data=mydata;
var P_INTEREST_RSC_XK msq_research_mean GQ_CAREER_RESEARCH_r;
class PMQ_SRS_race_ethnicity_r;
run;

*Asian/Pacific Islander vs. White;

*Table 4-20a;
*linear model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*asianpi asianpi research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*linear model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*asianpi asianpi research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*linear model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*asianpi asianpi research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*linear model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*asianpi asianpi research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*quadratic model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr asianpi time*asianpi timesqr*asianpi research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;
```

```
*quadratic model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr asianpi time*asianpi timesqr*asianpi research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*quadratic model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr asianpi time*asianpi timesqr*asianpi research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*quadratic model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr asianpi time*asianpi timesqr*asianpi research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*spline model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot asianpi time*asianpi timeknot*asianpi
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*spline model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot asianpi time*asianpi timeknot*asianpi
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*spline model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot asianpi time*asianpi timeknot*asianpi
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*spline model--Autoregressive;
proc mixed data=mydata_long order=data;
```

```
class R_ID t;
model y=time timeknot asianpi time*asianpi timeknot*asianpi
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*after covariance structures are selected in each of the three mean
models using REML, ML is used to compare the mean models;
*Table 4-20b;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time time*asianpi asianpi research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timesqr asianpi time*asianpi timesqr*asianpi research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Table 4-29;
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timeknot asianpi time*asianpi timeknot*asianpi
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Black vs. White;
*Table 4-21a;

*linear model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*black black research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female hispanic asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*linear model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*black black research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female hispanic asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;
```

```
*linear model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*black black research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female hispanic asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*linear model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*black black research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female hispanic asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*quadratic model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr black time*black timesqr*black research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*quadratic model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr black time*black timesqr*black research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*quadratic model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr black time*black timesqr*black research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*quadratic model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr black time*black timesqr*black research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*spline model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
```

```
model y=time timeknot black time*black timeknot*black research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*spline model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot black time*black timeknot*black research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*spline model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot black time*black timeknot*black research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*spline model--Autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot black time*black timeknot*black research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*after covariance structures are selected in each of the three mean
models using REML, ML is used to compare the mean models;
*Table 4-21b;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time time*black black research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female hispanic asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timesqr black time*black timesqr*black research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Table 4-30;
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
```

```
model y=time timeknot black time*black timeknot*black research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Hispanic vs. White;
*Table 4-22a;

*linear model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hispanic hispanic research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*linear model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hispanic hispanic research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*linear model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hispanic hispanic research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*quadratic model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hispanic time*hispanic timesqr*hispanic
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*quadratic model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hispanic time*hispanic timesqr*hispanic
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*quadratic model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
```

```
model y=time timesqr hispanic time*hispanic timesqr*hispanic
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*quadratic model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hispanic time*hispanic timesqr*hispanic
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*spline model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hispanic time*hispanic timeknot*hispanic
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*spline model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hispanic time*hispanic timeknot*hispanic
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*spline model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hispanic time*hispanic timeknot*hispanic
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*spline model--Autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hispanic time*hispanic timeknot*hispanic
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*after covariance structures are selected in each of the three mean
models using REML, ML is used to compare the mean models;
*Table 4-22b;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
```

```
model y=time time*hispanic hispanic research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black asianpi native othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timesqr hispanic time*hispanic timesqr*hispanic
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Table 4-31;
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timeknot hispanic time*hispanic timeknot*hispanic
research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Native vs. White;
*Table 4-23a;

*linear model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*native native research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic asianpi othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*linear model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*native native research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic asianpi othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*linear model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*native native research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic asianpi othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*linear model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
```

```
model y=time time*native native research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic asianpi othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*quadratic model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr native time*native timesqr*native research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic asianpi
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*quadratic model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr native time*native timesqr*native research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic asianpi
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*quadratic model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr native time*native timesqr*native research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic asianpi
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*quadratic model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr native time*native timesqr*native research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic asianpi
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*spline model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot native time*native timeknot*native research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic asianpi
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*spline model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot native time*native timeknot*native research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic asianpi
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
```

```
repeated t / subject=R_ID type=cs r rcorr;
run;

*spline model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot native time*native timeknot*native research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic asianpi
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*spline model--Autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot native time*native timeknot*native research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic asianpi
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*after covariance structures are selected in each of the three mean
models using REML, ML is used to compare the mean models;
*Table 4-23b;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time time*native native research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic asianpi othermultiple AGE_AT_TEST
par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timesqr native time*native timesqr*native research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic asianpi
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

Table 4-32;
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timeknot native time*native timeknot*native research_prog
hs_lab_xk hs_prog_xk coll_lab_xk female black hispanic asianpi
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Students with high school lab research vs. others;
*Table 4-14;
proc means data=mydata;
var P_INTEREST_RSC_XK msq_research_mean GQ_CAREER_RESEARCH_r;
class hs_lab_xk;
run;
```

```
*Table 4-24a;

*linear model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hs_lab_xk hs_lab_xk research_prog hs_prog_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*linear model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hs_lab_xk hs_lab_xk research_prog hs_prog_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*linear model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hs_lab_xk hs_lab_xk research_prog hs_prog_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*linear model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hs_lab_xk hs_lab_xk research_prog hs_prog_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*quadratic model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hs_lab_xk time*hs_lab_xk timesqr*hs_lab_xk
research_prog hs_prog_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*quadratic model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hs_lab_xk time*hs_lab_xk timesqr*hs_lab_xk
research_prog hs_prog_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*quadratic model--Toeplitz;
```

```
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hs_lab_xk time*hs_lab_xk timesqr*hs_lab_xk
research_prog hs_prog_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*quadratic model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hs_lab_xk time*hs_lab_xk timesqr*hs_lab_xk
research_prog hs_prog_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*spline model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hs_lab_xk time*hs_lab_xk timeknot*hs_lab_xk
research_prog hs_prog_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*spline model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hs_lab_xk time*hs_lab_xk timeknot*hs_lab_xk
research_prog hs_prog_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*spline model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hs_lab_xk time*hs_lab_xk timeknot*hs_lab_xk
research_prog hs_prog_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*spline model--Autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hs_lab_xk time*hs_lab_xk timeknot*hs_lab_xk
research_prog hs_prog_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*after covariance structures are selected in each of the three mean
models using REML, ML is used to compare the mean models;
*Table 4-24b;
```

```
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time time*hs_lab_xk hs_lab_xk research_prog hs_prog_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timesqr hs_lab_xk time*hs_lab_xk timesqr*hs_lab_xk
research_prog hs_prog_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Table 4-33;
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timeknot hs_lab_xk time*hs_lab_xk timeknot*hs_lab_xk
research_prog hs_prog_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Students with high school classroom-based programs vs. others;
*Table 4-15;

proc means data=mydata;
var P_INTEREST_RSC_XK msq_research_mean GQ_CAREER_RESEARCH_r;
class hs_prog_xk;
run;

*Table 4-25a;

*linear model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hs_prog_xk hs_prog_xk research_prog hs_lab_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*linear model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hs_prog_xk hs_prog_xk research_prog hs_lab_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*linear model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
```

```
model y=time time*hs_prog_xk hs_prog_xk research_prog hs_lab_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*linear model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hs_prog_xk hs_prog_xk research_prog hs_lab_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*quadratic model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hs_prog_xk time*hs_prog_xk timesqr*hs_prog_xk
research_prog hs_lab_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*quadratic model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hs_prog_xk time*hs_prog_xk timesqr*hs_prog_xk
research_prog hs_lab_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*quadratic model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hs_prog_xk time*hs_prog_xk timesqr*hs_prog_xk
research_prog hs_lab_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*quadratic model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hs_prog_xk time*hs_prog_xk timesqr*hs_prog_xk
research_prog hs_lab_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*spline model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hs_prog_xk time*hs_prog_xk timeknot*hs_prog_xk
research_prog hs_lab_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
```

```
repeated t / subject=R_ID type=un r rcorr;
run;

*spline model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hs_prog_xk time*hs_prog_xk timeknot*hs_prog_xk
research_prog hs_lab_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*spline model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hs_prog_xk time*hs_prog_xk timeknot*hs_prog_xk
research_prog hs_lab_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*spline model--Autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hs_prog_xk time*hs_prog_xk timeknot*hs_prog_xk
research_prog hs_lab_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*after covariance structures are selected in each of the three mean
models using REML, ML is used to compare the mean models;
*Table 4-25b;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time time*hs_prog_xk hs_prog_xk research_prog hs_lab_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timesqr hs_prog_xk time*hs_prog_xk timesqr*hs_prog_xk
research_prog hs_lab_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Table 4-34;
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timeknot hs_prog_xk time*hs_prog_xk timeknot*hs_prog_xk
research_prog hs_lab_xk coll_lab_xk female black hispanic asianpi
native othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
```

```
run;

*Students with college lab research vs. others;
*Table 4-16;
proc means data=mydata;
var P_INTEREST_RSC_XK msq_research_mean GQ_CAREER_RESEARCH_r;
class coll_lab_xk;
run;

*Table 4-26a;

*linear model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*coll_lab_xk coll_lab_xk research_prog hs_lab_xk
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*linear model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*coll_lab_xk coll_lab_xk research_prog hs_lab_xk
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*linear model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*coll_lab_xk coll_lab_xk research_prog hs_lab_xk
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*linear model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*coll_lab_xk coll_lab_xk research_prog hs_lab_xk
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*quadratic model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr coll_lab_xk time*coll_lab_xk timesqr*coll_lab_xk
research_prog hs_lab_xk hs_prog_xk female black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*quadratic model--compound symmetry;
```

```
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr coll_lab_xk time*coll_lab_xk timesqr*coll_lab_xk
research_prog hs_lab_xk hs_prog_xk female black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*quadratic model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr coll_lab_xk time*coll_lab_xk timesqr*coll_lab_xk
research_prog hs_lab_xk hs_prog_xk female black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*quadratic model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr coll_lab_xk time*coll_lab_xk timesqr*coll_lab_xk
research_prog hs_lab_xk hs_prog_xk female black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*spline model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot coll_lab_xk time*coll_lab_xk timeknot*coll_lab_xk
research_prog hs_lab_xk hs_prog_xk female black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*spline model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot coll_lab_xk time*coll_lab_xk timeknot*coll_lab_xk
research_prog hs_lab_xk hs_prog_xk female black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*spline model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot coll_lab_xk time*coll_lab_xk timeknot*coll_lab_xk
research_prog hs_lab_xk hs_prog_xk female black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*spline model--Autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
```

```
model y=time timeknot coll_lab_xk time*coll_lab_xk timeknot*coll_lab_xk
research_prog hs_lab_xk hs_prog_xk female black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*after covariance structures are selected in each of the three mean
models using REML, ML is used to compare the mean models;
*Table 4-26b;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time time*coll_lab_xk coll_lab_xk research_prog hs_lab_xk
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timesqr coll_lab_xk time*coll_lab_xk timesqr*coll_lab_xk
research_prog hs_lab_xk hs_prog_xk female black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Table 4-35;
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timeknot coll_lab_xk time*coll_lab_xk timeknot*coll_lab_xk
research_prog hs_lab_xk hs_prog_xk female black hispanic asianpi native
othermultiple AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Students with high school lab research vs. students with only college
lab research;
*Table 4-36a;

*linear model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hscol hscol research_prog hs_prog_xk female black
hispanic asianpi native othermultiple AGE_AT_TEST par_ed par_occ/S
CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*linear model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hscol hscol research_prog hs_prog_xk female black
hispanic asianpi native othermultiple AGE_AT_TEST par_ed par_occ/S
CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;
```

```
*linear model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*hscol hscol research_prog hs_prog_xk female black
hispanic asianpi native othermultiple AGE_AT_TEST par_ed par_occ/S
CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*quadratic model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hscol time*hscol timesqr*hscol research_prog
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*quadratic model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hscol time*hscol timesqr*hscol research_prog
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*quadratic model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hscol time*hscol timesqr*hscol research_prog
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*quadratic model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr hscol time*hscol timesqr*hscol research_prog
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*spline model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hscol time*hscol timeknot*hscol research_prog
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*spline model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
```

```
model y=time timeknot hscol time*hscol timeknot*hscol research_prog
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*spline model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hscol time*hscol timeknot*hscol research_prog
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*spline model--Autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot hscol time*hscol timeknot*hscol research_prog
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*after covariance structures are selected in each of the three mean
models using REML, ML is used to compare the mean models;
*Table 4-36b;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time time*hscol hscol research_prog hs_prog_xk female black
hispanic asianpi native othermultiple AGE_AT_TEST par_ed par_occ/S
CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timesqr hscol time*hscol timesqr*hscol research_prog
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Table 4-37;
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timeknot hscol time*hscol timeknot*hscol research_prog
hs_prog_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*MD/PhD maticulants vs. MD-only matriculants;
*Table 4-17;

proc means data=mydata;
```

```
var P_INTEREST_RSC_XK msq_research_mean GQ_CAREER_RESEARCH_r;
class research_prog;
run;

*Table 4-27a;

*linear model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*research_prog research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*linear model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*research_prog research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*linear model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*research_prog research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*linear model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time time*research_prog research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*quadratic model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr research_prog time*research_prog
timesqr*research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black
hispanic asianpi native othermultiple AGE_AT_TEST par_ed par_occ/S
CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*quadratic model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr research_prog time*research_prog
timesqr*research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black
```

```
hispanic asianpi native othermultiple AGE_AT_TEST par_ed par_occ/S
CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*quadratic model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr research_prog time*research_prog
timesqr*research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black
hispanic asianpi native othermultiple AGE_AT_TEST par_ed par_occ/S
CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*quadratic model--autoregressive;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timesqr research_prog time*research_prog
timesqr*research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black
hispanic asianpi native othermultiple AGE_AT_TEST par_ed par_occ/S
CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*spline model--unstructured;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot research_prog time*research_prog
timeknot*research_prog research_prog hs_lab_xk hs_prog_xk coll_lab_xk
female black hispanic asianpi native othermultiple AGE_AT_TEST par_ed
par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*spline model--compound symmetry;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot research_prog time*research_prog
timeknot*research_prog research_prog hs_lab_xk hs_prog_xk coll_lab_xk
female black hispanic asianpi native othermultiple AGE_AT_TEST par_ed
par_occ/S CHISQ;
repeated t / subject=R_ID type=cs r rcorr;
run;

*spline model--Toeplitz;
proc mixed data=mydata_long order=data;
class R_ID t;
model y=time timeknot research_prog time*research_prog
timeknot*research_prog research_prog hs_lab_xk hs_prog_xk coll_lab_xk
female black hispanic asianpi native othermultiple AGE_AT_TEST par_ed
par_occ/S CHISQ;
repeated t / subject=R_ID type=toep r rcorr;
run;

*spline model--Autoregressive;
proc mixed data=mydata_long order=data;
```

```
class R_ID t;
model y=time timeknot research_prog time*research_prog
timeknot*research_prog research_prog hs_lab_xk hs_prog_xk coll_lab_xk
female black hispanic asianpi native othermultiple AGE_AT_TEST par_ed
par_occ/S CHISQ;
repeated t / subject=R_ID type=ar(1) r rcorr;
run;

*after covariance structures are selected in each of the three mean
models using REML, ML is used to compare the mean models;
*Table 4-27b;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time time*research_prog research_prog hs_lab_xk hs_prog_xk
coll_lab_xk female black hispanic asianpi native othermultiple
AGE_AT_TEST par_ed par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timesqr research_prog time*research_prog
timesqr*research_prog hs_lab_xk hs_prog_xk coll_lab_xk female black
hispanic asianpi native othermultiple AGE_AT_TEST par_ed par_occ/S
CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;

*Table 4-38;
proc mixed data=mydata_long order=data method=ml;
class R_ID t;
model y=time timeknot research_prog time*research_prog
timeknot*research_prog research_prog hs_lab_xk hs_prog_xk coll_lab_xk
female black hispanic asianpi native othermultiple AGE_AT_TEST par_ed
par_occ/S CHISQ;
repeated t / subject=R_ID type=un r rcorr;
run;
```