INVESTIGATING THE RELATIONSHIP BETWEEN THE USE OF ADVANCED PLACEMENT CREDIT AND PERFORMANCE IN SUBSEQUENT COLLEGE COURSES

by

Sheila Hurt

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Educational Studies

West Lafayette, Indiana

May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Yukiko Maeda, Chair

Department of Educational Studies

Dr. Anne Traynor

Department of Educational Studies

Dr. Kristina Wong Davis

Vice Provost for Enrollment Management

Dr. Alina Alexeenko

Department of Aeronautics and Astronautics

Approved by:

Dr. Janet Alsup

Dedicated to the UIA Fellows Family without whom none of this would have happened

ACKNOWLEDGMENTS

There are so many people to thank for helping me to get across the finish line, but I'll start by expressing my sincere thanks to my advisor, Yukiko Maeda, who always knew when to push and when to give me space. I look forward to continuing our research collaborations in the future! I am also grateful to my committee members: to Kris Wong Davis for being my champion (as promised!), to Anne Traynor for her insightful methodology suggestions, and to Alina Alexeenko for helping get the results in front of people who might use them. Frank Dooley and Jenna Rickus got me started on this research topic, and have been fantastic supervisors, encouraging me to combine my professional and scholarly pursuits. Other wonderful Purdue people who made this dissertation possible include Shawn Bauldry, whose help with both propensity modeling and demystifying R was invaluable; the EMAR team (especially Julie Huser) who came through every time I presented them with another enormous data request; my colleagues Sandy Monroe, Christina King, Molly Gilbert, and Heather Servaty-Seib, who provided emotional support along the way and a great surprise celebration at the end; and John Gipson, who traveled this path ahead of me and helped me figure out how to navigate. Of course I also want to thank my supportive family members: my parents and sister who believed from the start that I could do this, my brother-in-law Dan who taught me enough Excel tricks that I actually *could* do it, and my three children who were (usually) patient and understanding every time I had to spend another weekend or evening on the computer. Finally, I want to thank my husband, Jon-Paul, who held down the home front for the past four years, and was always willing to listen to me talk through some methodological challenge; I could not have done this without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Research on the Advanced Placement (AP) program generally shows that students scoring 4s and 5s on AP exams outperform their non-AP peers in subsequent college courses. However, faculty often advise students with AP credit to repeat prerequisite courses in college before attempting advanced coursework, and there are few studies that provide empirical evidence about outcomes related to the use of AP credit. I compared grades of 26,843 students in 34 STEM courses using two-level cross-sectional multilevel modeling and found that students with AP credit in biology, calculus, chemistry, or physics significantly outperformed non-AP students after controlling for high school GPA and SAT Math scores, whether they used their AP credit to fulfill course pre-requisites or not. Additionally, I investigated outcomes of 10,152 students who had earned AP credit for course pre-requisites, depending on whether or not they chose to use their AP credit or repeat it at the college level prior to taking subsequent courses. I found that contextual factors, such as the specific target course and the student's home college, were highly influential in determining the propensity to use AP credit. Measures of prior academic achievement also influenced the propensity to use AP credit, but most demographic factors did not. After applying propensity weights, I found no causal effect of using AP credit on subsequent course grades. The use of two-level cross-sectional multilevel modeling showed that the effect of using AP credit on subsequent course grades varied significantly across courses. The results of this study show that students who use AP credit to move directly into subsequent college STEM courses do not earn lower grades in those courses as a result of their decision to use AP credit.

CHAPTER 1. INTRODUCTION

Advanced Placement (AP) courses and exams have become a fixture in American high schools, with parents, educators, and policymakers encouraging more and more students to enroll in hopes that participation in AP courses will improve their odds of first being admitted to and eventually being successful in college (Klopfenstein & Thomas, 2010). The College Board first offered AP exams in the early 1950s as a collaboration between elite college preparatory schools and university faculty in which high schools designed their own curriculum to prepare students for a common exam graded by college professors (Lichten, 2000). In the early years, faculty received the actual answer books of their enrolled students and evaluated their readiness for advanced work on an individual basis; eventually the AP exams became respected enough that institutions awarded credit based on exam scores alone (Casserly, 1986). Since its inception, the AP program has grown substantially both in the number of subjects offered and the availability of AP-designed courses in high schools (Ackerman et al., 2013). Today there are 38 AP courses and exams available, and at least one AP course is offered at over 22,000 high schools in the United States; in the high school graduating class of 2019, 1.25 million students took at least one AP exam (College Board, n.d., *AP program results: Class of 2019*).

One reason for this significant growth in AP participation is that a positive relationship between the AP program and better college outcomes has been accepted as "conventional wisdom" by the general public (Sadler & Tai, 2007, p. 5). The College Board claims that participating in the AP program benefits students in three ways: it helps students get accepted into college, be more successful once admitted, and graduate more quickly, saving time and money (College Board, n.d., *Benefits of AP*). Indeed, there is evidence for the first claim, as selective college and university admissions practices reward participation in AP courses, either directly or through giving

additional GPA points for the presumably more challenging courses (Geiser & Santelices, 2004; Klopfenstein & Thomas, 2010). As the emphasis on AP has grown in college admissions, parents and educational policymakers have pressured high schools to offer more AP courses (Geiser & Santelices, 2004), and federal and state governments have enacted policies to increase AP participation among high school students (Lichten, 2000).

The College Board's second claim, that AP students are more successful in college, was supported by early research showing that AP students earned higher grades in college courses (Casserly, 1986) and graduated at higher rates than their non-AP peers (Dougherty et al., 2006). The general public has typically interpreted such research as causal, inferring that the expansion of AP offerings would lead to greater student success (Klopfenstein & Thomas, 2009). Scholars have called those findings into question, however, since much of the initial research on the AP program appears to have had some methodological shortcomings. For example, early studies often involved simple comparisons of outcomes between AP and non-AP students, and did not account for important differences between those two groups that are also highly correlated with student success (Sadler & Tai, 2007). More recently, AP researchers have attempted to account for some possible confounders of the relationship between AP participation and college academic success, such as measures of prior academic achievement and student demographics (Ackerman et al., 2013; Hargrove et al., 2008). In Warne's (2017) review of the AP literature, he concluded that while researchers generally found positive results based on AP performance even after controlling for differences in student backgrounds, the reported effect sizes were much smaller than those reported in less methodologically rigorous research.

There is even less evidence to support the College Board's third claim, that successful AP students will graduate from college more quickly than they otherwise would have done. Few AP

studies have investigated this possible benefit (Klopfenstein, 2010), but in one example, University of California institutional researchers found little relationship between AP credits earned and time to degree (Eykamp, 2006). Evans (2019) reviewed how college students in a national sample used their AP credit and found that very few shortened their time to degree at all.

If students are not using AP credit to speed graduation, it may be the case that an academic advisor or faculty member recommended they repeat coursework instead of moving ahead (Sadler & Sonnert, 2010). Faculty may consider that AP students are not adequately prepared for advanced college coursework. In particular, students in STEM majors often choose to repeat calculus in college after earning AP credit, following the advice of their advisors or other STEM students (Sadler & Sonnert, 2018). In fact, most STEM faculty surveyed by the National Research Council (2002) thought students should take introductory classes in college even if they had earned AP credit, because the faculty doubted the equivalence of AP courses to college courses with regard to depth and rigor. While most universities do award credit based on AP exam scores (Ackerman et al., 2013), and legislation mandating the awarding of AP credit for scores of 3 or higher has been enacted in 22 states (College Board, n.d., *State and systemwide AP credit and placement policies*), some selective institutions have stopped granting credit for AP exam scores, or have raised exam score thresholds so only students who score 5s receive credit (Burkholder & Wieman, 2019; Drew, 2011).

It is therefore clear that there are diverging views about the benefits of the AP program, with parents' and secondary school administrators' belief that success in the AP program leads to success in college (Klopfenstein & Thomas, 2009) running counter to the skepticism of college faculty (National Research Council, 2002). This discrepancy is understandable given our incomplete understanding of the effects of AP participation and performance on student outcomes

in college. For even when scholars attempt to control for confounding factors, studies that treat AP course-taking as exogenous (meaning that the factors influencing a student's choice to take an AP course are completely independent of their eventual success in college) may report biased results (Clark et al., 2012). While a recent study (Conger et al., 2021) has shown it is possible to conduct a randomized controlled trial experiment to assess AP effects, such studies are rare and costly, so most AP research is non-experimental.

Beyond the difficulty in estimating the causal effects of AP participation and performance, possibly the largest gap in the existing body of knowledge on the relationship between AP and college academic outcomes is the lack of studies that account for students who choose not to use their earned AP credit (De Urquidi et al., 2015). Additional research is also needed that considers the variability of effects across different subjects (Warne, 2017), including courses outside the cognate department that granted the AP credit. Therefore, this study is designed to increase the breadth of knowledge about the AP program in three ways.

First, I included the group of students who earn AP credit but choose not to use it, instead repeating the equivalent course at the college level, in order to investigate outcomes associated with using AP credit. This study focused on those students who have earned credit based on their AP exam scores and must decide if they want to repeat the material, either to deepen their understanding or to have a greater chance of earning higher grades given their prior exposure to the material (Burkholder & Wieman, 2019; Sadler & Sonnert, 2018), or if they want to use their AP credit and either reduce their time to degree or free up time for other opportunities (Evans, 2019). Most studies that investigate college student achievement at the course level compare students who earned and used AP credit for the course pre-requisite to students without AP credit who enrolled in the pre-requisite course in college; they either ignore or explicitly exclude students

who earned AP credit but chose to repeat it (Patterson & Ewing, 2013; Shaw et al., 2013). This approach is reasonable for College Board studies that aim to evaluate whether AP exam grades are valid indicators of student readiness for placement into subsequent coursework (Ewing, 2006). However, while such studies may inform future changes to institutional policies regarding the granting of AP exam credit, they are not very helpful for students who have already earned exam credit based on existing university policies, and must decide whether or not to use that credit. Given the College Board's claims that AP students are more successful than non-AP students (n.d., *Discover the Benefits of AP*), it is meaningful for both students and institutions to understand whether the future success of AP students is associated with their choice to use (or not use) their earned AP credit. Two studies (i.e., De Urquidi et al., 2015; Hansen et al., 2006) that investigated whether students should accept AP credit or repeat the course in college provided helpful insights, however, neither was sufficient to answer the question broadly because both were limited to one subject.

Second, this study builds on others that investigated variations in AP effects across courses through the use of multilevel modeling. Studies that include comparisons of subsequent course grades between AP and non-AP students and report results by subject area show markedly different results across subjects (Dodd et al., 2002; Patterson & Ewing, 2013). De Urquidi et al. (2015) limited their investigation to calculus courses, but found different results depending on whether the subsequent course was second- or third-semester calculus; therefore, even reporting results by subject area may obscure differences across courses. There are differences across AP exams in terms of the level of skill needed to be successful, the demographics of students who take the exams, and of course the subject-specific content, so studies of AP effects should not assume that AP effects are the same across all subject areas (Warne, 2017).

Part of considering variation across AP courses involves the inclusion of grades AP students earn in subsequent courses that are outside the cognate department that granted their AP credit. Most commonly, studies that investigate grades earned by AP students in subsequent courses limit their sample of courses to those in the same subject, as in Wyatt et al. (2018). This approach ignores important STEM fields such as engineering since there are no AP engineering courses, but AP courses in calculus, chemistry, and physics often serve as pre-requisites for college-level engineering coursework (Patterson & Ewing, 2013). This study included courses in the four AP cognate departments (which include biology as well as the three listed previously) but also courses in animal science, statistics, and eleven engineering disciplines. The use of multilevel modeling in this study allowed for a common estimate of the average AP effect while allowing that effect to vary across this wide range of courses.

The application of a statistical approach like multilevel modeling is also needed to account for the nested nature of AP data, in which students are clustered in courses and schools or colleges; additionally, its use allows for the consideration of course-level predictors and possible cross-level interactions (Warne, 2017). Nevertheless, relatively few studies have used multilevel modeling to study AP data. Among those that have applied multilevel models to account for clustering effects, Dougherty et al. (2006) clustered students by high school, accounting for differences in school characteristics including the percentage of students taking AP courses in each high school, and Shaw et al. (2013) nested students by college, to account for grading differences across institutions. In this study, I used two-level cross-sectional multilevel modeling to cluster students in college courses, in order to investigate possible course-level variation as identified by De Urquidi et al. (2015). While little research has been conducted to explore the variation due to course characteristics, course difficulty level is a possible avenue for exploration (Wladis et al., 2017);

the course's historic DFW rate (the percentage of students who earn Ds, Fs, or withdraw from the course) is one measure of course difficulty that I included as a possible predictor.

Finally, this study adds to the collective understanding of AP students' college success by improving the extent to which we can draw causal inferences from estimates of AP effects through the use of inverse propensity weights. Propensity score modeling is one approach scholars can take to isolate AP effects, given that students who pass AP exams tend to be highly motivated, high-achieving students who are already likely to succeed in college regardless of their AP participation (Sadler, 2010a). Propensity scores represent the predicted probability that any student in the sample would select into the study's treatment condition, such as choosing to use earned AP credit in college (Warne et al., 2015). One advantage of using propensity score models is that they do not rely on the outcomes of students who are very unlikely to use AP credit in order to estimate the effects on those who do choose to use it (Long et al., 2012). There are multiple ways to include propensity for treatment in an analysis, such as one-to-one matching, stratified matching, or the use of inverse propensity weights. While its use is not yet widespread in AP research (Warne, 2017), a few AP studies have used simple stratified matching (e.g., Dodd et al., 2002; Hargrove et al., 2008), or more formal propensity models (e.g., Clark et al., 2012; Patterson & Ewing, 2013) to account for pre-college student differences. Similarly, this study used propensity weights to account for factors that could increase the likelihood of choosing to use AP credit, such as AP exam scores, measures of prior academic achievement, demographic characteristics, and contextual factors related to the student's home college and the specific courses taken. The use of propensity weights has advantages over propensity score matching, including a more straightforward calculation of standard errors (Morgan & Winship, 2015), and this approach avoided the significant loss of data that can occur when implementing a one-to-one matching

approach (e.g., Patterson & Ewing, 2013). Thus, using propensity weights in the multilevel analysis allowed for a more accurate estimate of causal effects than is possible in most existing AP research.

This study was also built upon the findings and insights obtained from a preceding study (Hurt & Maeda, under review). In that study, we used multilevel modeling to compare target course grades across three groups of students: those with no AP credit for the course pre-requisite, those who used AP credit as a pre-requisite, and those with AP credit who chose to repeat the pre-requisite course in college before attempting the target course. That study found that both AP groups outperformed non-AP students after controlling for high school grades and SAT scores. On average across courses, the students who repeated AP credit earned slightly higher grades than those who used it; the positive association between course grades and using AP credit varied across courses and was stronger in more difficult courses (Hurt & Maeda, under review). One limitation of this prior study is that it did not account for potentially meaningful factors that would lead a student to choose to use or repeat earned AP credit, such as AP exam scores. In addition to the use of propensity weights to mitigate that limitation, this study extended the prior work by including target courses outside the cognate department that granted AP credit, and target courses with multiple pre-requisites. Additionally, Hurt and Maeda (under review) indicated that a focus on STEM courses would improve the data analysis since the grade distributions in STEM courses tend to be wider than is typical for social science and humanities courses. And because early success in STEM courses is associated with degree progression and retention in STEM majors (Ackerman et al., 2013; De Urquidi et al., 2015), it is important for both individual AP students and institutions interested in promoting student success to understand how the use (or non-use) of

AP credit to satisfy introductory STEM courses in particular is associated with success in subsequent courses.

In summary, this study aimed to advance the domain knowledge regarding AP and academic success in college by focusing not only on any potential benefits of participating in the AP program or earning AP credit, but by estimating the effect of actually using it as a replacement for college courses. The College Board claims that students who pass AP exams are prepared to be successful in subsequent courses (n.d., *AP credit-granting recommendations*), but college faculty and academic advisors frequently recommend that students exercise caution and repeat AP credit at the college level (Sadler & Sonnert, 2010). Students who repeat credit are not able to realize one of the potential benefits of the AP program, reducing their time to degree, which has important financial consequences for students (Klopfenstein, 2010). This study provides empirical evidence regarding the effects of using AP credit so students and their advisors can make data-informed decisions. Thus, this study aims to answer the following research questions.

1. How do grades in STEM courses compare across three groups of students, those without AP credit for course pre-requisites, those who used AP credit for course pre-requisites, and those who had earned AP credit but chose not to use it?
2. Which factors predict the propensity to use earned AP credit as a pre-requisite for subsequent STEM courses?
    a) Are the findings consistent when focusing only on the use of AP Calculus credit as a pre-requisite for subsequent STEM courses?
3. What is the effect of using AP credit as a pre-requisite on subsequent course grades for students who choose to use their credit?
    a) To what extent does the effect vary across courses, and can course difficulty predict any of this variation?
    b) To what extent does the effect vary when focusing only on courses requiring AP Calculus as a pre-requisite?

CHAPTER 2. LITERATURE REVIEW

Introduction

The purpose of chapter two is to provide an overview of AP literature and identify important gaps in what is known about possible AP effects on student success in college. The chapter begins with an overview of AP history and growth, and evaluates three claims made by the College Board about AP effects: that AP students are more likely to be admitted to college, that they are more successful in college, and that they will graduate from college more quickly than non-AP students. It will then introduce the phenomenon of AP students repeating earned AP credit in college rather than using it exclusively for placement into advanced courses. Finally, it will explain three major gaps in AP literature: 1) understanding effects related to the use of AP credit, which can be addressed by including students who repeat credit; 2) accounting for and exploring variations in AP effects across subjects or courses, which can be addressed with the use of multilevel modeling; and 3) accounting for non-random assignment into AP groups in order to identify causal effects, which can be addressed with the use of propensity score modeling.

Early history of the Advanced Placement program

The College Board, a non-profit organization that currently administers the Advanced Placement program, was first established in 1900 as an association of elite colleges with the purpose of writing, administering, and scoring subject matter exams that were used for college admission purposes (Lacy, 2010). In 1952, a group of college and preparatory school faculty published the *General Education in School and College* report, which aimed to strengthen the connection between secondary and postsecondary education (Ewing, et al., 2010). The authors argued that it was a waste of time for gifted students to repeat the same material in general

education subjects such as English, history, and science (Lacy, 2010). At the same time, an initiative known as the Kenyon Plan aimed to address this concern by training high school teachers across the country to provide a college-level education to gifted secondary students, who then demonstrated their learning via end-of-year achievement tests (Lacy, 2010). In the 1955-56 academic year, the College Board assumed control of the Kenyon Plan (Lacy, 2010). The first Advanced Placement (AP) program exams, a new name for the Kenyon Plan achievement tests, were administered in 1956 (Drew, 2011). At that time, 104 high schools, 130 colleges, and 1,229 students participated in the AP program (Flowers, 2008).

Initially, the College Board created the exams, but individual faculty at the student's university reviewed and graded them, so the faculty had direct control over whether or not students were deemed ready for advanced coursework at their institution (Casserly, 1986; Lichten, 2000). As the AP program grew, college faculty deferred to the judgment of College Board exam readers, and credit would be automatically awarded by the University Registrar depending on the exam score (Casserly, 1986). The College Board's role in developing high school curricula for the AP program has also grown over time. At first, defining the AP curriculum was entirely up to individual high school teachers (Drew, 2011); today the College Board defines standards for each AP subject, and high school teachers submit their curricular plans for College Board review and approval (McCoy et al., 2020). In partnership with college faculty, the College Board conducts periodic studies of college curricula to ensure continued alignment between the content of introductory college courses and the high school AP courses that are meant to be equivalent (Ewing et al., 2010). College faculty are responsible for setting institutional policies around awarding credit, advanced placement, or both based on AP exam scores (Hansen et al., 2006), but the original

focus on the program was placing students into advanced courses so they could continue studying a subject without the need to repeat introductory material (Casserly, 1986; Lichten, 2000).

The Advanced Placement program's growing influence on American education

Since its beginnings in the 1950s, the AP program has grown substantially in the number of subjects offered as well as in the number of participating high schools and students (Ackerman, et al., 2013). Over the first decade of the 21st century, the number of students taking exams and the number of exams taken each increased approximately ten percent every year (Murphy & Dodd, 2009). This growth corresponded with a national movement towards a more rigorous, standards-based high school curricula (Judson, 2017). The high school graduating class of 2019 included 1.25 million students from over 22,000 high schools who took at least one AP exam out of the 38 exams available (College Board, n.d., *AP program results: Class of 2019*).

This substantial and sustained growth is due to the conventional wisdom that the AP program is excellent and worthy of support (Sadler & Tai, 2007). The AP program is "viewed as an incontrovertible indicator of educational excellence by educators and politicians alike" (Sadler, 2010a, p. 3), and there is a widespread belief that AP participation saves students time and money on college (Klopfenstein, 2010). A College Board report states that "one of the fundamental underpinnings of the AP Program is that students who perform well on AP examinations will be successful in college" (Morgan & Klaric, 2007, p. 1). Evidence of this conventional wisdom is found in the extensive public support for AP that began in the 1980s (Lacy, 2010). The United States Department of Education has promoted the AP program by providing grants to states that subsidize examination fees for low-income students in order to increase AP participation (U.S. Department of Education, 2006). The Education Commission of the States recommends that all states mandate a minimum number of AP courses offered at every high school (Klopfenstein &

Thomas, 2009), and legislation mandating the awarding of college credit for AP exam scores of 3 or higher has been enacted in 22 states (College Board, n.d., *State and systemwide AP credit and placement policies*). In the 2009 American Recovery & Reinvestment Act, the U.S. Department of Education highlighted the AP program as a means to promote rigorous standards and improve college outcomes (Chajewski et al., 2011). Both President Bush and President Obama included the AP program as part of their STEM education policies, calling for nearly twice the number of AP teachers in STEM subjects (Sadler, 2010a), and making a connection between AP STEM participation and greater numbers of students leaving college prepared to enter STEM careers (Tai et al., 2010).

College admissions offices have also accepted the "conventional wisdom" about the AP program to a large extent, with AP participation heavily weighted in admissions decisions through multiple mechanisms (Geiser & Santelices, 2004; Klopfenstein & Thomas, 2009). Colleges rely on AP as a measure of a student's readiness for college in part because both the curriculum and the evaluation of student achievement are standardized across the country (Conley, 2007). The emphasis on AP in college admissions in turn bolsters support for AP among parents and educational policymakers, who have pressured high schools to offer more AP courses (Geiser & Santelices, 2004). High schools are now rated and ranked, and the extent and success of a school's AP program contributes to those ratings (Duffet & Farkas, 2009; Judson, 2017). For example, the Challenge Index, published annually in *Newsweek*, ranks high schools by the ratio of AP tests taken by the school's students to the number of their graduating seniors (Duffet & Farkas, 2009; Shaw et al., 2013). The Fordham Institute surveyed high school AP teachers about possible explanations for the growth of the AP program; 90% of teachers agreed that students were enrolling in AP classes to improve their college applications, and 76% agreed that schools were

adding AP courses with the goal of improving their rankings and reputations, even though only 17% thought it was a good idea to rank schools based on AP participation rates (Duffet & Farkas, 2009).

Evaluating the benefits of the AP program

While there may be widespread belief that the AP program is beneficial, scholars prefer to take a more empirical approach to assessing the experiences and outcomes of AP students (Sadler, 2010a). The College Board touts three main benefits for students who participate in the AP program: 1) that they are more likely to be admitted to college, 2) that they will be more successful in college, and 3) that they will save time and money in pursuit of a college degree (College Board, n.d., *Benefits of AP*). Across the AP literature, there seems to be a consensus in support of the first claim, wide disagreement on the second, and insufficient evidence about the third.

Are AP students more likely to be admitted to college?

The use of AP in admissions began in the 1980s, and only highly selective colleges considered AP as a factor in decisions (Geiser & Santelices, 2004). National surveys of college admissions officers that ask about the importance of various factors did not start including any mention of AP until 2000; historically, AP was a tool for placement in college courses rather than linked to college admission (Shaw et al., 2013), so AP participation could only be seen as an indicator of a student's plans to enroll in college. Two studies (Chajewski et al., 2011; Wyatt et al., 2015) found that AP students had a significantly higher probability of attending a four-year college than non-AP students, but neither study made a case that AP participation caused (or even preceded) intent to enroll in college.

By 2004, nearly all selective colleges considered AP in admissions decisions (Geiser & Santelices, 2004), and at some institutions, students may even be penalized for not taking an AP course if it was available at their high school (Shaw et al., 2013). AP course participation rather than performance on AP exams is typically considered in college admissions because it is common for students to take AP courses during their senior year, and exam scores are not available until well after students have been admitted to college (Ackerman et al., 2013; Klopfenstein & Thomas, 2010). AP participation factors into college admissions decisions through different mechanisms: as a qualitative marker for a rigorous high school curriculum, as a quantitative metric (e.g., awarding admissions points for every AP course taken, adding points to AP courses when calculating high school GPA), or as a means of comparing students based on how many AP courses they took relative to the number of courses available at their school (Ackerman et al., 2013, Geiser & Santelices, 2004; Klopfenstein & Thomas, 2010; Shaw et al., 2013). While there is not empirical evidence that AP participation affects student intent to enroll in college, there is no question that AP participation is a factor in admissions decisions; therefore if a student's goal is to attend a selective institution, it is clear that AP participation will increase their likelihood of admission.

Do AP students experience better college outcomes?

The second possible benefit of the AP program is more challenging to evaluate. There is certainly a great deal of research showing that AP students out-perform non-AP students in college, in terms of attending more selective institutions (Mattern et al., 2009), earning higher grades in individual college courses (Breland & Oltman, 2001; Casserly, 1986; Godfrey & Beard, 2016; Wyatt et al., 2018), earning higher GPAs in their first year of college or after four years (Ackerman et al., 2013; Duffy, 2010; Mattern et al., 2009; Scott et al., 2010), returning to college after the first year (Duffy, 2010; Mattern et al., 2009), STEM degree persistence and attainment (Morgan &

Klaric, 2007; Shaw & Barbuti, 2010), college graduation rates (Ackerman, et al., 2013; Dougherty et al., 2006; Duffy, 2010; Morgan & Klaric, 2007), and post-college income (Flowers, 2008). Several of the studies cited above are published by the College Board, which regularly attempts to assess the extent to which AP exam performance is a valid substitute for completing the equivalent college course; these validity studies do not attempt to establish causation (Ewing, 2006). However, the public interprets the link between AP and student outcomes as causal even if studies are only correlational (Klopfenstein & Thomas, 2009).

From the early days of the AP program, scholars investigating AP student outcomes were aware that direct comparisons between AP and non-AP students were inadequate (Bergeson, 1967; Burnham & Hewitt, 1971), because the two groups are different in important ways that affect all the outcomes listed above. AP students are more likely than non-AP students to have well-educated parents (Duffy, 2010; Geiser & Santelices, 2004; Sadler, 2010b), higher family income (Duffy, 2010), and a better academic preparation for high school (Clark et al., 2012; Warne, 2017). Additionally, AP students score higher than non-AP students on other indicators of academic achievement that are measured before or concurrently with AP participation, such as state high school mathematics and reading assessment scores (Godfrey et al., 2014), high school grades, and SAT or ACT scores (Chajewski et al., 2011; Duffy, 2010; Mattern et al., 2009; Patterson & Ewing, 2013; Sadler & Tai, 2007; Warne, 2017). Along with these important pre-college differences between AP and non-AP students, there are potentially confounding factors related to the high school courses students take. For example, AP courses are more prevalent in wealthier communities (Sadler & Sonnert, 2010; Warne, 2017), and in schools that emphasize college preparation (Dougherty et al., 2006; Geiser & Santelices, 2004; Sadler & Sonnert, 2010; Shaw et al., 2013). AP courses also typically have fewer students and fewer classroom management

challenges than non-AP courses, and are taught by more experienced high school teachers with greater subject matter knowledge (Klopfenstein & Thomas, 2009; Sadler, 2010b). The central challenge with comparing AP and non-AP students is that AP participation is voluntary, and students who are highly motivated self-select into AP courses and exams (Dougherty et al., 2006; Sadler, 2010a). Simple comparisons of student college outcomes do not, therefore, provide convincing evidence that AP participation itself caused the higher achievement demonstrated by AP students.

Both College Board and independent AP scholars have attempted to account for differences between AP and non-AP students by including covariates in their analyses. Covariates that are frequently controlled for in AP studies include measures of academic achievement such as high school GPA, SAT or ACT scores, and AP exam scores, as well as student demographic characteristics such as gender, race or ethnicity, and various measures of parent education and family income (Ackerman et al, 2013; Burns et al., 2019; Dougherty et al., 2006; Duffy, 2010; Geiser & Santelices, 2004; Klopfenstein & Thomas, 2009; Mattern et al., 2009; Morgan & Klaric, 2007; Sadler & Tai, 2007; Shaw & Barbuti, 2010; Wyatt et al., 2015). AP studies also sometimes account for school-related variables at the student level, such as whether the high school is public or private (Sadler & Tai, 2007), or the strength of the curriculum offered (Chajewski et al., 2011; Geiser & Santelices, 2004). Klopfenstein & Thomas (2009) included all of the following school-related variables: the percent of students receiving free or reduced-price lunch, the percentage of students taking college entrance exams, the student/teacher ratio, school size, and the percent of inexperienced teachers on the faculty. After attempting to control for differences between AP and non-AP students, scholars typically have found a small but still positive relationship with multiple college outcomes.

One important distinction to make when reviewing AP research is whether the independent variable represents participation in the AP program (i.e., taking courses or exams) or performance on AP exams, as measured by exam scores which range from 1 to 5 (Ackerman et al., 2013). For example, Ackerman et al. (2013) found a monotonically increasing relationship between the number of AP exams completed and first-year college GPA (participation), but the relationship was stronger if they only plotted exams on which students scored a 4 or higher (performance). Similarly, Geiser and Santelices (2004) found that AP participation itself contributed almost nothing to a prediction of second-year college GPA, but that AP exam scores were one of the best predictors of that outcome metric.

Given that distinction, it may not be surprising that studies found no relationship between AP participation and college GPA or retention (Duffy, 2010; Klopfenstein & Thomas, 2009). After controlling for SAT Math scores, Burkholder and Wieman (2019) found no difference between students who had taken AP physics and those who had not on their final grades in the equivalent college physics course. In fact, students who do not score a 3 or higher on any completed AP exams actually earn slightly lower first-year college GPAs than non-AP students (Mattern et al., 2009; Wyatt et al., 2015).

However, when AP performance is the independent variable rather than AP participation, results show a more consistent advantage for successful AP students over non-AP peers. Students who score 3 or higher on AP exams earn higher first-year GPAs than non-AP students (Godfrey et al., 2014; Mattern et al., 2009; Wyatt et al., 2015), and outcomes on course grades and first-year GPA are even stronger for students who scored 4s or 5s on their AP exams (Morgan & Klaric, 2007; Sadler & Tai, 2007). Interestingly, both AP participation and AP performance have been shown to have positive relationships with college retention and graduation, although the magnitude

of the relationship varies across studies. After controlling for measures of prior academic achievement, AP students had higher retention rates (Mattern et al., 2009), graduation rates (Mattern et al., 2009; Wyatt et al., 2015), and greater persistence in STEM majors (Shaw & Barbuti, 2010) than non-AP students. Still, students with exam scores of 3 or higher have even better retention and graduation rate outcomes than those whose scores were all less than 3 (Mattern et al., 2009; Wyatt et al., 2015).

Even when studies report statistically significant differences for AP students, effect sizes are typically small. For example, after including other control variables in a regression analysis, adding AP exam scores increased the variance explained in first-year college GPA by 6.6% (Ackerman et al., 2013), and in second-year college GPA by 1.4% (Geiser & Santelices, 2004). Mattern et al. (2009) investigated the relationship between exam scores of 3 or higher across four AP subjects on both first-year GPA and college retention; effect sizes ranged from $d = 0.13$ to 0.21 for GPA, and from $d = 0.24$ to 0.42 for retention (Mattern et al., 2009). Similarly, Wyatt et al. (2015) estimated that first-year GPAs for students with at least one AP exam score of 3 or higher were 0.15 points higher than those of non-AP students on a 4-point scale; they also found that the same group of students had an approximately 11% higher probability of graduating in four years than their non-AP peers.

To summarize what is known about the second possible benefit of the AP program, better college outcomes, AP participation alone does not seem to have much of an effect on grade-related college outcomes after controlling for covariates. There may be a relationship between participation and longer-term outcomes such as retention and graduation, but it is also possible that relationship could be explained by unmeasured characteristics such as student motivation. Students who achieve high levels of performance on AP exams do seem to outperform non-AP

students across multiple college outcomes, although as Warne (2017) concluded, reported effect sizes tend to be much lower in studies that attempt to control for differences between AP and non-AP students than in studies that do not.

Do AP students reduce their time to degree?

It is challenging to evaluate the College Board's second claim, that AP students have better outcomes in college, because there have been so many studies with varying results; the third claim is difficult to evaluate mostly because there has been so little AP research related to the time it takes students to earn college degrees (Klopfenstein, 2010). The College Board maintains that the AP program helps students save time and money in college, because credits earned based on AP exam scores allow students to graduate earlier (College Board, n.d., *Benefits of AP*). Whether or not this is true for some students, the limited research about this topic does not indicate that most AP students benefit in this way.

Once again, AP participation is not a good predictor of positive outcomes, as AP course-taking alone is not associated with time to degree at all (Godfrey et al., 2014; Klopfenstein, 2010). Studies that consider AP performance report mixed results. Within the University of California system, growth in the number of AP credits awarded to incoming students who passed their AP exams was not followed by a reduction in the average time to degree (Eykamp, 2006). Similarly, passing one exam or earning higher exam scores had no effect on the likelihood of graduating from college in four or five years (Godfrey et al., 2014; Klopfenstein, 2010). One study did find that earning credit based on at least one AP exam was a significant predictor of reduced time to degree, but earning more credits had no additional effect (Burns et al., 2019). Smith et al. (2017) found that earning credit had a small effect on the likelihood of graduating in four years rather than five or six, but only if the student's exam scores were high enough to earn credit at the college where

they enrolled (i.e., students who earned 3s did not experience any benefits if their institutions required 4s). Overall, the available evidence does not support the College Board's claim that AP students save time and money because they graduate faster.

Warne at al. (2015) proposed two explanations for why the AP program does not appear to promote faster graduation: first, most students do not earn enough AP credit to meaningfully shorten their time in college. A national study of how students used AP credit found that only one out of every five students with ten AP credits graduates one term early (Evans, 2019). Students who earn enough AP credit to enter college as sophomores (approximately thirty credits) are more likely to graduate in three years than non-AP students, but only a very small group of students has the opportunity to enroll in, and successfully prepare for, that number of AP exams (Klopfenstein, 2010). The second possible explanation that Warne et al. (2015) proposed for why AP students do not graduate faster is that students may choose to repeat the corresponding introductory courses in college even after earning AP credit, which negates any potential opportunity to reduce their time to degree.

Whether and why students repeat AP credit in college

Over a quarter of introductory college calculus students earned a 3 or higher on their AP Calculus exam but enrolled in the course again in college (Bressoud et al., 2013; Sadler & Sonnert, 2018). Why might students choose to retake a course after earning AP credit? Students may decide to retake courses in order to earn higher grades (National Research Council, 2002) or to strengthen their mastery of the material (Sadler & Sonnert, 2010). Indeed, repeated exposure could be beneficial across multiple subjects. For example, students who completed advanced high school physics (AP or honors) and then took the equivalent introductory course again in college earned higher grades than students who encountered the material for the first time at the college level

(Burkholder & Wieman, 2019), and students who repeated AP Language & Composition by taking an introductory college composition course were judged to have stronger writing skills in a sophomore-level course than those who used their AP credit to place out of the introductory course (Hansen et al., 2006).

While students may choose to repeat AP credit for their own reasons, others are advised or even required to do so by college faculty, academic advisors, or other students (Sadler & Sonnert, 2010, 2018; Scott et al., 2010). Students repeating earned AP credit in introductory STEM courses reported that they did so because of advice from faculty and advisors, or because they did not pass a placement exam required by their department in addition to the AP exam score (Sadler & Tai, 2007).

Officially, an institution's faculty has the authority to decide whether to grant placement or credit for AP exam scores, and most institutions do award AP credit (Ackerman et al., 2013; National Research Council, 2002). The College Board encourages institutions to grant credit for scores of 3 or higher, but only about half of schools do so; most, however, will grant credit for scores of 4 or 5 (National Research Council, 2002). Faculty at state institutions that are mandated to accept scores of 3 for credit can still choose to grant only elective credit and require higher scores for direct course equivalencies (Lichten, 2000). Institutions without state mandates can limit the use of AP credit to an even greater degree. For example, in 2002, Harvard announced it would only grant credit for scores of 5 because their faculty had found that students who earned credit with 4s were performing well below class norms in subsequent courses (Duffy, 2010). And in 2007, MIT stopped accepting AP credit for biology altogether, because their faculty thought that even students who scored 5s on their AP Biology exams did not have the problem-solving skills needed to succeed at the college level (Drew, 2011). Other selective institutions restrict the number of AP

subject areas that are eligible for credit or cap the total amount of credit a student can be granted (Conger et al., 2021).

University faculty who recommend that students repeat AP credit in college are concerned that AP courses are not equivalent to college-level courses in depth or rigor (National Research Council, 2002). Biology, chemistry, and physics faculty panels concluded that most students would benefit from retaking AP courses at the college level, although the mathematics faculty panel disagreed, saying that most AP students who earned credit should use it and move on to the next course (National Research Council, 2002). College courses go at a faster pace and cover content that is not included in typical AP courses (Conley, 2007; Eykamp, 2006). Most high school science laboratory facilities are not comparable to those found in colleges, so AP lab science courses cannot approximate the college experience (Drew, 2011). Beyond content coverage, faculty argue that AP courses are focused more on learning procedures than on engaging deeply with the concepts of a discipline (Hansen et al., 2006; National Research Council, 2002; Wade et al., 2016). One study tested the assumption that AP courses are not equivalent to college courses by comparing AP physics exam scores to scores on a commonly used physics conceptual test (FCME) and college course grades. There was no significant correlation between AP exam scores and grades, and FCME scores were better predicted by SAT Math scores than AP exam scores; the authors concluded that the benefits of taking an AP physics course were largely limited to familiarity with physics terms (Burkholder & Wieman, 2019). One counterpoint to the skepticism of college faculty regarding the equivalence of AP courses to college courses is that the strongest and best-prepared students may earn AP credit and move ahead to the next subsequent course; faculty who teach introductory classes only encounter weaker or less confident AP students and therefore may underestimate the quality of the AP curriculum (Sadler & Sonnert, 2018).

Faculty resistance to students using AP credit to graduate faster may go beyond straightforward concerns about whether those students are prepared for more advanced coursework. An Association of American Colleges & Universities survey of 451 academic affairs administrators indicated that faculty also see the growing numbers of students earning college credit for work completed in high school as a threat to their traditional authority to decide what counts as college-level learning (Johnstone & Del Genio, 2001). Faculty are concerned that students who bring in AP credit will use it to satisfy general education requirements and never take another class in their subject area (Morgan & Klaric, 2007; National Research Council, 2002). This worry is in part about wanting to attract gifted students to their majors (Sadler & Sonnert, 2010), but also about budgets, since departments may be dependent on demand for introductory courses to justify their funding (Johnstone & Del Genio, 2001). In one study of AP calculus students, 39% of those who passed the AP calculus exam never enrolled in any mathematics course in college (Sadler & Sonnert, 2018).

In summary, the evidence supporting the College Board's claims about the three major benefits of the AP program is mixed. There is general agreement that participating in AP courses will help a student be admitted to college. The second claim, that AP students are more successful in college (whether success is measured by grades or longer-term metrics such as retention and graduation) has some support, but questions remain due to potential environmental, academic, and/or motivational differences between students who seek AP credit and those who do not. This present study contributes to the effort to answer these questions. The third claim, that AP students can save time and money in college, needs more investigation; this study adds to the limited research base on how students choose to use earned AP credit once in college.

Why additional research on Advanced Placement is needed

It is clear that there are diverging views about the effects of AP participation or performance on college outcomes. The conventional wisdom accepted by the public is that students benefit from participating in the AP program (Klopfenstein & Thomas, 2009; Sadler, 2010a; Sadler & Tai, 2007), but college faculty and researchers who study AP independent of the College Board are skeptical about those benefits (National Research Council, 2002; Johnstone & Del Genio, 2001; Sadler, 2010a). Regardless of the accuracy of the College Board's claims, a clear understanding of how participation in the AP program is associated with future student outcomes is important for policy and practice at both the college and high school levels. First, at the college level, although critics may question the role AP plays in admission decisions, it is evident that AP participation is indeed heavily weighted in college admissions through multiple mechanisms (Geiser & Santelices, 2004; Klopfenstein & Thomas, 2010). If AP participation is not a useful indicator of future college success, then perhaps it should not be factored into admissions decisions. Additionally, depending on institutional policies, AP exam results are used to place students directly in advanced college courses, as originally intended (Lichten, 2000), and appropriate placement can have meaningful consequences for students. A student's first grade in college mathematics, for example, is correlated with student graduation, so the choice of whether to accept AP credit or repeat it matters for student success (De Urquidi et al., 2015).

Second, a better understanding of AP effects could influence secondary education policy. There are costs associated with promoting AP courses at the high school level such as the direct costs of paying for teacher training, course materials, and exam fees, but also the indirect costs including larger class sizes and less-experienced teachers for non-AP students (Clark et al., 2012; Klopfenstein & Thomas, 2010). The growth of the AP program has been expensive for secondary schools; spending on AP courses has outpaced spending on regular or remedial courses on a per-

student basis (Duffet & Farkas, 2009). If there truly is no causal effect of AP participation or performance, then it does not make sense for the public to subsidize the AP program (Klopfenstein, 2010). Additionally, there has been an effort in recent years to expand access to the AP program to a wider population of students, including those who have historically be underrepresented in advanced classes; if there are no independent effects of the AP program, then broadening access will not improve outcomes for these students (Warne, 2017).

Third, we need to ask new questions about AP effects. Prior research has focused on outcomes related to participating in the AP program (e.g., Chajewski et al., 2011) or performing well on AP exams (e.g., Mattern et al., 2009), but much less research has been done regarding the use of AP credit in college (Evans, 2019). At the K-12 level it makes sense to investigate whether or how students benefit from choosing to participate in AP, but at the college level the choices are different. Faculty and institutions set policies regarding the granting of AP credit (National Research Council, 2002), and students and their academic advisors make decisions about whether to use earned AP credit (Sadler & Sonnert, 2018); both groups would benefit from an understanding of student outcomes related to the use of AP credit. In summary, we need a more complete understanding of a range of AP effects to address diverging views of the AP program, to support future policy decisions at the K-12 level, and to inform choices made at the college level, both by institutions and by individual students.

Remaining gaps in Advance Placement literature

In order to develop a more complete understanding of AP effects, it is crucial to address three knowledge gaps in the current AP literature base: 1) understanding effects related to the use of AP credit by including students who repeat credit, 2) a systematic investigation of the variation of AP effects across subjects or courses, and 3) the ability to draw causal inferences about AP

effects on college success. First, few studies investigated outcomes for students who choose to repeat AP courses in college even when they earned college credit based on their exam scores. If repeating courses is a reason that AP students do not typically reduce their time to graduation, then the implications of this choice need to be better understood. Second, independent research is needed on the possible variability of AP effects across subjects. There is a wide variety of AP subjects, and each requires different academic skills and knowledge; assuming the effect is the same across subjects oversimplifies a complicated question (Warne, 2017). The use of multilevel modeling to account for the type of clustering effects frequently observed in educational data, such as how students are clustered in AP subject areas, would allow for an exploration of how AP effects may vary across different college courses (Warne, 2017). Finally, the evident academic and socioeconomic differences between students who self-select into AP programs and those who do not make drawing causal inferences about AP effects challenging. (Dougherty et al., 2006; Sadler, 2010a). In non-experimental studies, variables such as prior academic achievement and motivation directly influence both the outcome of interest (such as college success) and selection into treatment (such as AP participation) so including covariates in regression analyses is not sufficient to identify treatment effects (Li et al., 2013); propensity score analysis is one way to account for the non-random assignment to AP groups before arguing that AP has any causal relationship to student outcomes (McCaffrey et al., 2013). Next, I will discuss each of these three gaps in the AP literature in greater detail, describing relevant prior studies and considering the next steps that should be taken to address each of the gaps.

Including AP "repeaters" to understand the effects of using AP credit

There is little research on outcomes related to the use of AP credit. Two studies have investigated how students choose to use AP credit, for example whether they use it to graduate

faster, take additional advanced courses, or complete a second major (Evans, 2019; Eykamp, 2006). A third study compared outcomes related to time to degree of students just above and below raw score thresholds associated with official AP exam scores with regression discontinuity analysis; Smith et al. (2017) found that students benefitted from earning high exam scores only if their colleges granted credit for those scores. This implies that the college credit earned, rather than the knowledge gained from the AP course, is the more important factor in predicting faster graduation.

Additionally, three studies examined outcomes related to the use of AP credit by comparing students who did use their credit to others who chose to repeat AP credit in college. First, Murphy & Dodd (2009) designed their study to include a separate group of AP "repeaters," but decided that sample sizes were too small to analyze in all subjects except Calculus. They found no significant difference in first-year GPAs or mathematics GPAs between students who repeated first-semester calculus and a matched group of non-AP students (Murphy & Dodd, 2009). Second, De Urquidi et al. (2015) studied outcomes of AP calculus students, comparing grades in the first college calculus taken depending on whether students scored 4s or 5s on either of the two AP calculus exams (AB, which covers first semester calculus, or BC, which covers a full year of college calculus). They found no advantage to students who chose to repeat earned credit, although results varied somewhat by first mathematics course taken and SAT Math scores (De Urquidi, et al., 2015). Third, Hansen et al. (2006) evaluated writing samples from three groups: students who used AP credit as the pre-requisite for a sophomore-level English course, non-AP students who took a first-year composition course prior to the sophomore-level course, and students who earned AP credit and also took the composition course. They considered the first two groups to be equivalent, but rated the third group as superior to the others; however, the authors did not control for any student background characteristics, such as high school grades or test scores (Hansen et

al., 2006). In summary, the effect of repeating AP credit is inconclusive; two studies found no significant differences associated with repeating earned AP credit, while the third study that did report better outcomes for AP repeaters failed to include common covariates, thereby possibly confounding the results of the study.

Other than these three studies, AP scholars have often chosen not to include AP repeaters as a separate group. But given that at least some students do choose to repeat earned AP credit, studies that do not evaluate AP repeaters separately must either exclude them from the study altogether (e.g., Patterson & Ewing, 2013), or include them with AP students who take introductory courses because they did not earn credit by exam (e.g., Sadler & Tai, 2007). The first option is the most common way to account for repeaters: excluding this group of students completely when comparing AP and non-AP student outcomes. For example, the College Board conducts studies to evaluate the performance of students who use AP credit when they enroll in the next subsequent course, so it would not make sense to include students who chose not to use their credit (Casserly, 1986; Patterson & Ewing, 2013). The method sections of such College Board studies typically explain that students are only included in the AP group if they earned credit by exam, and took the intermediate course in a subject without previously taking the introductory course at the college level (for example, Morgan & Klaric, 2007; Wyatt et al., 2018).

Unlike these validity studies that compare outcomes in subsequent courses, other AP studies have investigated outcomes in introductory college courses, and choose the second option described above. These studies combine AP repeaters with non-AP students and AP students who did not earn college credit; because they focus on introductory courses, they exclude the group of students who earned AP credit and used it to move directly into the next subsequent course. Two studies compared grades in introductory college science courses and included students with AP

exam scores of 3, 4, and 5 in their sample; presumably some of those students had earned credit and were repeating the introductory course, but the authors did not mention this issue at all (Burkholder & Wieman, 2019; Sadler & Sonnert, 2010). Authors of a third study took a similar approach, but in this case, the inclusion of AP repeaters was an intentional part of the study design. Specifically, Sadler & Tai (2007) chose to investigate outcomes in introductory biology, chemistry, and physics courses rather than intermediate courses in part because prior studies had ignored the group of students who earned AP credit but chose not to use it. They did not separate out students who had earned credit from those who did not, but instead just used AP exam score as a variable in their regression model predicting course grades (Sadler & Tai, 2007).

Unlike studies of AP performance, studies investigating AP participation only do not differentiate between students who earn credit and those who do not, so there is no need to consider repeaters. For example, Hargrove et al. (2008) grouped students by high school experience (AP course, dual credit course, or neither), and Klopfenstein and Thomas (2009) counted the number of AP courses students took in high school; both studies were focused on AP participation rather than performance. Additionally, studies that consider more general outcomes (e.g., retention, graduation) rather than course performance typically do not mention AP repeaters; students may be grouped by AP exam score but not by the courses they chose to take in college (Ackerman et al., 2013; Dougherty et al., 2006).

Unfortunately, there are potential problems with choosing to include AP repeaters in a sample without accounting for them as a separate group. Geiser & Santelices (2004) observed that the phenomenon of repeating AP credit could confound analyses of first-year grades in two possible ways: first-year AP students could be earning higher grades due to repeating introductory courses, or earning lower grades because they were taking more challenging intermediate courses.

They chose to mitigate this potential problem by focusing on second-year grades rather than by differentiating AP repeaters from those who used their credit (Geiser & Santelices, 2004). Rather than simply ignoring the group of students who repeat AP credit, combining them with other groups, or choosing different outcome measures to avoid possible confounding effects, future AP studies should account for this group of students separately. Doing so will allow for an expanded understanding of AP effects, including outcomes associated with the use of AP credit at the college level.

Investigating variation in AP effects

The second major need for additional research on AP concerns the problem of considering AP as a single entity, when in fact there are currently 38 different AP courses and corresponding exams (College Board, n.d., *AP program results: Class of 2019*). It is common for studies investigating first-year GPA or graduation rates to identify students by average AP exam score regardless of AP subject (e.g., Ackerman et al., 2013; Geiser & Santelices, 2004), or to differentiate students by whether they had scored a 3 or higher on any individual AP exam (Dougherty et al., 2006; Wyatt et al., 2015). However, combining subjects in this way may mask differences among the various AP courses students take, and there is fairly extensive evidence that such differences do exist.

Evidence of variation in AP effects by subject area

Ackerman et al. (2013) mostly reported results in aggregate, but when they calculated the point-biserial correlation between whether a student earned credit in a given exam and their eventual first-year GPA, significant correlations across 24 exams ranged from nearly zero (r = 0.03) to a weak correlation (r = 0.23). Similarly, Klopfenstein and Thomas (2009) mostly studied

outcomes based on AP participation across subjects, but when they did find significant differences between AP and non-AP groups, they looked more closely to identify which subjects were driving that result. For example, they found the entire effect of AP on retention for Hispanic students was driven by participation in AP science courses, while for White students, AP Government accounted for all of the difference in first-year GPA between AP and non-AP students (Klopfenstein & Thomas, 2009). Warne et al. (2019) found that completion of AP Calculus did have a small but positive relationship with the likelihood of a student choosing a career in a STEM field, but completion of AP Statistics had no such relationship. The authors concluded that these results reaffirmed the need to study AP courses individually and not as a homogeneous program (Warne et al., 2019).

There are also studies that report results separately by subject; those that do so on a large scale are nearly all sponsored by the College Board. These include validity studies investigating course performance of AP students in subsequent courses (e.g., Godfrey & Beard, 2016; Patterson & Ewing, 2013; Wyatt et al., 2018), as well as studies that investigate broader outcomes such as overall subject GPA (Godfrey et al., 2014; Murphy & Dodd, 2009; Patterson et al., 2011), or first-year GPA, retention, and graduation (Hargrove et al., 2008; Mattern et al., 2009). Unsurprisingly, studies investigating these broader outcomes that encompass performance across all subject areas do not show much variation by AP exam taken. In general, the farther removed the outcome is from performance on a specific AP exam, the less variation is seen in results across subjects; overall, AP students have better outcomes than non-AP students on these broader measures such as GPA, retention, and graduation.

For example, Hargrove et al. (2008) compared first-year GPAs and four-year graduation rates of AP exam-takers and non-AP students separately by exam subject (i.e., AP Calculus

students compared to non-AP Calculus students, AP History students compared to non-AP History students); they found that AP students outperformed non-AP students in these non-subject-specific across to a similar extent, regardless of the AP exam they had taken. Mattern et al. (2009) reported AP Biology, Calculus, English, and History students who scored 3 or higher all outperformed non-AP students in both first-year GPA (effect sizes ranged from 0.13 to 0.21) and retention (effect sizes ranged from 0.24 to 0.42).

Subsequent course grades and overall subject GPA are the outcome measures most closely linked to performance on an AP exam in the relevant subject area, and those results do vary by subject. Table 1 below summarizes results based on some of the most frequently taken AP exams from eight College Board studies investigating one or both of these two outcomes. One important note to keep in mind: AP Biology, Chemistry, and Physics courses underwent large-scale curricular reform between 2012 to 2015, with the goal of a greater emphasis on scientific reasoning (McCoy et al., 2020). All of the studies described below included students who completed their AP courses prior to this curricular reform, but future studies may show different results based on the revised AP curricula.

Table 1: Summary of College Board Results by Subject

Subject	Study	Outcome Variable: Course Grade	Outcome Variable: First-Year Subject GPA
Biology	Dodd et al. (2002)	AP exam passers < non-AP in one year of study; other 3 years no difference	
	Godfrey & Beard (2016)	No difference	
	Morgan & Klaric (2007)	AP4/AP5 > non-AP; AP3 no difference	
	Murphy & Dodd (2009)	No difference	No difference
	Patterson & Ewing (2013)	No difference	
Calculus	Dodd et al. (2002)	AP exam passers > non-AP in all four years of the study	
	Godfrey & Beard (2016)	Calc AB students > non-AP	
	Godfrey et al. (2014)		AP exam scores were significant predictor of subject GPA (std. regr. coef. <0.1).
	Morgan & Klaric (2007)	AP3/AP4/AP5 > non-AP	
	Murphy & Dodd (2009)	AP exam > non-AP (small ES) for BC Calc; AB calc no difference	AP exam passers > non-AP (small ES) for both BC and AB Calculus
	Patterson & Ewing (2013)	AP > non-AP (mean difference on GPA scale was .192 for AB and .229 for BC)	
	Patterson et al. (2011)		AP > non-AP. Regr. coef: AP3=.196, AP4=.211, AP5=.361
Chemistry	Godfrey & Beard (2016)	AP > non-AP in one subsequent course, but AP < non-AP in another subsequent course	
	Morgan & Klaric (2007)	No difference	
	Murphy & Dodd (2009)	No difference	No difference
	Patterson & Ewing (2013)	AP > non-AP (mean difference on GPA scale was .318)	
Economics	Morgan & Klaric (2007)	AP4 < non-AP; AP3/AP5 no difference	
	Murphy & Dodd (2009)	No difference	AP exam passers > non-AP (small ES)
	Patterson & Ewing (2013)	No difference	
	Wyatt et al. (2018)	AP students earned significantly higher grades than non-AP	

Table 1 continued

English	Dodd et al. (2002)	AP exam passers > non-AP in 2 years of study; other 2 years no difference	
	Godfrey & Beard (2016)	AP > non-AP in 3 out of 4 courses; 4th course no difference	
	Godfrey et al. (2014)		AP exam scores were significant predictor of subject GPA (std. regr. coef. <0.1)
	Morgan & Klaric (2007)	AP3/AP4/AP5 > non-AP students	
	Murphy & Dodd (2009)	No difference	AP exam passers > non-AP (small ES)
	Patterson et al. (2011)		AP > non-AP. Regr. coef: AP3=.082, AP4=.112, AP5=.143
	Wyatt et al. (2018)	AP > non-AP	
History	Godfrey et al. (2014)		AP exam scores were significant predictor of subject GPA (std. regr. coef. for AP 4/5 = 0.23)
	Morgan & Klaric (2007)	AP4/AP5 > non-AP; AP3 no difference	
	Murphy & Dodd (2009)	No difference	AP exam passers > non-AP (small ES)
	Patterson & Ewing (2013)	No difference	
	Patterson et al. (2011)		AP > non-AP. Regr. coef: AP3=.160, AP4=.205, AP5=.304

1. Cells that are shaded indicate there was no attempt to account for pre-college differences between AP and non-AP students.
2. Murphy & Dodd (2009) used the same data set as Dodd et al. (2002), but matched students differently and included additional outcome variables.
3. "No difference" means that no significant difference was found between AP and non-AP students. All other outcomes reported were significant at $p < .05$.

Each of the eight College Board studies summarized in Table 1 included other subjects as well; none of the eight reported the exact same results across all AP subjects. Overall, results for AP Calculus and AP English were almost entirely positive, although effect sizes varied and there were a few non-significant results. The other subjects all had mixed results, from some studies showing very positive effects, several showing no effect, and a few even finding that AP students performed worse than their non-AP peers. These results support the contention that AP subjects should not be assumed to be equivalent, and the fact that there are so many multi-subject studies

might seem to indicate that this need is already being addressed. However, there are two reasons that additional research is needed for a more complete understanding of AP effects.

First, all of these studies summarized above were sponsored by the College Board. Sadler (2010a) and Warne (2017) have both argued that research conducted by independent scholars and subject to peer review is needed to complement the College Board's internally produced research reports. Secondly, almost no studies of course performance or subject GPA consider courses outside the cognate department that grants AP credit even though students use their AP credit as pre-requisites for courses outside the same cognate department as well. Patterson and Ewing (2013) included engineering courses in their list of subsequent courses for AP Chemistry, Physics, and Computer Science; for AP Biology they included courses in animal science and anatomy and physiology (if they were offered outside the biology department); similarly, Patterson et al. (2011) included engineering course grades as an outcome for students who had completed AP Calculus or any AP natural science course. However, none of the other studies summarized in Table 1 included courses outside the cognate department.

The use of multilevel modeling to account for differences across groups

If there is meaningful variation in AP effects across courses or subjects, then cross-sectional multilevel modeling is a useful approach because it allows for an exploration of how effects vary across groups; it is also helpful because it accommodates the naturally occurring cluster effects of education data (Warne, 2017). In higher education environments, student learning often occurs in settings where cluster effects are organically created by cohorts, sections, instructions, or courses. For example, if the outcome variable is student course grades, cross-sectional multilevel modeling can account for differences in grading across instructors, courses, or institutions (Sadler & Sonnert, 2010). Research on AP that uses multiple regression relies on the

assumption that observations are independent, but education data are typically nested; violating the independence assumption with clustered data will underestimate standard errors and increase the likelihood of Type 1 errors (Raudenbush & Bryk, 2002; Warne, 2017). One of the questions that is answered with multilevel modeling is how much of the variation in outcomes is associated with different levels in the model (Raudenbush & Bryk, 2002). In AP research, students would be considered level one, and courses, high schools, or colleges could be treated as the level two cluster in cross-sectional models. The proportion of variance in the outcome that is found between groups, sometimes referred to as the cluster effect, is identified by the intraclass correlation coefficient (or ICC), which is the ratio of group-level variance (τ^2) to total variance ($\tau^2 + \sigma^2$), where σ^2 represents within-group variation (Raudenbush & Bryk, 2002). Even when the ICC is as low as 0.05, accounting for the clustering effect via the use of multilevel modeling protects against Type I errors (Huang, 2018).

A second question that the use of cross-sectional multilevel modeling can answer is directly related to the concern that AP effects have been shown to vary by subject (e.g., Patterson & Ewing, 2013). If there are group-level differences, then it follows that some group-level predictors could explain those differences; in two-level cross-sectional multilevel modeling, means-as-outcomes models use level two predictors to estimate the variation in cluster-level means (Raudenbush & Bryk, 2002). While little research has been conducted to explore variation in grades due to course characteristics, course difficulty level is a possible avenue for exploration as a level two predictor (Wladis et al., 2017).

Finally, cross-level interaction effects in cross-sectional multilevel models indicate if the variation in the effect of a level one predictor across courses can be explained by any level two predictors (Raudenbush & Bryk, 2002). For example, if there is an effect of AP credit usage on

college grades, and if that effect is different across different subjects or courses, then it is possible that course difficulty level (or some other course-related predictor) can explain some of that variation. Overall, multilevel modeling is a useful approach for AP studies because it allows a model to include differences beyond the student level. Cross-sectional multilevel modeling that accounts for group effects and cross-level interactions can provide a more precise estimate of AP effects, as well as allowing for an exploration of how the effect varies across groups (Warne, 2017).

Cross-sectional multilevel modeling in prior AP research

At least six studies have employed cross-sectional multilevel modeling to investigate AP effects, with a range of outcomes at both the high school and college levels. Table 2 below includes information about the methods and model information shared by each of these six studies.

Table 2: Description of AP Studies Using Cross-Sectional Multilevel Modeling

Study	Outcome Variable	*N/K* at each level	Level 2/3 Predictors	Results Reported Beyond Student Level
Dougherty et al., 2006	Probability of college graduation	L1 (student): 54,556 L2 (high school): *k* not shared but all Texas HS with >15 students in each ethnicity group	L2: Percent low-income, average 8th grade math score, percent taking an AP course, percentage by ethnicity	L2 parameter estimates and cross-level interaction effect estimates
McKillip & Rawls, 2013	SAT Scores	L1 (student): calculus=9,499; chemistry=10,730; English=51,175 L2 (high school): *k* not shared but national sample	None	ICC Random intercepts only; fixed slopes
Patterson et al., 2011	Subject GPA (multiple subjects)	L1 (student): range 13,214 – 115,324 L2A (high school): range 3,488 – 7,857 L2B (college): range 65-110	L2A: Urbanicity, size, public/ private, number of AP courses offered L2B: public/ private, size, selectivity	Reported percentage of variance explained at each level
Sadler & Sonnert, 2010	Intro science course grades	L1 (student): 4207 L2 (course/ instructor): 124 L3 (college): 55	None	None
Sadler & Sonnert, 2018	Intro calculus course grades	L1 (student): 6207 L2 (course): 216 L3 (college): 133	None	Reported percentage of variance explained at each level Random intercepts only; fixed slopes
Shaw et al., 2013	First-year GPA	L1 (student): 74,501 L2 (college): 125	None	ICC

As a whole, results from these studies support the choice to use cross-sectional multilevel modeling when investigating possible AP effects. First, most of the studies reported that between 8 and 23 percent of the variance in outcomes was attributable to cluster effects, such as high schools, colleges, or college courses, which indicates that multilevel modeling is an appropriate analytic method to account for nesting effects (Hedges & Hedberg, 2007). Specifically, across

three AP subject areas (calculus, chemistry, and English), McKillip and Rawls found that AP participation explained 22-23% of the variance in SAT scores; Sadler and Sonnert (2018) reported that 18% of the variance in introductory calculus grades could be attributed to either the course or college levels; Patterson et al. (2011) found that group-level effects accounted for over 10% of the variance in subject GPA across all eight subjects included in their study; and Shaw et al. (2013) reported an intraclass correlation of 0.836, indicating that over 8% of the variance in first-year GPA was due to college-level effects. Second, while Dougherty et al. (2006) did not specify the percentage of variance in the probability of college graduation found at the student or high school levels, they did report multiple significant cross-level interaction effects, indicating that the effects of AP-related student predictors varied based on high school factors such as average mathematics achievement levels and the percentage of low-income students enrolled. Researchers recommend the application of multilevel modeling to align with the nested data structure, and the potential cluster effects should not be ignored even when the ICC is low (Huang, 2018).

While multilevel modeling is an attractive methodological choice for AP studies, these models require complex data assumptions and relatively large sample sizes at each level. Of these six studies, three chose to cluster students by high school (Doughterty et al., 2006; McKillip & Rawls, 2013) or college (Shaw et al., 2013); Patterson et al. (2011) employed a cross-classified multilevel model in which students were clustered in both high schools and colleges. The other two (Sadler & Sonnert, 2010, 2018) chose to use three-level structures, with students clustered in courses, which were themselves clustered in universities. This latter choice seems questionable given that based on their studies' sample sizes, each university-level group would have only two instructors on average; one of the two studies (Sadler & Sonnert, 2018) reported that 47% of institutions had only one instructor, rendering the third level meaningless in those cases. Another

possible concern about the decisions made regarding group levels is that not all studies reported the number of groups, or the range of group sizes. For example, McKillip and Rawls (2013) drew their sample of almost 10,000 AP Calculus students from every high school in the country that offers AP Calculus in their study; they did not report the number of high school-level clusters, so there is no way to know the average number of students in each group. In contrast, while Dougherty et al. (2006) did not report the number of Texas high school groups in their study, they did specify that they only included high schools with at least 500 total students, and at least 15 students from whichever racial or ethnic group was serving as the independent variable.

In addition, one other conclusion from Table 2 above is that most multilevel AP studies have reported exclusively on student-level (or, level one) results. Few studies reported any details about main effects of level two predictors, or how those predictors explained any of the variation in the effects of level one predictors across groups; in fact, more than half did not have any level two or level three variables included in the model at all, and only one study (Dougherty et al., 2006) allowed level one predictors to randomly vary across groups. In general, the application of multilevel modeling in AP research seems to have been used more as a means of controlling for group-level differences rather than maximizing the benefit of this methodological approach by exploring or attempting to explain those group-level differences.

Allowing for causal inferences of AP effects

The third major need for additional research on the AP program is to expand on the limited number of studies that draw causal inferences from non-experimental data (Warne et al., 2015). As previously discussed, AP students are quite different from non-AP students in terms of their academic and family backgrounds, and issues of self-selection (choosing to take an AP course, or choosing to use earned AP credit) make it difficult to isolate the effects of the AP program (Clark

et al., 2012). Even after controlling for prior academic achievement and personal background characteristics, any estimates of AP effects are likely to be biased upwards by unobserved factors such as student motivation or work ethic variables (Clark et al., 2012). Propensity score modeling aims to reduce this selection bias by predicting selection into treatment groups (such as enrolling in an AP course) rather than directly predicting outcomes (McCaffrey et al., 2013).

Of course, the accuracy of the propensity score depends on the extent to which selection into treatment is based only on observed characteristics, but in this regard, propensity models have an advantage over standard regression approaches because they do not impose a linear functional form (Long et al., 2012). A typical regression model that attempts to predict an outcome (*Y*) based on a treatment condition (*D*), controlling for confounding variables (*X*), would simultaneously estimate the linear relationship of *D* and *X* on *Y*, whereas a propensity or matching estimator nonparametrically balances the confounding variables (*X*) across both treatment and control groups, such that the remaining differences in observed outcomes (*Y*) between groups can be attributed solely to the treatment (*D*) (Morgan & Winship, 2015). In AP studies, this means creating a comparison group of non-AP students that is more similar to the group of AP students; if students could be matched on all possible confounding variables, then the difference in outcomes between the two groups could be attributed to their AP status (Patterson & Ewing, 2013).

The goal of propensity score analysis is to identify treatment effects, but there are different possible treatment effects to consider. The average treatment effect (ATE) is an estimate of the effect of a given treatment across all individuals in a population, whereas the average treatment effect for the treated (ATT) is an estimate of the effect of the same treatment only on the subset of the population who typically experiences the treatment (Morgan & Winship, 2015). This distinction has meaningful differences for how results are interpreted. For example, if we want to

know the effect of earning AP credit on the outcomes of students who actually earn AP credit, then the ATT would be most relevant. A disadvantage to the ATT, however, is that it cannot be used to predict what the effect of earning AP credit would be if the program were expanded to new groups of students who are unlike the current treatment group (McCaffrey et al., 2013). There are a variety of ways to incorporate the general idea of accounting for propensity to be in a given treatment group, from simple matching techniques to far more sophisticated models. Studies investigating AP effects have taken three general approaches to the use of propensity modeling: stratified matching, more sophisticated matching techniques, and the use of propensity weights.

Propensity score analysis in prior AP research

Three AP studies employed stratified matching prior to comparing outcomes between AP and non-AP students. Both Dodd et al. (2002) and Murphy and Dodd (2009) matched students by stratifying high school rank into five categories and SAT total scores into 100-point increments, and then pairing AP and non-AP students who were in the same rank/test score grouping. Hargrove et al. (2008) used similar SAT score groupings but combined them with students' free or reduced lunch status rather than high school rank. There are two concerns with the methods employed by these studies. First, neither Dodd et al. (2002) nor Hargrove et al. (2008) provided any data about how successful the matching process was, in terms of how similar the groups became after matching or if they differed on other predictors of their outcome variable. Murphy & Dodd (2009) did note that 88% of their matched pairs differed by less than 60 points on the SAT, but that means 12% differed by more than 60 points, and even within the 88% of closer matches, there were likely to have been "matched" pairs who were not very similar. Along with the lack of evidence about the success of the matching process, there is an even greater concern with how Dodd et al. (2002) chose to match AP and non-AP students. Dodd et al. (2002) compared outcomes across three

groups: AP students who scored high enough on exams to earn college credit, AP students who did not earn credit, and non-AP students. Unfortunately, Dodd et al. (2002) chose to match the non-AP students to the lower-scoring group of AP students. It is perhaps not surprising then, that when Murphy & Dodd (2009) analyzed the same data set several years later but instead matched non-AP students to successful AP students, they found far fewer and smaller AP effects than the original study.

A more rigorous approach to matching starts with calculating propensity scores that can include far more variables than it would be possible to match on exact values; subjects are instead matched on their overall propensity to be part of the treatment group (Warne, 2017). There are multiple matching options, and several of these have been applied in studies of AP effects. In a study of the effect of AP Calculus and AP Statistics enrollment on STEM career interest, Warne et al. (2019) included pre-high school measures of mathematics and science achievement, as well as academic and career interests as predictors in a logistic regression model to create propensity scores; both AP and non-AP students were then stratified into five groups based on their propensity scores. Balance was assessed via a series of ANOVAs in which each covariate (e.g., middle school math grades) was a dependent variable; covariates were considered balanced because the main effect for treatment and the treatment by stratum interaction effect (η^2) were both less than 0.01 (Warne et al., 2019).

Another matching method used in an AP-related study is kernel matching, in which each treatment student is matched to every single control student, but the control students are weighted by how close their propensity scores are to the treatment students (Long et al., 2012). In their study on the effects of taking rigorous high school courses (including but not exclusively AP courses) on multiple college outcomes, Long et al. (2012) created a propensity model that included

measures of prior academic achievement, demographic characteristics, educational needs (such as limited English proficiency or the need for disability accommodations), and high school characteristics. They described performing "a variety of balancing tests," and noted that fewer than 5% of the standardized differences between treatment and control groups could be considered "large" after matching (Long et al., 2012, p. 317).

Two AP studies used 1:1 nearest neighbor matching based on propensity scores and assessed balance after matching by comparing group mean differences on all covariates (McKillip & Rawls, 2013; Patterson & Ewing, 2013). One difference between their approaches is that Patterson and Ewing (2013) created a multilevel propensity model that included a random intercept effect for each high school, and then used the resulting matched pairs to assess AP effects via a simple comparison of group means. In contrast, McKillip and Rawls (2013) used a single-level regression model to generate propensity scores, and then compared outcomes of the match pairs using multilevel modeling. Due to their strict matching requirements, both studies had to exclude students from treatment and control groups who they were unable to match. McKillip and Rawls (2013) excluded between 29-37% of treatment students depending on the subject area; for Patterson and Ewing (2013) the percentage of excluded treatment students ranged from 35-79%.

Clark et al. (2012) also used nearest-neighbor matching, but as a robustness check, they employed five other matching techniques (nearest-neighbor without replacement with propensity scores sorted ascending and descending, and radius matching using calipers of 0.0001, 0.0005, and 0.00001); their final estimate of the treatment effect was similar regardless of the matching technique used. Post-matching balance was assessed by comparing mean differences on ten predictor variables within bands of propensity scores, and around 80% of the time there were no

statistically significant differences between treatment and control groups after matching (Clark et al., 2012).

The final approach to propensity modeling used in AP studies in inverse probability weighting of propensity scores. This method accounts for differences in the probability of a given student being part of the treatment group by weighting some subjects more heavily than others in the outcomes model (Sadler & Sonnert, 2010). As with propensity score matching, the first step is to create a propensity score model that can be used to generate the weights. Sadler and Sonnert (2010) used a polytomous logistic regression model to estimate the probability of each student belonging to one of multiple categories of high school course-taking in each of three STEM subject areas: never taking a course in the subject, taking a regular course in the subject, taking an honors course, taking the AP course but no AP exam, taking the AP course and scoring 1 or 2 on the exam, and taking the AP course and scoring 3 or above on the exam. This single-level propensity model included one high school-level predictor, and was eventually used to weight cases in a multilevel outcomes model (Sadler & Sonnert, 2010). ANOVAs were used to test for significant differences on any of the independent variables across all groups; only one variable was still unbalanced after applying propensity weights (Sadler & Sonnert, 2010).

In summary, the most common approach to applying propensity score modeling to the study of AP effects is to create matched pairs based on propensity scores, but stratified matching (on specific variables or on propensity scores) and propensity weights have also been used, and in general resulted in the loss of fewer cases (and subsequent loss of statistical power) than is seen in studies that employ one-to-one matching. AP researchers have investigated a student's propensity to enroll in an AP course (Clark et al., 2012; Long et al., 2012; Warne et al., 2019), to take an AP exam (McKillip & Rawls, 2013), to earn credit based on AP exam success (Patterson & Ewing,

2013), or a combination of AP course-taking and exam performance (Sadler & Sonnert, 2010). In each of these cases, students selected into the "treatment" of interest while still in high school, when they enrolled in AP courses and completed AP exams. Thus far no study has investigated a student's propensity to use earned AP credit rather than repeating it, which is needed in order to understand whether the use of AP credit has any effect, positive or negative, on subsequent course grades.

Summary

The College Board's Advanced Placement program has become a major force in American secondary education, enjoying widespread support among parents and educational policymakers (Sadler, 2010a). At the same time, there is skepticism on the part of college faculty and independent scholars about the extent to which AP students actually benefit from their experience (Hansen et al., 2006; Sadler & Tai, 2007). Additional research on possible AP effects is needed because of its importance in both secondary and post-secondary education policy and practice. Specifically, few studies address the group of students who choose to repeat AP credit in college, so there is limited empirical evidence regarding the effects of using AP credit on college student success. And despite evidence that AP effects vary by subject area, few studies have accounted for or explored possible explanations for this variation. Finally, the optional nature of the AP program means that students choose whether to enroll in AP courses at the high school level, and decide whether to use earned AP credit at the college level; this self-selection into treatment means that it is difficult to make causal inferences about AP effects.

This study was designed to address those needs by 1) including AP repeaters as a separate group in order to investigate the effects of using AP credit; 2) using cross-sectional multilevel modeling to account for the nested nature of educational data and to explore possible variation in

AP effects across courses; and 3) using propensity score modeling to account for non-random assignment of students into AP groups, thereby allowing for greater confidence in any estimation of a causal relationship between the AP program and student outcomes.

CHAPTER 3: METHOD

The primary goal of this study is to estimate the causal relationship between the use of AP credit and grades in subsequent STEM courses. Specific research questions to address this goal are:

1. How do grades in STEM courses compare across three groups of students, those without AP credit for course pre-requisites, those who used AP credit for course pre-requisites, and those who had earned AP credit but chose not to use it?
2. Which factors predict the propensity to use earned AP credit as a pre-requisite for subsequent STEM courses?
    a) Are the findings consistent when focusing only on the use of AP Calculus credit as a pre-requisite for subsequent STEM courses?
3. What is the effect of using AP credit as a pre-requisite on subsequent course grades for students who choose to use their credit?
    a) To what extent does the effect vary across courses, and can course difficulty predict any of this variation?
    b) To what extent does the effect vary when focusing only on courses requiring AP Calculus as a pre-requisite?

In this chapter, I will describe the research design, the research context, sources of data, variables I used and how they were operationalized in the study, and the series of analyses completed.

Research design

This study used a quantitative design that includes a secondary analysis of existing cross-sectional institutional data. The study used two-level cross-sectional multilevel modeling for the primary statistical models to explore possible variation across courses, the extent to which various level one (student) and level two (course) predictors explain variation in target course grades, and possible cross-level interaction effects. I also used propensity weights to strengthen any causal

inferences that may be drawn from the results. There are three main parts to this study to accomplish the broader research goal: a preliminary analysis that used multilevel modeling only to answer research question 1, the primary analysis that used propensity score modeling and multilevel modeling to answer research questions 2, 3, and 3a, and a mathematics case study that addressed research questions 2a and 3b.

Research context

This study took place at a selective public research university in the Midwest. Students at this institution earn free elective credit for all AP exam scores of 3 or higher as required by state law, but the score required for credit equivalent to introductory courses varies by department. Students generally need exam scores of 4 or 5 to place into advanced STEM courses, and the Physics department only grants credit for scores of 5. The College of Engineering and the College of Science are the largest two colleges at this institution, and their students comprise the majority of students enrolled in the courses included in this study. Engineering students are initially admitted to First-Year Engineering, and then have to be admitted to an Engineering professional school after completing a common set of introductory courses. Students admitted to Engineering and Science have higher academic profiles, on average, than students admitted to other Colleges within the institution.

Data

The initial data set included 28,741 undergraduate students who completed one or more of 34 STEM courses for a grade between fall 2015 and summer 2019. Data for this study were provided by the University Registrar. The retrieval of appropriate data for the current investigation began with identifying the STEM courses for which students could be granted credit based on AP

exam scores in biology, chemistry, mathematics, and physics. Those courses (referred to as "AP courses") were used to develop a list of "target courses," the subsequent courses for which an AP course could serve as a pre-requisite. Target courses could be in the same cognate department as the AP course or a different department, e.g., if students earned AP credit for Calculus 1, they could use that credit to meet the pre-requisite for Calculus 2, or courses outside the Mathematics department in physics or engineering disciplines. There were 44 courses on the initial list of target courses, but ten were eliminated (primarily from the College of Agriculture) because fewer than four students enrolled in the course had earned AP credit but not used it, leaving 34 target courses for both the preliminary and primary models. Of these, 24 courses required calculus as a pre-requisite, making them eligible for the mathematics case study; one advanced physics course, however, was eliminated because after randomly selecting one target course for students who enrolled in more than one, there was only one student in the course who had not used AP credit. This left 23 target courses for the mathematics case study.

If students repeated a course, only the first attempt was included, and if students completed multiple target courses, one course was randomly selected for the study so that any given student would be counted only once. Just over 36% of the sample had completed only one target course during the time period of the study, nearly 17% of the students completed two courses, and the remaining 47% of the sample completed between three and twelve of the target courses. Additionally, three of the target courses (General Biology 2 and both versions of Theoretical Calculus 2) could also themselves be AP courses. Students who earned 5s on the AP Biology exam, or either 4s or 5s on the AP Calculus BC exam, would earn credit for those target courses, and could choose to repeat them. However, I did not include any students in the sample who had earned AP credit for the target course itself, only the course pre-requisites. So if a student earned a 4 on

the AP Biology exam, granting them credit for General Biology 1, and then they subsequently enrolled in General Biology 2, I included them whether or not they repeated the AP credit by enrolling in the first biology course in college. But if they earned a 5 on the exam and then chose to repeat one or both courses in the biology sequence, I would only include them in the sample if they took subsequent advanced biology courses. They would not be part of the cluster of students enrolled in General Biology 2 as a target course. That means this study does not include students who chose to use part, but not all, of their AP credit; this decision was made because my primary interest in this study is the course performance of the students who were encountering the material in target courses for the first time.

The initial data set also included 1,898 students who were missing SAT scores, high school GPAs, or both, which amounted to 6.6% of the original sample. I had to eliminate these students from the sample for two reasons. First, the SAT Math score and high school GPA variables were both included in the propensity models; students without values for those variables could not contribute equally to the model that created propensity weights, nor could they be assigned weights without data on all the factors. Second, I needed a consistent sample composition across all models (i.e., unconditional model, the model with only key AP predictors, and models with additional student predictors) in order to make inferences about the population that the sample represented throughout the analysis, and to assess model fit. After eliminating these students, the final sample included 26,843 students for the preliminary analysis. Pearson's correlation coefficients between the occurrence of missing SAT scores, high school GPA data, or both, and all of the independent variables showed mostly low (under $r = 0.1$) correlations, with the exception of the correlation between missing SAT scores and the international student indicator ($r = .312$, $p < .001$). Pearson's correlation coefficients between the occurrence of missing SAT scores, high school GPA data, or

both, and the dependent variable were under 0.03. These results suggest that excluding the missing data should not bias the results, as the missingness is related to observed variables (international student status) but not the outcome variable. Excluding students with missing SAT scores or high school GPA data resulted in a loss of approximately 21% of the original number of international students, leaving a total of 4,128 international students in the sample. No other group of students (i.e., defined by the remaining independent variables) lost more than 10% of the original sample due to missing data.

From final sample of 26,843 students with no missing data, 10,152 students who had earned AP credit were later included in the sample for the primary analysis that answered research questions 2, 3, and 3a. Of those students with any AP credit, 9,411 students (93%) were included in the case study focusing on students who earned AP calculus credit that answered research questions 2a and 3b. For each phase of the study, I repeated the random assignment of students who had completed multiple target courses, so any individual student could have been part of different AP groups in the three phases of the study depending on whether they had earned and subsequently used AP credit prior to taking the specific target course to which the student was assigned.

Table 3 shows the final list of target courses, the AP courses that could serve as pre-requisites for each, historical DFW rates for each course, and total enrollment from all students in the sample. All courses showing that Calculus 1 or 2 is a required pre-requisite were included in the mathematics case study except for Advanced Physics B. Note that the enrollment totals in Table 3 include students enrolled in more than one target course; subsequent tables include the final number of students randomly selected for each course for each phase of the study.

Table 3: Target Course Enrollment, DFW Rate, and AP Course Pre-Requisites

Course ID	Target Course	Total Enroll.	DFW Rate	AP Courses Required
1	Aeronautical Engr	1,158	0.15	Calculus 2, Engr. Physics 1
2	Agric/Biol Engr	265	0.03	Chemistry 2
4	Animal Science	1,171	0.13	Chemistry 1
7	General Biol 2	4,701	0.08	Biology 1
8	Microbiology	951	0.25	Biology 2, Chemistry 2
9	Cellular Biol	994	0.07	Calculus 2, Chemistry 2
10	General Biol 3	778	0.22	Biology 2, Chemistry 2
12	Biomedical Engr	407	0.01	Chemistry 2
13	Civil Engr	938	0.28	Calculus 2, Engr. Physics 1
14	Chemical Engr	799	0.21	Calculus 1, Chemistry 1, Engr. Physics 1
15	Honors Chem	171	0.02	Calculus 1
16	Organic Chem A	2,032	0.13	Chemistry 2
17	Organic Chem B	728	0.27	Chemistry 2
18	Organic Chem C	851	0.05	Chemistry 2
19	Organic Chem D	195	0.13	Chemistry 2
20	Electric/Comp Engr	4,837	0.25	Calculus 2, Engr. Physics 1
21	Environmental Engr	235	0.07	Calculus 2, Chemistry 2, Engr. Physics 1
26	Industrial Engr A	1,344	0.17	Calculus 2
27	Industrial Engr B	2,817	0.10	Calculus 2
28	Applied Calc 2	5,852	0.22	Calculus 1
29	Theoretical Calc 2A	6,765	0.26	Calculus 1
30	Theoretical. Calc 2B	4,180	0.15	Calculus 1
31	Calculus 3	12,591	0.18	Calculus 2
32	Linear Algebra	7,192	0.17	Calculus 2
33	Mechanical Engr A	4,224	0.20	Calculus 2, Chemistry 1
34	Mechanical Engr B	4,185	0.23	Calculus 2, Engr. Physics 1
35	Materials Sci Engr	2,646	0.09	Calculus 1, Chemistry 1
37	Nuclear Engr	475	0.14	Calculus 2, Engr. Physics 1
39	Life Science Physics	97	0.04	Biology 2, Calculus 1, Chemistry 1
40	Engr Physics 2A	5,450	0.16	Calculus 1, Engr. Physics 1
41	Engr Physics 2B	3,241	0.11	Calculus 1, Engr. Physics 1
42	Adv Physics A	224	0.05	Calculus 2, Engr. Physics 2
43	Adv Physics B	249	0.18	Calculus 2, Engr. Physics 2
44	Statistics	3,285	0.13	Calculus 2

Variables

The data set included both level one (student) and level two (course) variables as described below.

Dependent variable: Final course grades

Following prior AP research (e.g., Patterson & Ewing, 2013), final grades earned in the target courses served as the dependent variable, because a primary goal of the AP program is to prepare students for advanced college courses, and the College Board regularly engages college faculty to ensure that learning objectives of AP courses are closely aligned to those of introductory college courses in the same subjects (Ewing et al., 2010). Grades ranged from A+ to F and were converted to a numerical scale consistent with how the institution calculates GPA (4.0 for A+ and A grades, 3.7 for A-, 3.3 for B+, etc.).

Student predictors

Advanced Placement variables

The primary independent variables for each part of the study were dichotomous variables indicating students' AP status. *AP User* (coded as 1 in the first dummy variable) refers to students who earned AP credit for at least one pre-requisite and enrolled directly in the target course, and *AP Non-User* (coded as 1 in the second dummy variable) refers to students who earned AP credit for at least one pre-requisite but repeated it at the institution prior to enrolling in the target course. Students who earned AP credit for more than one target course pre-requisite were included in the AP Non-User group only if they repeated all pre-requisite courses for which they earned AP credit. For example, if a target course requires both biology and chemistry as pre-requisites and the student earned AP credit for both, the student was considered an AP User if he or she used AP

credit for one course but repeated the other. Students with zeros on both of these dichotomous variables comprised a *Non-AP* reference group.

Subsequent analyses that included only students who had earned AP credit each had only one AP indicator. *AP User* was coded the same way for the primary analysis as in the preliminary analysis (i.e., 1 = student who fulfilled at least one course pre-requisite with AP credit, 0 = student who repeated all earned AP credit). In the mathematics case study, *APM User* was coded as 1 for students who used earned AP credit in calculus as a pre-requisite and 0 for students who earned AP Calculus credit but repeated the course in college.

Individual students could be in different AP groups for each phase of the study. For example, Student X earned AP credit for Calculus 1 and Biology 1, and completed three target courses: Calculus 2, Biology 2, and Animal Science. Student X used their Biology AP credit but repeated Calculus 1 before taking Calculus 2. For the preliminary phase of the study, the Animal Science course was randomly selected for Student X, and they were considered a Non-AP student because they did not have AP credit for Chemistry 1, the only AP course pre-requisite for Animal Science. For the primary phase of the study, Biology 2 (one of the two target courses for which Student X had earned AP credit for course pre-requisites) was randomly selected for Student X; this time, Student X was considered an AP User because they used their AP credit for Biology 1. For the case study, the only relevant target course Student X completed was Calculus 2, and they were considered an APM Non-User because they repeated Calculus 1.

Two other AP-related variables are indicators of AP exam scores. For the primary analysis of students who had earned AP credit, students could have earned AP credit on more than one AP exam that would serve as a pre-requisite for the target course. I used the *AvgScore* variable to indicate the average score the student earned on relevant AP exams (i.e., if a course required

calculus and physics as pre-requisites, and the student earned AP credit for both, then the student's scores on the calculus and physics AP exams would be averaged). Scores on other exams that were not relevant for the target course (including other STEM exams) were not included in the average score. Additionally, I included exam scores in the average whether the student used the credit earned or not. Finally, if a student took a relevant exam but scored too low to earn credit, then that score was not included in the average.

For the mathematics case study, I used *APMScore* to indicate the student's AP calculus exam score. If a target course required credit for calculus 2, then the student's score on the AP Calculus BC exam was used for this variable. If a target course only required credit for first-semester calculus, however, a student could have earned that credit by scoring a 4 or 5 on the AP Calculus AB exam, the AP Calculus BC exam, or on the AB subscore of the BC exam. If a student had more than one relevant calculus exam score (e.g., a student may have earned a 4 on the AP Calculus AB exam their junior year of high school and then earned a 5 on the AB subscore when they took the BC Calculus exam their senior year), then I used the highest relevant score.

Prior academic achievement

I used core high school GPA and standardized test scores (SAT or ACT) as prior academic achievement indices. Core high school GPA is calculated by the institution's Office of Admissions and includes high school mathematics, English, laboratory science, foreign language, and social studies courses measured on a 4-point scale. SAT and ACT scores are submitted by students to the institution. Because students in the sample reported both older and newer versions of SAT scores as well as ACT scores, all test scores were converted to the most recent SAT scores using concordance tables published by the College Board (n.d., *Concordance*). SAT Math and SAT Total scores were too highly correlated to use both in the model ($r = 0.811$, $p < 0.001$), so I used only

the SAT Math scores as in other AP research focusing on STEM courses (e.g., Burkholder & Wieman, 2019; Sadler & Tai, 2007). To make it easier to interpret model results, I rescaled SAT Math scores, dividing them by 10 so they ranged from 20 to 80 rather than 200 to 800.

Demographics

The following five demographic indicators were coded with dummy variables for analysis: gender (female = 1, male = 0), international student (yes = 1, no = 0), underrepresented minority student (URM; yes = 1, no = 0), Pell grant recipient for at least one year within the time frame of the study (yes = 1, no = 0), and first generation student (yes = 1, no = 0). Students at this institution are considered first generation if they indicate on their application that neither parent graduated from college with a bachelor's degree.

Student college

Colleges and majors have different policies or practices, and college faculty and advisors may tend to provide different guidance regarding the choice to use AP credit or not. It may be that the propensity to use AP credit varies by college within the institution (i.e., the student's academic home rather than the college offering the course). To explore whether college enrollment was a factor in the propensity to use AP credit, I created dummy-coded variables to represent First-Year Engineering (*FYES*), the various Engineering professional schools (*EngrS*), and the College of Science (*SciS*), with the reference group representing all other colleges. These indicators represented the student's own academic home at the time they enrolled in the target course, not the college or school that owned the course itself. While I did not create any variable related to how far along students were in their degree plan, at this institution, future engineers are admitted to First-Year Engineering, and typically seek admission to one of the Engineering professional

schools at the start of their sophomore year. This means that FYE students are typically in their first year of college, Engineering professional school students are mostly sophomores and above, and College of Science students would be a mix of first year and upper-division students.

Course predictors

Course difficulty

Historical DFW rate represents the percentage of students who earned Ds, Fs, or Ws in any section of the course offered during the fall 2015 to summer 2019 time period, and provides an estimate of course difficulty, which one prior study suggested as a possible source of variability in course grades to explore (Wladis et al., 2017). Values range from 1% to 27%, with an average DFW rate of 15%, and were included in Table 3 above. Because I used historic DFW rate rather than the DFW rate of study participants, the values do not change across models in this study.

Other course predictors

Historical DFW rate was the only course predictor that was included in any of the final models, but I explored other possible course-level predictors as well, which were later removed from the models for parsimony. Of the 34 courses included in the sample, 17 have more than one pre-requisite course, so I used a dummy variable to indicate if the course had multiple pre-requisites (*Mult_PreR;* yes = 1, no = 0). In the preliminary and primary analyses, I also used a dummy variable to indicate whether the target course required calculus as a pre-requisite (*Math_PreR;* yes = 1, no = 0). For the mathematics case study, I explored the use of a dummy variable to indicate if the target course required another science pre-requisite course in addition to calculus (*Sci_PreR*; yes = 1, no = 0). Finally, I used dummy variables to indicate whether the

course was offered by the College of Engineering (*EngrC*; yes = 1, no = 0), or the College of Science (*SciC*; yes = 1, no = 0).

Tables 4, 5, and 6 below provide descriptive statistics of student demographic and prior academic achievement variables by AP groups, and Tables 7, 8, and 9 include the key student-level academic achievement variables (high school GPA, SAT Math scores, and, in Tables 8 and 9 only, Average AP Exam scores) organized by course, because there are evident differences in student academic background across target courses. All six tables group students by the primary independent variables indicating AP status. More specifically, Tables 4 and 7 provide data for the preliminary model that included all students (RQ 1), Tables 5 and 8 show the subset of students with AP credit who were included in the primary model (RQs 2, 3, and 3a), and Tables 6 and 9 summarize descriptive statistics for the mathematics case study (RQs 2a and 3b).

Table 4: Preliminary Analysis: Descriptive Statistics

	AP User (*N*=5,363)		AP Non User (*N*=2,001)		No AP Credit (*N*=19,479)	
	M	*SD*	*M*	*SD*	*M*	*SD*
HS GPA	3.71	0.28	3.67	0.27	3.57	0.33
SAT Math	745	47	723	51	674	75
Female	0.26	0.44	0.29	0.46	0.40	0.49
URM	0.05	0.22	0.05	0.23	0.09	0.29
International	0.10	0.30	0.07	0.26	0.18	0.38
First Gen	0.11	0.32	0.12	0.33	0.19	0.40
Pell Recipient	0.14	0.35	0.14	0.35	0.20	0.40

Note: All variables other than HS GPA and SAT Math are dummy coded, so the mean value equals the percentage of students in that group who are in the demographic category represented by the variable name, e.g., the AP User group is 26% female.

Table 5: Primary Analysis: Descriptive Statistics

	AP User (*N*=7,019)		AP Non User (*N*=3,133)	
	M	*SD*	*M*	*SD*
HS GPA	3.70	0.28	3.66	0.28
SAT Math	740	48	718	51
Avg AP Score	4.66	0.47	4.45	0.51
Female	0.28	0.45	0.29	0.46
URM	0.05	0.23	0.06	0.23
International	0.10	0.29	0.06	0.24
First Gen	0.12	0.32	0.12	0.32
Pell Recipient	0.15	0.35	0.14	0.35
FYE Student	0.36	0.48	0.39	0.49
Engineering Student	0.29	0.46	0.26	0.44
Science Student	0.19	0.40	0.17	0.38

Note: All variables other than HS GPA, SAT Math, and Avg AP Score are dummy coded, so the mean value equals the percentage of students in that group who are in the demographic category represented by the variable name, e.g., the AP User group is 28% female.

Table 6: Case Study: Descriptive Statistics

	AP User (*N*=6,283)		AP Non User (*N*=3,128)	
	M	*SD*	*M*	*SD*
HS GPA	3.70	0.28	3.66	0.28
SAT Math	742	47	721	50
AP Calc Score	4.72	0.45	4.51	0.50
Female	0.26	0.44	0.27	0.44
URM	0.05	0.22	0.06	0.23
International	0.10	0.30	0.07	0.25
First Gen	0.12	0.32	0.12	0.32
Pell Recipient	0.14	0.34	0.14	0.35
FYE Student	0.39	0.49	0.41	0.49
Engineering Student	0.28	0.45	0.27	0.44
Science Student	0.20	0.40	0.17	0.38

Note: All variables other than HS GPA, SAT Math, and Avg AP Score are dummy coded, so the mean value equals the percentage of students in that group who are in the demographic category represented by the variable name, e.g., the AP User group is 26% female.

Table 7: Preliminary Analysis (All Students): Descriptive Statistics by Course and AP Status

	Non-AP					AP Users					AP Non-Users				
	n	HS GPA		SAT Math		*n*	HS GPA		SAT Math		*n*	HS GPA		SAT Math	
		M	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*
Aeronautical Engr	135	3.64	0.28	715	56	87	3.76	0.21	765	33	21	3.68	0.18	743	44
Agric/Biol Engr	54	3.70	0.29	676	57	9	3.85	0.31	733	53	3	3.90	0.17	723	15
Animal Science	588	3.46	0.38	576	67	12	3.90	0.11	706	50	7	3.73	0.44	673	95
General Biol 2	2481	3.51	0.36	611	76	105	3.65	0.33	699	61	56	3.75	0.24	689	60
Microbiology	287	3.55	0.37	628	74	86	3.82	0.24	721	51	8	3.86	0.20	703	54
Cellular Biol	97	3.72	0.28	697	55	60	3.81	0.21	752	46	31	3.69	0.24	725	36
General Biol 3	197	3.56	0.33	625	69	43	3.76	0.30	727	62	4	3.87	0.14	720	63
Biomedical Engr	49	3.64	0.29	695	59	12	3.88	0.17	750	45	3	3.85	0.22	760	35
Civil Engr	143	3.58	0.32	691	52	25	3.80	0.19	745	38	9	3.71	0.21	716	67
Chemical Engr	55	3.64	0.31	691	47	73	3.85	0.18	746	45	23	3.65	0.33	721	53
Honors Chem	7	3.85	0.13	710	67	36	3.86	0.23	758	46	0				
Organic Chem A	680	3.56	0.34	634	73	71	3.76	0.25	732	46	15	3.77	0.26	757	35
Organic Chem B	257	3.59	0.32	640	78	23	3.75	0.26	746	50	8	3.81	0.21	756	44
Organic Chem C	114	3.72	0.30	712	54	33	3.86	0.17	728	49	42	3.85	0.18	752	41
Organic Chem D	31	3.46	0.38	664	69	25	3.78	0.26	718	50	5	3.39	0.13	720	53
Electric/Comp Engr	659	3.62	0.29	713	53	292	3.71	0.28	758	38	82	3.76	0.24	747	38
Environmental Engr	27	3.75	0.18	677	49	7	3.90	0.11	754	40	7	3.89	0.09	731	21
Industrial Engr A	179	3.62	0.29	710	53	42	3.59	0.32	753	43	15	3.52	0.38	737	34
Industrial Engr B	598	3.63	0.32	706	55	89	3.73	0.26	761	33	37	3.64	0.27	432	48
Applied Calc 2	3493	3.48	0.35	646	66	322	3.63	0.31	694	54	272	3.60	0.30	689	57
Theoretical Calc 2A	1760	3.55	0.34	693	62	502	3.67	0.29	730	48	299	3.64	0.28	715	51
Theoretical. Calc 2B	932	3.63	0.31	707	53	240	3.61	0.31	727	47	392	6.66	0.27	722	44
Calculus 3	2119	3.63	0.30	712	58	1000	3.70	0.27	752	41	209	3.72	0.25	744	43
Linear Algebra	1122	3.61	0.32	711	55	475	3.70	0.27	757	39	79	3.71	0.24	739	44
Mechanical Engr A	529	3.60	0.33	710	53	257	3.75	0.26	756	40	56	3.68	0.25	744	35

Table 7 continued

Mechanical Engr B	481	3.65	0.28	710	56	171	3.74	0.26	759	39	70	3.72	0.28	737	40
Materials Sci Engr	265	3.60	0.32	701	59	403	3.73	0.26	746	44	77	3.58	0.29	717	45
Nuclear Engr	81	3.60	0.34	714	58	24	3.74	0.24	753	38	4	3.42	0.19	683	67
Life Science Physics	7	3.41	0.43	679	63	25	3.64	0.32	716	50	3	3.84	0.18	747	12
Engr Physics 2A	774	3.65	0.28	706	56	209	3.70	0.28	751	42	79	3.74	0.22	737	42
Engr Physics 2B	532	3.59	0.33	706	60	190	3.67	0.28	757	40	44	3.65	0.27	733	51
Adv Physics A	34	3.55	0.36	684	72	16	3.75	0.29	753	33	1	3.13	n/a	590	n/a
Adv Physics B	37	3.63	0.32	701	67	17	3.77	0.23	751	56	5	3.65	0.36	740	37
Statistics	675	3.58	0.34	712	61	382	3.68	0.31	760	40	35	3.66	0.30	751	35

Table 8: Primary Analysis (AP Students): Descriptive Statistics by Course and AP Status

	AP Users							AP Non-Users						
		HS GPA		SAT Math		AP Exam Score			HS GPA		SAT Math		AP Exam Score	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*	*n*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Aeronautical Engr	97	3.73	0.23	763	35	4.85	0.33	18	3.64	0.32	741	44	4.67	0.49
Agric/Biol Engr	7	3.83	0.25	746	50	4.29	0.49	6	3.80	0.29	742	45	4.50	0.55
Animal Science	18	3.88	0.14	677	66	3.50	0.71	28	3.77	0.29	673	63	3.04	0.19
General Biol 2	174	3.65	0.33	693	59	4.00	0.00	95	3.70	0.27	687	59	4.00	0.00
Microbiology	107	3.79	0.26	720	51	4.39	0.47	10	3.88	0.18	692	56	4.75	0.42
Cellular Biol	72	3.80	0.25	737	48	4.58	0.44	39	3.68	0.28	720	45	4.37	0.48
General Biol 3	53	3.78	0.26	723	60	4.71	0.41	7	3.84	0.11	704	62	4.21	0.39
Biomedical Engr	12	3.92	0.10	749	47	4.33	0.49	6	3.76	0.33	745	50	4.33	0.52
Civil Engr	33	3.67	0.30	741	43	4.71	0.43	10	3.76	0.25	709	56	4.60	0.52
Chemical Engr	118	3.78	0.24	741	45	4.66	0.42	35	3.76	0.22	712	50	4.46	0.49
Honors Chem	37	3.88	0.21	757	38	4.84	0.37	2	3.93	0.09	730	57	4.00	0.00
Organic Chem A	90	3.76	0.29	724	51	4.29	0.46	18	3.74	0.25	731	46	4.33	0.49
Organic Chem B	32	3.77	0.26	734	48	4.28	0.46	14	3.84	0.20	735	49	4.36	0.50
Organic Chem C	37	3.74	0.30	452	43	4.68	0.47	42	3.83	0.19	743	37	4.38	0.49
Organic Chem D	38	3.75	0.27	717	58	4.50	0.51	8	3.44	0.31	720	59	4.13	0.35
Electric/Comp Engr	240	3.71	0.29	758	41	4.84	0.36	66	3.73	0.26	746	40	4.70	0.46
Environmental Engr	12	3.85	0.15	742	39	4.46	0.45	9	3.88	0.13	731	55	4.33	0.50
Industrial Engr A	29	3.60	0.32	755	39	4.66	0.48	11	3.47	0.38	727	31	4.73	0.47
Industrial Engr B	109	3.76	0.25	756	39	4.81	0.40	46	3.67	0.33	733	52	4.57	0.50
Applied Calc 2	405	3.66	0.30	697	54	4.47	0.50	331	3.62	0.30	690	56	4.32	0.47
Theoretical Calc 2A	1167	3.67	0.28	726	49	4.63	0.48	636	3.63	0.28	712	51	4.46	0.50
Theoretical. Calc 2B	574	3.65	0.29	730	47	4.58	0.49	866	3.65	0.28	717	46	4.48	0.50
Calculus 3	1042	3.69	0.27	754	41	4.73	0.44	243	3.70	0.26	740	45	4.63	0.48
Linear Algebra	488	3.72	0.27	755	39	4.77	0.42	86	3.69	0.26	741	43	4.60	0.49
Mechanical Engr A	332	3.75	0.24	749	44	4.58	0.46	73	3.64	0.26	736	36	4.42	0.50

Table 8 continued

Mechanical Engr B	180	3.73	0.23	761	3.39	4.87	0.33	50	3.74	0.23	740	40	4.64	0.48
Materials Sci Engr	554	3.82	0.26	740	44	4.67	0.43	216	3.59	0.29	721	46	4.45	0.50
Nuclear Engr	28	3.72	0.21	755	39	4.88	0.32	7	3.82	0.16	740	69	4.50	0.50
Life Science Physics	32	3.65	0.32	706	48	4.38	0.46	4	3.64	0.37	688	66	4.38	0.48
Engr Physics 2A	251	3.73	0.27	755	37	4.56	0.34	64	3.72	0.24	744	43	4.60	0.49
Engr Physics 2B	206	3.70	0.28	752	41	4.85	0.35	51	3.74	0.22	744	45	4.60	0.49
Adv Physics A	11	3.77	0.27	753	55	4.82	0.40	4	3.77	0.23	733	44	4.63	0.48
Adv Physics B	26	3.74	0.30	755	43	4.97	0.20	4	3.66	0.44	753	28	5.00	0.00
Statistics	408	3.67	0.31	760	39	4.80	0.40	28	3.65	0.30	739	45	4.61	0.50

Table 9: Case Study (AP Calculus Students): Descriptive Statistics by Course and AP Status

	AP Users							AP Non-Users						
		HS GPA		SAT Math		AP Exam Score			HS GPA		SAT Math		AP Exam Score	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*	*n*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Aeronautical Engr	74	3.73	0.24	761	36	4.74	0.44	23	3.64	0.26	754	42	4.65	0.49
Cellular Biol	68	3.79	0.25	750	45	4.75	0.44	25	3.74	0.24	740	42	4.64	0.49
Civil Engr	22	3.77	0.28	753	38	4.91	0.29	14	3.76	0.23	746	36	4.57	0.51
Chemical Engr	130	3.79	0.23	743	43	4.81	0.40	63	3.73	0.24	716	53	4.52	0.50
Honors Chem	55	3.85	0.20	755	43	4.85	0.36	5	3.85	0.18	732	37	4.40	0.55
Electric/Comp Engr	225	3.76	0.24	755	42	4.80	0.40	85	3.74	0.23	745	37	4.60	0.49
Environmental Engr	7	3.82	0.17	727	49	4.86	0.38	4	3.64	0.10	690	36	4.50	0.58
Industrial Engr A	31	3.58	0.32	750	46	4.71	0.46	13	3.68	0.31	724	39	4.77	0.44
Industrial Engr B	113	3.76	0.25	754	38	4.81	0.39	43	3.67	0.31	744	44	4.51	0.51
Applied Calc 2	437	3.67	0.30	698	53	4.49	0.50	373	3.62	0.30	692	57	4.36	0.48
Theoretical Calc 2A	1168	3.67	0.28	726	48	4.64	0.48	649	3.63	0.28	712	51	4.47	0.50
Theoretical. Calc 2B	631	3.66	0.29	730	47	4.61	0.49	898	3.66	0.28	718	46	4.49	0.50
Calculus 3	1054	3.70	0.28	754	42	4.73	0.44	284	3.72	0.25	742	42	4.65	0.48
Linear Algebra	457	3.72	0.26	760	36	4.78	0.42	97	3.74	0.25	738	47	4.55	0.50
Mechanical Engr A	213	3.75	0.25	759	37	4.82	0.38	84	3.73	0.23	735	41	4.63	0.49
Mechanical Engr B	157	3.75	0.25	759	40	4.83	0.38	62	3.70	0.30	748	41	4.60	0.50
Materials Sci Engr	536	3.72	0.27	745	42	4.77	0.42	211	3.60	0.29	723	46	4.50	0.50
Nuclear Engr	22	3.71	0.24	761	33	4.91	0.29	10	3.67	0.27	733	57	4.60	0.52
Life Science Physics	47	3.73	0.28	710	54	4.51	0.52	4	3.85	0.16	698	63	4.50	0.58
Engr Physics 2A	192	3.73	0.28	752	40	4.81	0.39	79	3.72	0.23	736	44	4.63	0.49
Engr Physics 2B	200	3.69	0.26	754	37	4.76	0.43	57	3.69	0.28	743	44	4.65	0.48
Adv Physics A	27	3.73	0.29	761	34	4.93	0.27	3	3.42	0.51	723	35	4.33	0.58
Statistics	417	3.68	0.31	759	40	4.83	0.38	42	3.70	0.30	739	52	4.60	0.50

Analyses

This study included three sets of analyses to address all of the research questions; each phase is described in detail below and summarized in Table 10 at the end of this chapter. Recall that each set of analyses involved different student samples. More specifically, the preliminary series of analyses included all 26,843 students and all 34 courses to address research question 1, indicating whether students who use AP credit to fulfill course pre-requisites performed as well or better than non-AP students who completed pre-requisite courses at the college level.

The primary phase of the analysis focused on a subset of the whole sample, i.e., students who had earned AP credit for one or more target course pre-requisites. I first estimated students' propensity for using earned AP credit to answer the second research question about which factors predict student choice regarding use or non-use of AP credit. I then created propensity weights for use in the subsequent multilevel model to answer the third research question in a way that allows for causal inferences to be drawn from the results.

Finally, to answer sub-parts of research questions two and three (i.e., RQs 2a and 3b), I repeated several steps of the primary analysis as a case study focusing on the use of AP Calculus credit. More high school students take one or both AP calculus exams than any other STEM AP exam (College Board, n.d., *AP Program Participation and Performance Data 2019*), and appropriate placement in the first mathematics course is important for future success in STEM majors (De Urquidi et al., 2015; Herzog, 2005). Therefore, a special focus on the effect of using AP credit in calculus is appropriate.

As discussed in Chapter 2, there are different possible treatment effects to consider, specifically an estimate of the average treatment effect for the treated (ATT), and an estimate of the treatment effect on the broader population of interest (ATE) (Morgan & Winship, 2015). In

this study, I created propensity models for both types of treatment effects for two reasons. First, I was primarily interested in the ATT, so as to establish the effect of using AP credit on the students who currently choose to use it. Second, because the answer would have broader policy implications, I decided to explore the effect of using AP credit on a broader group of students, including those who do not typically choose to use their credit.

Preliminary analysis

To answer the first research question, I began with a preliminary analysis that is similar to the preceding study (Hurt & Maeda, under review), conducting a series of analyses using two-level cross-sectional multilevel linear models (Raudenbush & Bryk, 2002) to examine the relationship between student AP status and grades in target courses after controlling for student- and course-level indicators. In the model, level one represented students and level two represented the target courses taken by the students; thus the model captured the nested effect of AP status on students within target courses. The analysis began with an unconditional model to quantify variation in the target course grade across the 34 courses in the sample. The intraclass correlation coefficient (ICC), which is the ratio of group-level variance (τ^2) to total variance ($\tau^2 + \sigma^2$) (Raudenbush & Bryk, 2002), indicated that 8.1% of the variance in target course grades was due to course-level characteristics. This is below 10%, a typical threshold that has been considered a guideline for justifying the use of multilevel modeling in social science research, but Huang (2018) argued that this threshold is a "myth," and that it is better to account for clustering effects even when they are responsible for much less than 10% of the total variance in the outcome. The first conditional model included the two dummy variables indicating AP status (i.e., *AP User* and *AP Non-User*); these both varied significantly across courses. Subsequent conditional models included high school GPA, rescaled SAT Math scores, and demographic variables at level one, and then

added course DFW rate at level two. No other course-level variables had significant main effects on target course grades, and there were no significant cross-level interaction effects between the AP status variables and DFW rate. Model 1 below is the final model for this part of the analysis. In this model, i = 1, 2, …, n students in course j; j = 1, 2, …, 34 courses; each γ coefficient represents the fixed effects of the associated predictors; γ_{10} and γ_{20} represent the average or common effect of the AP User and AP Non-User predictors, while u_{1j} and u_{2j} represent the random course-level effects of the AP User and AP Non-User predictors respectively; u_{0j} is a residual course-level error term for the conditional average of grades for the jth course; and r_{ij} is a residual error term for student i in course j.

Model 1

$$GRADE_{ij} = \gamma_{00} + \gamma_{01} * (DFW_j) + \gamma_{10} * (AP\_User_{ij}) + \gamma_{20} * (AP\_Non_{ij}) + \gamma_{30} * (Gender_{ij}) + \gamma_{40} * (URM_{ij}) + \gamma_{50} * (International_{ij}) + \gamma_{60} * (FirstGen_{ij}) + \gamma_{70} * (Pell_{ij}) + \gamma_{80} * (HSGPA_{ij}) + \gamma_{90} * (SATM_{ij}) + u_{0j} + u_{1j} * (AP\_User_j) + u_{2j} * (AP\_Non_j) + r_{ij}$$

None of the variables in this model were centered, because I only planned to interpret the results of the two AP indicators, which are both binary variables. Table 7 above provides information about the students randomly assigned to each course for this model, including the number and percentage in each of the key independent variables indicating AP status.

Primary analysis: Propensity score model

The purpose of using propensity weights is to account for potential confounders that affect both the probability of treatment (choosing to use AP credit for at least one course pre-requisite)

and the outcome (target course grades). I used the *twang* package (Ridgeway et al., 2020) in R which implemented generalized boosted regression modeling to create a single-level propensity model that included level one predictors and dummy variables for each course to account for course-level effects (Arpino & Mealli, 2011; Li et al., 2013). Generalized boosted models (GBMs) use an iterative and flexible estimation method that allows for complex and nonlinear relationships among variables (McCaffrey et al., 2013). The predictors in the propensity models include all the demographic variables included in Model 1, high school GPA, SAT Math score, average AP score, a set of dummy variables indicating the student's academic home (First Year Engineering, one of the Engineering professional schools, or the College of Science), and dummy variables for each course. In total, 44 variables were included in the propensity model. Model 2 represents the logistic regression model for propensity estimation.

Model 2

$$logit(T_i = 1) = \beta_0 + \Sigma\beta_s X_{si} + \Sigma\beta_c Z_{ci} + r_i$$

In Model 2 above, the probability of using at least one AP course ($T_i = 1$) is predicted with a vector of student-level factors noted as X_{si}, where s represents a student-level individual factor for student i, and a vector of fixed cluster effects noted as Z_{ci}, where ci represents the student-level effect associated with each course. The model also includes the residual individual error term r_i.

The next step was to use the GBM to estimate both ATT and ATE propensity weights, to ensure that treatment and control groups were as similar as possible on the pre-treatment covariates, in order to approximate the conditions of random assignment (Routon & Walker, 2019). ATT weights assign a value of 1 to any student outcome if the student did select into the treatment condition (use of AP credit), and a value of $p_i/(1- p_i)$ if the student did not select into the treatment

condition, where p_i represents the estimated probability of treatment for student i (Morgan & Winship, 2015). ATE weights assign a value of $1/p_i$ to treatment students (AP users) and a value of $1/(1- p_i)$ to students who did not select into the treatment condition (AP non-users) (Morgan & Winship, 2015).

Estimating the ATT requires that the distribution of propensity scores of treated students is contained within the range of propensity scores of untreated students; untreated students with propensity scores that fall below the range of treated students are not included in the analysis (Leite et al., 2015). In contrast, the ATE requires common support for both treated and untreated individuals, meaning that there are no values of pre-treatment variables (such as high school grades or SAT Math scores) that only occur among one group (McCaffrey et al., 2013). Because of these requirements, the analyses using ATT weights include all treatment students (AP users) but not all control students (AP non-users); the analyses using ATE weights included a wider range of control students but excluded some treatment students with the highest propensity scores. The ATT propensity model used for the primary analysis included all treatment students and approximately 46% of control students (see Table A1 in the appendix); the ATE propensity model included approximately 91% of the treatment students and 66% of control students (see Table A2 in the appendix). Figure 1 below shows the distribution and overlap of treatment and control students across the spectrum of propensity scores, and provides a visual representation of the extent to which treated and untreated students were included in the ATT or ATE analyses. Students who chose to use AP credit are clearly clustered around higher propensity scores, but the bulk of the control students are within the range of common support.

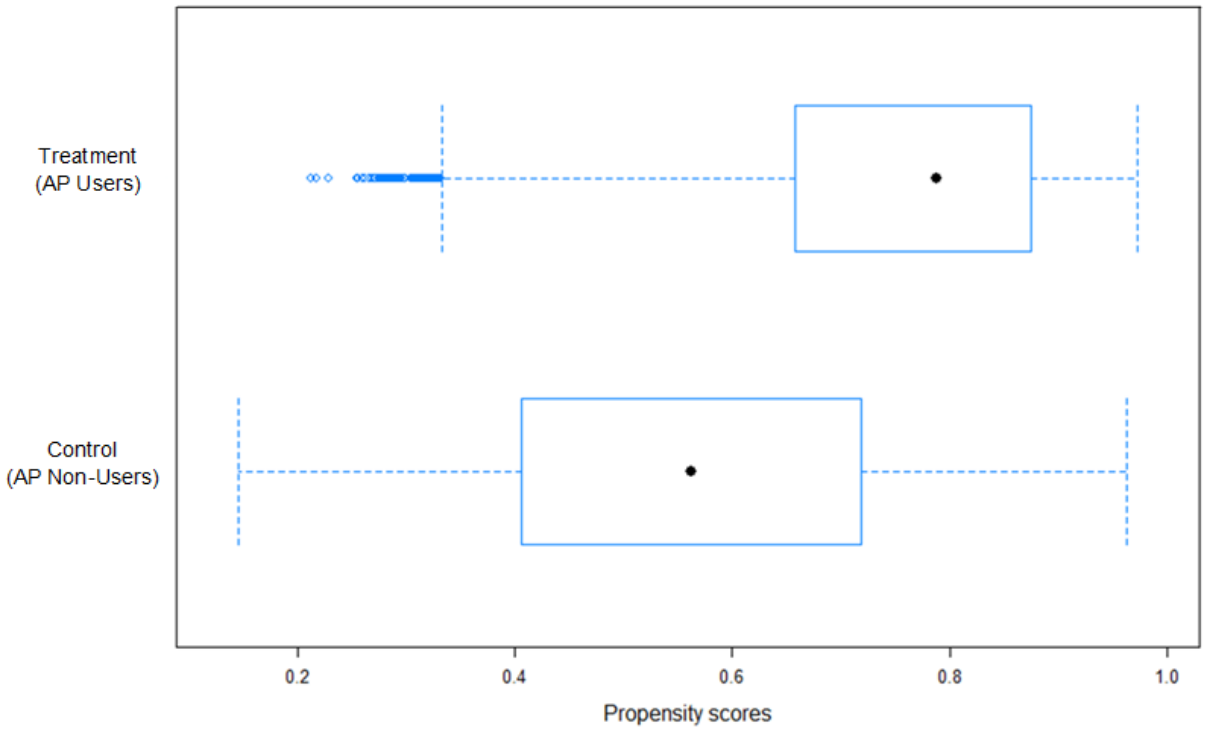

Figure 1: Primary analysis: Distribution of propensity scores for treatment and control students

Both weighting methods successfully balanced treatment and control groups, with only one variable (i.e., cluster effects for course 15) remaining a significant predictor of treatment status after balancing. This course, an advanced first-year chemistry course, is specifically designed for students with AP credit in calculus, so it is unusual for students to repeat their AP credit prior to enrolling in the course. The standardized effect size between treated and untreated students after weighting was very small (0.05), and the course has relatively low enrollment, so the lack of balance on this one predictor should not be a major concern. Tables A3 and A4 in the appendix include standardized mean differences between treatment and control students for each course, before and after weighting, for the ATT and ATE models, respectively.

Primary analysis: Outcomes model

The primary goal of this study is to estimate the causal relationship between use of AP credit and grades in subsequent courses. To address this question, I conducted an additional series of analyses using two-level cross-sectional multilevel linear models (Raudenbush & Bryk, 2002) incorporating propensity weights from the model described above, using the *lme4* package in R (Bates et al., 2020). As in the preliminary analysis, level one represented students and level two represented the target courses taken by the students. Table 8 above provides information about the students randomly selected to each course for this model, including the number and percentage in each of the key AP independent variables. I began with an unconditional model, and found that 6.2% of variation in the target course grades was accounted for by course differences. Next, I added the unweighted AP_User variable and found that it did significantly vary across groups. The final version of Model 3 included only one level one predictor, AP_User, weighted by either ATT or ATE propensity scores generated in the previous analysis. Because all of the student-level predictors were successfully balanced by the weighting process, I chose not to include any other level one predictors (i.e., to create a doubly robust model) so as to simplify interpretation. I tested two level two predictors, historical DFW rate and a dummy variable indicating the target course required calculus as a pre-requisite (*Math_PreR)*, as well as interaction effects between both level two variables and the weighted AP_User predictor. Of these, only DFW rate was a significant predictor of target course grades, so it is the only level 2 variable I included in the final model (Model 3, below). In this model, $i = 1, 2, \ldots, n$ students in course j; $j = 1, 2, \ldots, 34$ courses; $ATT\text{-}AP\_User_{ij}$ = treatment status weighted by the ATT weights estimated in Model 2, while u_{1j} represents the random course-level effects of the weighted AP User predictor; u_{0j} is a residual course-level error term for the conditional average of grades for the jth course, and r_{ij} = a residual

error term for student *i* in course *j*. I subsequently used the same model incorporating ATE weights as well.

Model 3

$$GRADE_{ij} = \gamma_{00} + \gamma_{01} * (DFW_j) + \gamma_{10} * (ATT - AP\_User_{ij}) + u_{0j} + u_{1j} * (ATT - AP\_User_{ij}) + r_{ij}$$

Mathematics case study

Following the primary analysis including all 34 STEM courses in the sample, I focused on AP calculus as a case study to see whether the effect of using AP calculus as a pre-requisite is similar to the broader effect of using any AP credit. To answer research questions 2A and 3B, I conducted an additional series of analyses that focused specifically on 23 courses for which AP calculus serves as a pre-requisite. First, I estimated students' propensity to use AP calculus credit specifically, using the same predictors as I used for Model 2 with one exception. Instead of using the student's average AP exam score, I used their highest relevant AP calculus score (*APMScore)*. Once again, I also included dummy variables for each course to account for fixed cluster effects.

Model 4

$$logit(T_i = 1) = \beta_0 + \Sigma\beta_s X_{si} + \Sigma\beta_c Z_{ci} + r_i$$

In Model 4 above, the probability of using AP calculus credit ($T_i = 1$) is predicted with a vector of student-level factors noted as X_{si}, where *s* represents a student-level individual factor for student *i*, and a vector of fixed cluster effects noted as Z_{ci}, where *c* represents the student-level effects associated with each course. The model also includes the residual individual error term r_i.

Figure 2 below shows the distribution of treatment and control students across the spectrum of propensity scores when the treatment represents using AP calculus credit. As with the first propensity model, students who chose to use AP calculus credit have higher average propensity scores than those who did not, but most control students were still within the range of common support; approximately the same percentages of treatment and control students were included in the ATT and ATE propensity models for the case study as in the primary analysis (see Tables A5 and A6 in the appendix). Both weighting methods successfully balanced treatment and control groups, with no remaining variables significantly predicting treatment status. Tables A7 and A8 in the appendix include standardized mean differences between treatment and control students for each course, before and after weighting, for the ATT and ATE models, respectively.

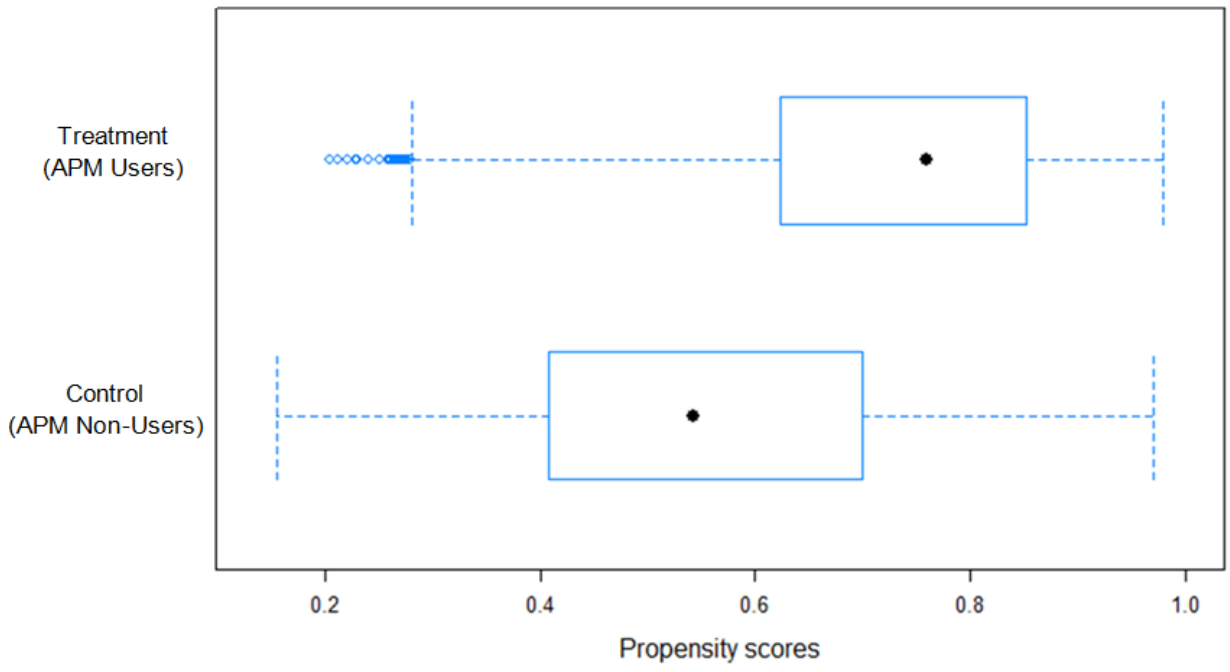

Figure 2: Case study: Distribution of propensity scores for both treatment and control students

Finally, I conducted an analysis with a two-level cross-sectional multilevel linear model (Model 5). The results with an unconditional model showed that 6.6% of the variation in target course grades was accounted for by courses. Next, I added the unweighted treatment indicator, APM_User. This time when I attempted to allow the effect of that variable to vary randomly across courses, the model failed to converge, so in subsequent models I kept APM_User as a fixed effect. I did not include any other level one predictors since the weighting process ensured they were sufficiently balanced across treatment and control groups (see Tables A7 and A8 in the appendix). When I tested the two proposed course-level predictors of DFW rate and a dummy variable indicating whether the target course also required another STEM course as a pre-requisite, only the former was a statistically significant predictor of target course grades. Therefore, Model 5 below is the full model for the case study analysis. In this model, $i = 1, 2, \ldots, n$ students in course j; $j = 1, 2, \ldots, 34$ courses; $ATT\text{-}APM\_User_{ij}$ = treatment status weighted by the ATT weights estimated in Model 4; u_{0j} is a residual course-level error term for the conditional average of grades for the jth course; and r_{ij} = a residual error term for student i in course j.

Model 5

$$GRADE_{ij} = \gamma_{00} + \gamma_{01} * (DFW_j) + \gamma_{10} * (ATT - APM\_User_{ij}) + u_{0j} + r_{ij}$$

I subsequently used the same model incorporating ATE weights as well. A summary of each phase of the study (i.e., preliminary, primary, case study), the samples of students and courses included in each, and the associated models answering each research question is found in Table 10 below.

Table 10: Summary of Study Analyses

Phase	Sample	Key Independent Variable(s)	Dependent Variable	Model	Research Question
Preliminary	All courses, All students	AP User, AP Non-User, Non-AP	Target course grades	1: Two-level cross sectional multilevel model	1
Primary	All courses, AP students	AP exam score, HS GPA, SAT, demographics	Use of AP Credit	2: Propensity model	2
Primary	All courses, AP students	AP User, weighted by ATT/ATE	Target course grades	3: Multilevel model with Model 2 weights	3, 3A
Case Study	Courses requiring calculus, AP Calculus students	AP exam score, HS GPA, SAT Math, demographics	Use of AP Calculus Credit	4: Propensity model	2A
Case Study	Courses requiring calculus, AP Calculus students	APM User, weighted by ATT/ATE	Target course grades	5: Multilevel model with Model 4 weights	3B

Model fit and diagnostics

There are six primary data assumptions for cross-sectional multilevel modeling; to ensure that my data met these assumptions, I ran diagnostics using the *lmerTest* (Kuznetsova et al., 2020) and *multilevelTools* (Wiley, 2020b) packages in R. The six primary data assumptions are as follows (Raudenbush & Bryk, 2002, p. 255):

1. Level one residuals are independently and normally distributed with a mean of zero and a common variance across all level two clusters.
2. Level one predictors and level one residuals are independent.
3. Random errors at level two are multivariate normally distributed.
4. Level two predictors and level two residuals are independent.
5. Level one residuals and level two residuals are independent.
6. Predictors at one level are independent of the residuals at the other level.

For the preliminary model, level one residuals were approximately normally distributed with a mean near zero. As was the case for all three phases of the study, the variance of level one residuals was not identical across all courses. This was due to the fact that course grades served as the outcome variable; in courses with lower DFW rates, the distribution of grades was much smaller than in courses with high DFW rates, so there was simply more room for error in some courses than others. Figure A1 in the appendix shows the box plots of residual variation across courses for the preliminary analysis. A violation of this assumption could lead to inaccurate estimates of the standard errors associated with γ terms and estimates of the variance components of the model (Raudenbush & Bryk, 2002). However the violation was not extreme in any of the three phases of this study, and none of the *p* values were close to 0.05, so I concluded this violation was not a major concern. All of the remaining data assumptions were met in the preliminary model with one exception; cluster averages for the level one AP Non-User predictor were significantly correlated with level two residuals, $r = 0.34$, $p = 0.05$. A violation of this assumption could mean the estimates for the fixed effect of the AP Non-User variable were biased (Raudenbush & Bryk, 2002); I do not think this is cause for major concern since the *p* values for this predictor in the final model were less than 0.01 so a Type I error is unlikely.

The diagnostics for the primary model and the case study model were nearly identical. Level one residuals were not quite as normally distributed as in the preliminary analysis; the distributions were negatively skewed, but means were both close to zero. Level one residual variances had a similar level of homogeneity across courses as in the preliminary model, i.e., not perfectly consistent but not so heterogenous as to cause major concerns (see Figures A2 and A3 in the appendix). All of the remaining data assumptions were met in both the primary model and the case study. Overall I concluded that it is safe to trust the results of the study without undue concern

that the data structure led to any major bias or instability in the models. Additionally, I compared model fit statistics for each version of the outcomes models in each phase of the study using Akaike's Information Criterion (AIC); the AIC balances model fit and complexity to estimate the relative quality of different models (Hamaker et al., 2011). Two models can be compared using the following formula (Hamaker et al., 2011):

$$\frac{AIC_1}{AIC_2} = \exp\left\{-\frac{1}{2}AIC_1 + \frac{1}{2}AIC_2\right\}$$

After completing each phase of the study, I compared AIC ratios of the models to assess the fit of the final model for the preliminary analysis, the primary analysis, and the case study.

CHAPTER 4: RESULTS

Introduction

The purpose of this chapter is to share the results of the analyses described in the previous chapter. It begins with the results of the preliminary analysis of the entire sample of students, which answers the first research question about how target course grades compare across three groups of students depending on whether they had earned, and then subsequently used, AP credit for course pre-requisites. Next, it will describe the results of the propensity model for the primary analysis, answering the second research question about which factors influenced the propensity of students to use earned AP credit. Then it will provide the results for the primary analysis outcomes model, answering the third research question regarding the effect of using AP credit on grades in subsequent courses, and the extent to which that effect varies across courses. Finally, the chapter will conclude with the results of both the propensity and outcomes models for the case study analysis, which answer follow-up parts of the second and third research questions regarding whether results are consistent when focusing only on the use of AP calculus as a STEM course pre-requisite.

Preliminary analysis

The first research question related to the preliminary analysis comparing grades in target courses across three groups: students who used AP credit for at least one course pre-requisite (AP Users), students who repeated any earned AP credit before completing a target course (AP Non-Users), and students without AP credit for course pre-requisites (No AP). As seen in the descriptive statistics provided in Table 11 below, average grades in target courses were highest for students who used their AP credit, followed by the group of students who repeated AP credit; the former

earned just above a B average and the latter earned between a B- and a B average. Students without any AP credit earned the lowest grades on average, halfway between a B and a C. All three groups had a wide range of grades, with all three standard deviations at a full grade point or higher. Table 11 also provides average grades by demographic group. Female students and international students earned higher grades than their male and domestic counterparts; underrepresented minority students, first generation students, and Pell grant recipients earned lower average grades than students not in those groups.

Table 11: Preliminary Analysis: Grade Distribution by AP and Demographic Groups

	N	*M*	*SD*	Skewness	Kurtosis
AP User	5,363	3.07	1.00	-1.201	1.108
AP Non-User	2,001	2.88	1.07	-0.972	0.461
No AP	19,479	2.55	1.10	-0.534	-0.315
Female	9,700	2.74	1.05	-0.655	-0.105
Male	17,143	2.65	1.13	-0.663	-0.237
URM	2,211	2.35	1.14	-0.377	-0.499
Non-URM	24,632	2.71	1.09	-0.696	-0.122
International	4,128	2.88	1.09	-0.919	0.234
Domestic	22,715	2.65	1.10	-0.629	-0.216
First Gen	4,601	2.42	1.15	-0.460	-0.498
Non-FG	22,242	2.74	1.08	-0.709	-0.083
Pell Recipient	4,938	2.53	1.15	-0.546	-0.426
Non-Pell	21,905	2.72	1.09	-0.692	-0.112

The results of the four multilevel models (MLMs) with Grade as the outcome variable (see Table 12) address the first research question. The estimates obtained with the second model, which included dummy predictors for student AP status but no other predictors, suggest that both AP groups outperformed the non-AP reference group by approximately half a letter grade, with AP Users earning slightly higher grades than AP Non-Users. Recall that the reference group in these

four models is the No AP group, whose average grade is represented by the intercept of each model, assuming zeros on all the predictor variables. The coefficient for the AP User indicator was 0.58, $t(23.59) = 12.57$, $p < 0.01$, and the coefficient for the AP Non-User indicator was 0.47, $t(8.32) = 13.21$, $p < 0.01$; the effect of both predictors varied significantly across courses.

After including SAT Math scores, high school GPA, and several demographic variables in Model 3, the gaps between both AP groups (indicated by the dummy variables) and the reference group were reduced approximately by half, but were still significant. On average, the gap between AP and non-AP students was 0.27 for AP Users, $t(23.44) = 6.99$, $p < 0.01$, and 0.26 for AP Non-Users, $t(47.99) = 9.58$, $p < 0.01$. The results of adding other predictors in Model 3 are mostly consistent with the descriptive statistics seen in Table 4 and Table 7; students in both AP groups had higher high school GPAs and SAT Math scores, and were less likely to be underrepresented minority students, first generation students, or Pell grant recipients. Once these potentially confounding variables were accounted for, the effect of being in an AP group was reduced. Historic DFW rate, the one course-level variable added in Model 4, had a significant main effect on target course grade, but there was no significant cross-level interaction effect between DFW rate and either of the two variables indicating AP status. This means the AP effect on course grade is independent of the course difficulty represented by the DFW rate, and the average effect of DFW rates on grades (i.e., the higher the DFW rate, the lower the average course grade) applies equally to all courses. After the addition of DFW rate at the course level, the estimates of the effects of each AP group did not change substantially.

Overall, the answer to the first research question is that after controlling for multiple covariates, students who had earned AP credit for course pre-requisites earned approximately one-quarter of a grade higher in target courses than non-AP students, whether they used their AP credit

or repeated it. The difference in the effects associated with the AP User and AP Non-User variables was negligible after the addition of other independent variables, suggesting no meaningful difference in grades between the two AP groups, on average. In the preliminary analysis, the final model (Model 4 in Table 12) identified variables that were significant predictors of final grades other than AP status, but it did not include AP exam scores, because over 70% of students in the sample (see Table 4) had not earned AP credit for the required course pre-requisite(s). The AIC ratio showed that Model 4 was a significantly better fit to the data than Model 3.

Table 12: Preliminary Analysis: Multilevel Regression Estimates Across Four Models of Student Grades in Target Courses

	Model 1 Unconditional Model		Model 2 Key Predictors: AP Groups		Model 3 Student Characteristics		Model 4 Full Model	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Fixed Effects								
Intercept (γ_{00})	2.82**	0.06	2.64**	0.06	-2.85**	0.11	-2.53**	0.14
DFW (γ_{01})							-2.19**	0.63
AP_User (γ_{10})			0.58**	0.05	0.27**	0.04	0.28**	0.04
AP_Non (γ_{20})			0.47**	0.04	0.26**	0.03	0.26**	0.03
Gender (γ_{30})					0.01	0.01	0.01	0.01
URM (γ_{40})					-0.05*	0.02	-0.05*	0.02
International (γ_{50})					0.16**	0.02	0.16**	0.02
FirstGen (γ_{60})					-0.21**	0.02	-0.21**	0.02
Pell (γ_{70})					-0.05*	0.02	-0.05*	0.02
HSGPA (γ_{80})					0.87**	0.02	0.87**	0.02
SATM (γ_{90})					0.03**	0.00	0.03**	0.00
Variance Estimates								
Intercept (τ^2_0)	0.32		0.33		0.37		0.31	
AP_User slope (τ^2_1)			0.22		0.18		0.18	
AP_Non slope (τ^2_2)			0.12		0.06		0.06	
Within-student (σ^2)	1.07		1.04		0.96		0.96	
AIC	79905.99		78563.69		74411.08		74402.47	

** $p < .01$; * $p < .05$

Primary analysis

Propensity model

The second research question addresses the propensity to use earned AP credit, and which factors predict whether students will choose to use their credit without repeating the pre-requisite course in college. In addition to all the student-level predictors from the preliminary analysis, the propensity model for the primary analysis included average AP exam score, as the students in this subset of the original sample all had AP exam scores. The propensity model also included dummy variables representing student enrollment in First-Year Engineering, one of the Engineering professional schools, or the College of Science (with all other colleges serving as the reference group), and dummy variables for each course to provide estimates of fixed student-level effects associated with the course. In all, the propensity model included 44 predictors, 12 of which combined for nearly 95% of the relative influence on propensity to use AP credit (see Table 13 below).

Table 13: Primary Analysis: Relative Influence of Variables on Estimating Propensity Scores

Variable	Variable Description	Relative Influence (%)
C30	Course variable: Theoretical Calculus 2B	33.58
AvgScore	Average relevant AP exam score	16.69
SATM	SAT Math scores	11.86
EngrS	Dummy variable: Engr. Prof. School Student	6.32
SciS	Dummy variable: College of Science Student	5.55
C29	Course variable: Theoretical Calculus 2A	5.08
C28	Course variable: Applied Calculus 2	4.36
HSGPA	High School GPA	4.14
FYES	Dummy variable: FYE Student	2.83
C31	Course variable: Calculus 3	2.13
C18	Course variable: Organic Chemistry C	1.22
International	Dummy variable: International student	1.05

The fixed cluster effects for Theoretical Calculus 2B had twice the influence of any other factor on a student's propensity to use AP credit; the standardized mean difference in target grades between treatment and control students for that class was -0.710 prior to applying propensity weights. Average AP score, SAT Math score, and high school GPA were also among the most influential predictors of treatment status, along with the variables indicating students were enrolled in one of the Engineering professional schools, the College of Science, or First-Year Engineering at the time they took the target course. The other two versions of Calculus 2 were also influential, but the only demographic variable that accounted for at least 1% of the influence on propensity for treatment was whether the student was international rather than a domestic student. It is notable that Theoretical Calculus 2B was more than six times as influential as Theoretical Calculus 2A, despite the fact that the two courses are interchangeable in terms of how they can be used to fulfill degree requirements at this institution.

Table 14 below shows the pre- and post-weighting treatment and control group means for each of the 12 most influential variables. Higher AP exam scores, SAT Math scores, and high school grades all increased the propensity to use AP credit, as did being enrolled in one of the Engineering professional schools or the College of Science. International students had greater propensity to use AP credit than domestic students. All three Calculus 2 courses, however, were associated with a lower propensity to use AP credit, as was being enrolled in First-Year Engineering. Balancing treatment and control groups involves the use of ATT or ATE weights as described in Chapter 3. Because Table 14 shows the result of ATT weighting, the treatment group means do not change; control students, however, are weighted such that those control students who are more similar to treatment students on the variables included in the model have greater influence in the analysis. For example, while the pre-weighting control group had a mean of 0.276 on the

dummy variable associated with Theoretical Calculus 2B (compared to the treatment group's mean of 0.082), after weighting, the control group mean was 0.090, resulting in a non-significant difference between groups. The full balance table that includes all 44 predictors is available in the appendix (see Table A3); the procedure to balance treatment and control groups using ATE weights was similarly successful (see Table A4).

Table 14: Primary Analysis: ATT Balance Table Before and After Weighting

	Before Weighting						After Weighting					
	Treatment Group (AP Users)		Control Group (AP Non-Users)				Treatment Group (AP Users)		Control Group (AP Non-Users)			
Variable	*M*	*SD*	*M*	*SD*	*Std ES*	*p*	*M*	*SD*	*M*	*SD*	*Std ES*	*p*
C30	0.082	0.274	0.276	0.447	-0.710	<0.001	0.082	0.274	0.090	0.287	-0.031	0.093
AvgScore	4.660	0.468	4.454	0.514	0.439	<0.001	4.660	0.468	4.649	0.478	0.023	0.384
SATM	739.66	48.46	718.12	50.95	0.445	<0.001	739.66	48.46	737.38	48.53	0.047	0.090
EngrS	0.295	0.456	0.264	0.441	0.067	0.001	0.295	0.456	0.310	0.463	-0.034	0.223
SciS	0.194	0.396	0.170	0.376	0.062	0.003	0.194	0.396	0.178	0.383	0.042	0.211
C29	0.166	0.372	0.203	0.402	-0.099	<0.001	0.166	0.372	0.184	0.388	-0.049	0.060
C28	0.058	0.233	0.106	0.307	-0.206	<0.001	0.058	0.233	0.066	0.248	-0.034	0.107
HSGPA	3.702	0.279	3.660	0.280	0.150	<0.001	3.702	0.279	3.701	0.276	0.004	0.894
FYES	0.358	0.479	0.387	0.487	-0.062	0.004	0.358	0.479	0.361	0.480	-0.007	0.796
C31	0.148	0.356	0.078	0.268	0.199	<0.001	0.148	0.356	0.167	0.373	-0.051	0.138
C18	0.005	0.072	0.013	0.115	-0.112	<0.001	0.005	0.072	0.005	0.074	-0.002	0.901
Int’l	0.095	0.293	0.063	0.243	0.110	<0.001	0.095	0.296	0.101	0.302	-0.021	0.516

Note: See Table 13 for descriptions of each of the variables listed by abbreviations in this table

Overall, the answer to the second research question is that while measures of prior academic achievement were important factors predicting the propensity to use earned AP credit as expected, the decision to use AP credit is at least as much about the specific course as it is about student preparation. The three versions of Calculus 2 were associated with a much lower propensity to use AP credit, regardless of student academic characteristics. Additionally, the student's home college at the time the target course was completed is an important predictor of propensity scores, with all three dummy variables among the twelve most influential factors. Beyond these contextual factors, average AP exam scores and SAT Math scores were also highly influential, and had greater influence on propensity to use AP credit than high school GPA. Gender, underrepresented minority status, first-generation status or Pell grant recipient status were not influential factors for the decision; the only influential demographic variable was the dummy variable for International students.

Outcomes model

The third research question aims to identify the effect of using AP credit as a course pre-requisite on target course grades, by using propensity weights to account for pre-treatment differences in the student groups that could potentially affect both treatment selection and the eventual outcome. As seen in Table 15, the descriptive results show that students who used AP credit earned higher target course grades, on average, than AP non-users; their average grades in this phase of the study were nearly identical to the two group averages in the preliminary study, with AP users earning just above a B average and AP non-users earning between a B- and a B average. The differences in grades by demographic group were also similar to what was seen in the preliminary study, although overall, average grades for each group were approximately a third

of a letter grade higher (not surprising given that in this sample, all students had earned AP credit, and in the preliminary study, students with AP credit earned higher grades than non-AP students).

Table 15: Primary Analysis: Grade Distribution by AP and Demographic Groups

	N	*M*	*SD*	Skewness	Kurtosis
AP User	7,019	3.04	0.99	-1.10	0.90
AP Non-User	3,133	2.88	1.02	-0.85	0.35
Female	2,858	3.08	0.95	-1.09	0.99
Male	7,294	2.96	1.02	-0.98	0.57
URM	549	2.77	1.08	-0.71	-0.01
Non-URM	9,603	3.00	1.00	-1.03	0.75
International	865	3.20	0.96	-1.49	2.11
Domestic	9,287	2.97	1.00	-0.98	0.61
First Gen	1,186	2.78	1.13	-0.86	0.08
Non-FG	8,966	3.02	0.98	-1.03	0.75
Pell Recipient	1,479	2.91	1.04	-0.91	0.39
Non-Pell	8,673	3.00	1.00	-1.03	0.74

The results of the four MLM models are shown in Table 16 below. For this set of analyses, AP Non-Users (the control group) served as a reference group, thus the intercept of each model represents their average grade, assuming zeros on predictor variables. First, the effect of the unweighted AP User variable was significant, and indicated that students who used AP credit for at least one course pre-requisite earned, on average, one-tenth of a GPA point higher in target courses (Model 2). Once I applied the ATT propensity weights to Model 3, there was no longer any significant effect of using AP credit ($p = 0.77$). This held true after adding the only significant course-level predictor (DFW rates) in Model 4. The AIC ratio showed that Model 4 was a better fit to the data than Model 3, but Model 2 was the best fit; this is reasonable given that the weighted treatment indicator is not a significant predictor of target course grades.

Next, the second part of the third research question asked to what extent the effect of using AP credit varied across courses, and a comparison of Model 2 with and without fixing the effect of the AP User variable showed that there was significant variation in the effect across courses, χ^2 (2) = 14.63, $p < .001$. This indicates that for some courses, there could exist a causal effect of using AP credit. However, there were no significant cross-level interaction effects between the AP User predictor and course DFW rates, so course difficulty cannot explain the variation in the effect of using AP credit across courses. This implies that other unknown course-level differences may possibly explain the variation in the AP User effect across courses.

Students who use AP credit have, on average, higher AP exam scores, high school GPAs, and SAT Math scores, all of which were shown in the preliminary model to be associated with higher grades in target courses. After accounting for those and other differences between treatment and control students with the use of propensity weights, these results show, on average, there is no effect on grades of choosing to use earned AP credit to fulfill course pre-requisites.

Table 16: Primary Analysis: ATT-Weighted Multilevel Regression Estimates Across Four Models of Student Grades in Target Courses

	Model 1 Unconditional Model		Model 2 Key Predictor: AP User		Model 3 AP User ATT weights		Model 4 Full Model ATT weights	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Fixed Effects								
Intercept (γ_{00})	3.21**	0.05	3.13**	0.06	3.24**	0.06	3.54**	0.10
DFW (γ_{01})							-2.02**	0.57
AP_User (γ_{10})			0.10*	0.04	-0.02	0.05	-0.01	0.05
Variance Estimates								
Intercept (τ^2_0)	0.25		0.28		0.31		0.28	
AP_User slope (τ^2_1)			0.15		0.21		0.19	
Within-student (σ^2)	0.97		0.96		1.10		1.34	
AIC	28197.60		28177.31		29906.84		29897.01	

** $p < .01$; * $p < .05$

Using ATT weights allows for estimating the effect of treatment on those who currently select into treatment, and other students who are similar to them in terms of pre-treatment characteristics that were incorporated into the propensity model. Using ATE weights to assess the overall average treatment effect makes it possible to estimate the effect of using AP credit as a course perquisite on a broader population, as long as there is sufficient overlap in the range of propensity weights for the two groups (i.e., the effect estimate will not apply to control group students whose propensity for treatment is very low). Table 17 below shows the results of applying ATE weights to the same model as was shown in Table 16; the results are nearly identical except that the addition of DFW rates meant the model failed to converge, so the third model is the final version. Even when using ATE weights, there was no significant effect associated with the use of AP credit to fulfill course pre-requisites ($p = 0.74$). The AIC ratio showed that Model 3 was not as good a fit to the data, because the weighted treatment indicator was not a significant predictor of target course grades.

Table 17: Primary Analysis: ATE-Weighted Multilevel Regression Estimates Across Four Models of Student Grades in Target Courses

	Model 1 Unconditional Model		Model 2 Key Predictor: AP User		Model 3 AP User ATE weights		Model 4 Full Model ATE weights	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Fixed Effects								
Intercept (γ_{00})	3.21**	0.05	3.13**	0.06	3.21**	0.06		
DFW (γ_{01})							Failed to converge	
AP_User (γ_{10})			0.10*	0.04	-0.01	0.04		
Variance Estimates								
Intercept (τ^2_0)	0.25		0.28		0.30			
AP_User slope (τ^2_1)			0.15		0.18			
Within-student (σ^2)	0.97		0.96		1.34			
AIC	28197.60		28177.31		29681.53			

Interpretation of ATE results

The goal of using ATE weights as well as ATT weights was to estimate the effect of using AP credit on subsequent course grades if a broader range of students chose to use their credit beyond the group of students who typically chose to do so. While there was no effect of using AP credit after applying ATE weights, this does not mean that the average treatment effect applies to all students with AP credit. Figure 1 showed the distribution of propensity scores for students who used AP credit (treatment) and those who did not (control). One assumption of the potential outcomes framework that the use of propensity scores relies on is overlapping distributions; for all possible values of the covariates that influence propensity for treatment, there are both treated and control units (Arpino & Mealli, 2011). This means that the ATE only applies to students who fall

within the propensity score distribution for control students that overlaps with that of treatment students.

To understand which control students would be similar enough to treatment students that the ATE would apply to them, I reviewed the propensity model summary (see Table A2 in the appendix) which indicated that approximately 2,060 control group students had propensity scores that overlapped with the treatment group and were therefore included in the ATE propensity model. I compared this group (high-propensity control students) to current AP users, and to the group of control students who were not included in the ATE propensity model due to propensity scores that fell below the region of common support, to get a sense of what academic profile would indicate that a student should expect to see no difference in target course grades whether they used AP credit or not. I found that high school GPA values were nearly unchanged across all three groups: current AP users had an average high school GPA of 3.70, current non-users with higher propensity scores averaged 3.67, and lower propensity non-users had an average high school GPA of 3.63. There were greater differences in SAT Math scores and average AP exam scores. The three groups (current AP Users, higher-propensity AP Non-Users, and lower-propensity AP non-users) had average SAT Math scores of 740, 727, and 700 respectively; their average AP exam scores were 4.66, 4.53, and 4.28 respectively. Students whose academic profile is similar to that of the higher-propensity AP Non-User group should therefore expect to see no difference in grades whether they used AP credit or not, based on the results of the ATE-weighted outcomes model (Table 17, above). Additionally, given the fact that all three Calculus 2 courses were highly influential in the propensity model, I repeated both the propensity model and the outcomes model without any Calculus 2 courses to evaluate the validity of the findings. The results were unchanged, indicating that the conclusions were not overly influenced by those specific courses.

Case study

The propensity model and outcome model in the primary analysis both included all target courses in the sample. The remaining research questions asked whether the findings of those analyses would remain consistent when focusing only on the use of AP calculus as a course pre-requisite. Therefore, the final set of analyses serves as a case study about the use of AP calculus; the need for a separate analysis is supported by the results of the first propensity model, which showed that the three Calculus 2 courses were important predictors of the propensity to use AP credit.

Propensity model

The case study propensity model incorporated all of the same predictor variables as the primary propensity model, with two exceptions. First, rather than creating an average AP exam score, the model included the highest relevant AP calculus exam score. Second, because not all of the courses required calculus as a pre-requisite, there were only 22 dummy variables representing fixed course-level effects. As in the primary case study propensity model, there were twelve predictors that accounted for at least 1% of the propensity to use AP credit; this time those 12 predictors together accounted for over 97% of the relative influence on the propensity model (see Table 18 below).

Table 18: Case Study: Relative Influence of Variables on Propensity Scores

Variable	Variable Description	Relative Influence (%)
C30	Course variable: Theoretical Calculus 2B	29.49
SATM	SAT Math scores	15.64
APM Score	AP Calculus exam score	15.41
EngrS	Dummy variable: Engr. Prof. School student	6.10
SciS	Dummy variable: College of Science student	5.22
HSGPA	High school GPA	4.80
C28	Course variable: Applied Calculus 2	4.77
C31	Course variable: Calculus 3	4.38
C29	Course variable: Theoretical Calculus 2A	4.10
FYES	Dummy variable: FYE student	3.58
C32	Course variable: Linear Algebra	2.53
International	Dummy variable: International student	1.12

Overall, the case study propensity model is quite similar to the original propensity model; all three Calculus 2 course fixed effect predictors remained highly influential, as were SAT Math scores, AP Calculus exam scores, high school GPA, all three school/college dummy variables, and the International indicator. There were a few differences, as SAT Math scores gained approximately 4 percentage points of relative influence, overtaking AP Calculus exam scores for the second-most influential position. High school GPA also became relatively more influential than any course cluster effect other than Theoretical Calculus 2B, but the actual percentage changed very little. Additionally, the course indicator for Linear Algebra replaced the indicator for Organic Chemistry C. The indicator for Theoretical Calculus 2B enrollment was still the most influential, but somewhat less so than in the primary propensity model. Therefore, the answer to the second part of research question two is that the findings of the two propensity models are consistent; whether estimating the propensity to use any AP credit as a STEM course pre-requisite

or estimating the propensity to use AP Calculus as a STEM course pre-requisite, the same factors have the most influence on propensity scores.

Table 19 below shows the pre- and post-weighting treatment and control group means for each of the twelve most influential variables in the case study propensity model. As before, higher AP exam scores, SAT Math scores, and high school grades all increased the propensity to use AP credit, as did being enrolled in one of the Engineering professional schools or the College of Science. And once again, international students had greater propensity to use AP credit than domestic students. While all three Calculus 2 courses were still associated with a lower propensity to use AP credit, Calculus 3 enrollment was associated with a higher propensity to do so. First-Year Engineering students continued to have a lower propensity for using AP credit. The full balance table is available in the appendix (see Table A7); the procedure to balance treatment and control groups using ATE weights was similarly successful (Table A8).

Table 19: Case Study: ATT Balance Table Before and After Weighting

	Before Weighting						After Weighting					
	Treatment Group		Control Group		Std ES	*p*	Treatment Group		Control Group		Std ES	*p*
Variable	*M*	*SD*	*M*	*SD*			*M*	*SD*	*M*	*SD*		
C30	0.100	0.301	0.287	0.452	-0.621	<0.001	0.100	0.301	0.109	0.312	-0.029	0.144
SATM	742.10	47.04	720.63	50.07	0.456	<0.001	742.10	47.04	740.04	46.96	0.044	0.117
APMScore	4.715	0.451	4.511	0.500	0.453	<0.001	4.715	0.451	4.704	0.457	0.026	0.311
EngrS	0.278	0.448	0.266	0.442	0.027	0.214	0.278	0.448	0.294	0.456	-0.035	0.197
SciS	0.198	0.399	0.174	0.379	0.060	0.005	0.198	0.399	0.182	0.386	0.041	0.199
HSGPA	3.701	0.279	3.661	0.278	0.145	<0.001	3.701	0.279	3.702	0.277	-0.002	0.958
C28	0.070	0.254	0.119	0.324	-0.195	<0.001	0.070	0.254	0.076	0.266	-0.027	0.218
C31	0.168	0.374	0.091	0.287	0.206	<0.001	0.168	0.374	0.181	0.385	-0.035	0.292
C29	0.186	0.389	0.207	0.406	-0.055	0.014	0.186	0.389	0.204	0.403	-0.046	0.087
FYES	0.394	0.489	0.410	0.492	-0.033	0.138	0.394	0.489	0.399	0.490	-0.009	0.756
C32	0.073	0.260	0.031	0.173	0.161	<0.001	0.073	0.260	0.054	0.226	0.072	0.098
Int’l	0.098	0.298	0.066	0.249	0.108	<0.001	0.098	0.298	0.108	0.310	-0.031	0.395

Note: See Table 18 for descriptions of each of the variables listed by abbreviations in this table.

Outcomes model

The last analysis in this study applied the propensity weights from the model above to another cross-sectional multilevel model to estimate the effect of using AP Calculus credit on grades in target courses that required calculus as a pre-requisite. The distribution of grades by AP group was similar to the grade distribution in the primary analysis, although grades were slightly lower (see descriptive statistics in Table 20 below). Considering demographic groups, female students, underrepresented minority students, first generation students, and Pell recipients all earned slightly lower grades, on average, in classes requiring calculus as a pre-requisite than they did across all STEM courses in the primary model, although female students continued to earn higher grades than their male counterparts.

Table 20: Case Study: Grade Distribution by AP and Demographic Groups

	N	*M*	*SD*	Skewness	Kurtosis
AP User	6,283	3.03	0.99	-1.09	0.90
AP Non-User	3,128	2.85	1.02	-0.83	0.29
Female	2,500	3.04	0.95	-1.01	0.79
Male	6,911	2.95	1.02	-0.98	0.58
URM	509	2.74	1.06	-0.70	0.02
Non-URM	8,902	2.99	1.00	-1.01	0.70
International	825	3.23	0.91	-1.44	2.12
Domestic	8,586	2.95	1.01	-0.96	0.57
First Gen	1,087	2.73	1.15	-0.82	-0.04
Non-FG	8,324	3.00	0.98	-1.00	0.71
Pell Recipient	1,300	2.87	1.06	-0.91	0.35
Non-Pell	8,111	2.99	0.99	-1.01	0.70

The results of the final four MLM models are shown in Table 21 below. The effect of the unweighted APM User variable in Model 2 was significant, and indicated that students who used

AP Calculus credit for at least one course pre-requisite earned, on average, just over one-tenth of a GPA point higher in target courses. Consistent with the primary analysis, after applying ATT propensity weights to Model 3, there was no longer any significant effect associated with using AP Calculus credit ($p = 0.33$). This held true after adding the only significant course-level predictor (DFW rates) in Model 4; the other course-level predictor I considered adding to the model (an indicator of a second STEM course pre-requisite) was not significant. Finally, the last part of the third research question asked to what extent the effect of using AP Calculus credit varied across courses. Unlike in the primary analysis, the effect of using AP credit did not vary significantly across courses when only AP Calculus was studied.

Table 21: Case Study: ATT-Weighted Multilevel Regression Estimates Across Four Models of Student Grades in Target Courses

	Model 1 Unconditional Model		Model 2 Key Predictor: APM User		Model 3 APM User ATT weights		Model 4 Full Model ATT weights	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Fixed Effects								
Intercept (γ_{00})	3.14**	0.06	3.05**	0.06	3.12**	0.06	3.51**	0.11
DFW (γ_{01})							-2.49**	0.67
APM_User (γ_{10})			0.12**	0.02	0.02	0.02	0.02	0.02
Variance Estimates								
Intercept (τ^2_0)	0.26		0.25		0.26		0.20	
Within-student (σ^2)	0.97		0.97		1.10		1.10	
AIC	26219.97		26197.42		27791.56		27781.03	

** $p < .01$; * $p < .05$

While the exact estimate of possible AP effects was slightly different in the mathematics case study, overall the results are indeed consistent. After weighting by the propensity to use AP

Calculus credit, there was no significant effect on grades for those students who typically choose to use earned AP Calculus credit. I checked to see if this conclusion would still hold true using ATE weights instead, and as shown in Table 22, the results did not change.

Table 22: Case Study: ATE-Weighted Multilevel Regression Estimates Across Four Models of Student Grades in Target Courses

	Model 1 Unconditional Model		Model 2 Key Predictor: APM User		Model 3 APM User ATE weights		Model 4 Full Model ATE weights	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Fixed Effects								
Intercept (γ_{00})	3.14**	0.06	3.05**	0.06	3.10**	0.06	3.51**	0.12
DFW (γ_{01})							-2.66**	0.69
APM_User (γ_{10})			0.12**	0.02	0.00	0.02	0.00	0.02
Variance Estimates								
Intercept (τ^2_0)	0.26		0.25		0.26		0.20	
Within-student (σ^2)	0.97		0.97		1.36		1.36	
AIC	26219.97		26197.42		27574.55		27563.09	

Results summary

In summary, the results showed that students who earned AP credit prior to college perform better, on average, in subsequent STEM courses than students who did not earn AP credit, even after controlling for measures of prior academic achievement and demographic characteristics. Additionally, while students who used AP credit earned higher grades, on average, than students who repeated AP credit, that difference was mostly due to differences between AP Users and AP Non-Users on pre-treatment variables (including high school GPA, SAT Math scores, and AP exam scores) that were associated with both target course grades and the propensity to use earned

AP credit. After applying the use of ATT or ATE propensity weights, the model showed no significant causal difference in grades between students who used AP credit and those who repeated AP credit prior to enrolling in subsequent courses. In addition to academic background variables, a student's home school or college and their enrollment in specific courses, particularly Calculus 2, were among the most important factors predicting the propensity to use AP credit. This finding indicates that the decision to use AP credit or not is highly related to a student's specific context, and students may choose differently depending on the course, or on factors related to their academic home, even with the same levels of academic preparation as measured by high school grades and test scores. Both the findings of the propensity model and the outcomes model were consistent when focusing only on the use of AP Calculus in the case study.

CHAPTER 5. DISCUSSION

Introduction

In this chapter, I will first review the key findings of each phase of the study, connecting what was learned in this study to prior research on AP and college student success. Then, I will address the implications of the results, both for individual students who need to decide whether to use earned AP credit, and for college and university policies and practices regarding the use of AP credit. Next, I will discuss the scholarly contributions of the study, and some of the methodological decisions I made that could inform similar studies. Finally, I will discuss the potential limitations for drawing conclusions based on the results of the study, and suggest avenues for future research.

Key findings of the study

Preliminary analysis

The preliminary analysis compared grades of three groups of students: those who used AP credit to meet course pre-requisites, those who repeated pre-requisite courses after earning AP credit, and those who did not earn AP credit and had to take pre-requisite courses for the first time at the college level. The results showed that both groups of students with AP credit earned higher grades in target courses than non-AP students, even after controlling for differences in prior academic achievement and demographic background. While this phase of the analysis did not include AP exam scores, students in this study with AP credit nearly all earned exam scores of 4 or higher, due to institutional policy regarding the scores required to earn credit for pre-requisite courses. Therefore, the finding that they surpassed non-AP students by approximately a quarter of a letter grade is consistent with prior research showing students with AP scores of 4 or higher

significantly outperform non-AP students even after controlling for other variables (e.g., Ackerman et al., 2013; Morgan & Klaric, 2007). Notably, the initial estimates for both AP predictors were much higher prior to adding other variables to the model. The fact that the estimates of AP effects were reduced approximately by half after accounting for high school grades, SAT Math scores, and other covariates, supports the contention of AP scholars that simple comparisons between AP and non-AP students are insufficient due to important differences between the groups other than their AP status (Dougherty et al., 2006; Sadler, 2010a).

The preliminary analysis also showed that in addition to reducing the gap between AP and non-AP students, controlling for academic and demographic covariates reduced the gap between students who used their AP credit and those who repeated it. The initial model suggested that AP Users earned approximately a tenth of a letter grade higher than AP Non-Users, but this advantage disappeared in the final model that included other predictors. Given that on average, students who chose to use AP credit had higher high school grades and SAT Math scores than those who did not, it appears that all of the difference in grades between these two groups was accounted for by the higher academic profile of the AP User group. The few other studies that investigated outcomes of AP repeaters similarly found no difference in grades based on whether or not a student used their AP credit (De Urquidi et al., 2015; Murphy & Dodd, 2009).

Primary analysis

While the main purpose of identifying the student characteristics associated with propensity to use AP credit was to understand the causal effect of using AP credit on target course grades, the results of the propensity analyses were interesting on their own as well. For the primary propensity model, I investigated 44 predictors and found that 12 were responsible for nearly all of the influence on a student's propensity to use AP credit. Among the most important factors were

AP exam scores, SAT Math scores, and high school GPA, which is consistent with the limited earlier research on a student's choice to use earned AP credit (De Urquidi et al., 2015). One of the potentially surprising results is that demographic factors including gender, underrepresented minority status, first generation status, and Pell grant recipient status were not important factors in predicting the use of AP credit. Prior research has shown that low-income and underrepresented minority students have lower levels of college achievement than their higher-income and white counterparts (Burns et al., 2019; Daugherty et al., 2006), and that female students typically outperform male students in college (Burns et al., 2019; Conger et al., 2009), but a previous study on how students used AP credit (Evans, 2019) found little difference by race or gender in the number of AP credits earned after accounting for the institution attended and other academic characteristics. Similarly, in this study, the only demographic factor that was in the top twelve most influential predictors was whether the AP student was International or not, which accounted for just over 1% of the propensity to use AP credit. De Urquidi et al. (2015) found that gender was a small, but significant predictor of the use of AP credit, but gender was not an influential factor predicting that choice in this study. Educational research typically finds differences in academic outcomes by gender and ethnicity (Chajewski et al., 2011), but in this case, when the outcome was the propensity to use AP credit, there were no such differences. Even prior to the weighting procedure, the group of AP users did not differ significantly from the group of AP non-users based on gender, underrepresented minority status, first generation status, or Pell grant recipient status.

Another factor that the propensity models showed was important in predicting the use of AP credit is the student's home college or school. It is clear from prior research that local context matters when discussing the use of AP credit, whether due to informal advice from faculty and advisors (e.g., De Urquidi et al., 2015) or departmental policies requiring placement exams in

addition to AP credit in order to move directly to subsequent coursework (Sadler & Tai, 2007). Therefore, while this finding is new, it is not unexpected that the propensity to use AP credit would be influenced by factors related to the student's home college or school. Specifically, this study found that being enrolled in First-Year Engineering was associated with a lower propensity to use AP credit. At this institution, future engineers are admitted to FYE rather than directly into an Engineering professional school, and must compete with other students for admission into selective professional schools for their sophomore year. Students in FYE may have been motivated to retake AP courses because earning high grades in introductory courses was a high priority while their admission to an Engineering major was still uncertain, and they expected to earn higher grades by repeating familiar material. This is consistent with earlier research suggesting that new STEM students are advised to retake AP courses even when not required (Sadler & Sonnert, 2018). In contrast, students who were enrolled in one of the Engineering professional schools or the College of Science at the time they took the target course were more likely to use AP credit for course pre-requisites. Students in Engineering and Science would mostly be already admitted to their major of choice, and therefore may have felt less pressure to maximize their chances of earning higher grades by repeating a course.

The other five most important factors influencing propensity scores were the variables associated with specific target courses. By far the most important factor, with twice the influence of the second-strongest predictor (average AP exam scores), was enrolling in Theoretical Calculus 2B (a four-credit course that is otherwise equivalent to the five-credit Theoretical Calculus 2A in content and pre-requisites). The next two most influential factors associated with course enrollment were the two other versions of Calculus 2 offered at the institution (Theoretical Calculus 2A and Applied Calculus 2); all three courses were associated with significantly lower

propensity to use earned AP credit. Across the three versions of Calculus 2 combined, students were slightly more likely to use AP credit than not, but the percentage of students in each course who chose to repeat their credit was much higher than in other courses in the sample. This pattern was not consistent for all mathematics courses in the study, as enrolling in either Calculus 3 or Linear Algebra was associated with a higher propensity to use AP credit. Although there is nothing in the existing AP research that could help explain this finding about Calculus 2 as a predictor of lower propensity to use AP credit, I can propose two possible explanations. First, mathematics courses are required for all STEM careers (Warne et al., 2019), and college calculus in particular is a gateway for most students pursuing STEM majors (Ellis et al., 2016). If students think it is especially important to gain a deep understanding of calculus in order to be successful in their chosen majors, they may decide to retake Calculus 1 before moving on to Calculus 2; indeed, students are frequently advised to do just that (Sadler & Sonnert, 2018). Second, there are two AP Calculus courses: Calculus AB (Calculus 1), and Calculus BC (Calculus 1 and 2). As explained in Chapter 3, successful Calculus BC students who decided to use their credit for Calculus 1, but repeated Calculus 2, were not included in the sample of students enrolled in Calculus 2 as a target course, because they had encountered the material previously. Therefore, the only Calculus 2 students included in the study were those who had earned AP credit for Calculus 1 but not Calculus 2. It is possible that those students compared themselves to their more advanced peers, who had earned a full year of college calculus credit, and were less confident in their ability to be successful in Calculus 2 as their first college mathematics course. This possibility is supported by the fact that enrollment in Calculus 3 was also among the most influential factors in the propensity model, and those students (who must have taken the more rigorous Calculus BC course) were more likely to use AP credit. This finding about Calculus 2 requires further investigation to understand if it is

consistent across other institutions, if it is due to policy or practice at the university or department level, or if it is related to the methodological choices I made in this study.

The results of the outcomes models were simpler than the propensity models, but are perhaps of greater consequence for student success. The purpose of the primary analysis was to understand if there were differences in grades depending on whether students used their AP credit or repeated it prior to taking subsequent courses. I found three perspectives regarding this decision in the AP literature. First, there are faculty skeptics who assert that AP courses provide inadequate preparation for further study, and that students should repeat credit in college in order to be successful in subsequent coursework (e.g., Hansen et al., 2006; Scott et al., 2010). Second, Casserly (1986) suggested that repeating AP credit could lead to boredom and reduced motivation, and recommended that students use their AP credit to avoid these problems related to "underplacement." Finally, the concept of educational dose would suggest that additional exposure to pre-requisite material would lead to better outcomes, so even if AP students could pass subsequent courses, they would earn higher grades if they repeated pre-requisites in college (Warne et al., 2019). The results of the preliminary analysis could be interpreted as support for this third perspective. In that model, which did not include AP exam scores, there was almost no difference in grades between AP Users and AP Non-Users. Controlling for SAT Math scores and high school GPA had eliminated the initial advantage for AP Users, so it would be reasonable to expect that after accounting for the higher AP exam scores earned by AP Users, students who used their AP credit would earn lower grades in target courses. Instead, the primary outcomes model showed no effect associated with using AP credit. The non-significant result for this phase of the analysis actually is quite meaningful, because it suggests that the use of AP credit worked as originally intended, allowing students who were ready for advanced work in a subject to perform

just as well as if they had completed course pre-requisites in college (Lichten, 2000). There was no advantage or disadvantage to using AP credit; AP students were successful, on average, regardless of the choices they made.

The results of the outcomes models also showed that both before and after weighting based on propensity scores, the effect of using AP credit did vary across courses. This means that while on average there is no effect of using AP credit, it could be that some courses have a significant positive effect associated with the use of credit and others have a significant negative effect. I did find that historical DFW rate was a significant predictor of course grades after other variables were included in the model, but I was primarily interested in course-level variables to see if, for example, the use of AP credit in more difficult courses resulted in better or worse outcomes than the use of AP credit in easier courses. However, none of the course-related variables I tested could explain how or why the effect of using AP credit was different across courses in this study. Most of the overall variation in grades was due to differences at the student level rather than the course level, so it may be that there simply was not enough variation to explain with course-level predictors. While somewhat disappointing, this finding regarding how the AP effect varies across courses is at least consistent with prior studies that used cross-sectional multilevel modeling to investigate AP effects, given that most studies focused almost entirely on student-level effects and did not explain how effects varied across groups (e.g., Patterson et al., 2011). This study provided empirical support for a student-focused modeling approach in AP research by systematically investigating and quantifying the course effect on grades in subsequent courses.

Case study

It was not the original intent, but in the sample for the primary analysis, the courses that required calculus as a pre-requisite also tended to be the largest courses, and students in the sample

were far more likely to have credit for AP Calculus than for the other three subjects combined. Therefore, with few exceptions, such as relatively greater influence of SAT Math scores on the propensity to use AP credit, the results of the AP Calculus case study were consistent with the primary analysis. I thought that a focus on AP Calculus might provide some insight into the finding that students were less likely to use AP credit for Calculus 2. However, there were no course-level variables, related to Calculus 2 or otherwise, that could explain any differences in AP effects across courses.

Implications for policy and practice

The results of this study have implications for students with AP credit and for the colleges and universities they attend. First, students who choose to use their AP credit should feel confident that they are likely to be as successful in subsequent courses as they would have been had they repeated the AP credit first. The College Board promotes the AP program as a means of speeding time to degree (College Board, n.d., *Benefits of AP*), but graduating more quickly because of success in AP exams is only possible if students use the credit they earned. Students may believe that they have to choose between two possible benefits of the AP program, higher grades or reduced time to degree (Klopfenstein, 2010), but this study shows they are not likely to earn lower grades by using their AP credit. Academic advisors, who previously have had to provide guidance regarding the use of AP credit based on anecdotal knowledge, now have empirical evidence that students who typically choose to use AP credit are not likely to earn lower grades as a result of that choice, regardless of their demographic characteristics.

While it should be reassuring to students who use AP credit that they are not likely to see any negative effect on their grades for doing so, the results of this study have broader implications as well. The model that incorporated ATE weights showed that students with somewhat lower

academic profiles could choose to use AP credit without any expectation that their grades in target courses would be different than if they repeated their credit. Students and academic advisors trying to decide whether a given AP student is prepared to move on to the next subsequent course could expect that a student at this institution from any demographic background with SAT Math scores around 730, a high school GPA close to 3.7, and an average AP score of approximately 4.5 would do just as well whether they repeated AP credit or not. Unless they have unusually low SAT Math scores or high school grades, students who earn 5s on AP exams should feel very confident in using their AP credit, but earning a 4 on an AP exam is not on its own an indication that the student needs to repeat the AP course in order to be successful in subsequent courses. The decision to use or repeat AP credit should depend on each student's priorities, and may vary by course or by semester. However, students who want to move directly into advanced courses, either to graduate more quickly or to free up time to pursue other educational opportunities (Evans, 2019), should feel confident in doing so.

In addition, these results could also inform policy and practice regarding the use of AP credit at the institutional level, because a student's choice to use AP credit can be constrained by institutions. Institutional policy and practice can encourage or discourage the use of AP credit, by requiring 5s to award credit or requiring students to pass placement exams in addition to AP exams (Drew, 2011), by capping the total amount of AP credit a student can earn (Conger et al., 2021), or by strongly advising students to repeat AP courses (Sadler & Sonnert, 2010; Scott et al., 2010). This study included hundreds of students who were successful in target courses after earning 4s on the relevant AP exams, which is consistent with previous findings that students could benefit from earning AP credit across a range of preparation levels (Smith et al., 2017). In the absence of legislative mandates, faculty decide whether to grant credit for AP exams, and for which exam

scores (Johnstone & Del Genio, 2001). With the empirical evidence provided by this study, institutions that require 5s for credit (such as this institution requires for all AP physics exams) may consider granting credit for students with scores of 4 as well.

In addition to setting policies that restrict the awarding or use of AP credit, institutions also influence the decision to use AP credit if they require students to earn high grades in their first year in order to be admitted to competitive majors, as students may choose to repeat AP credit to bolster their GPAs rather than moving into subsequent courses. Allowing, let alone encouraging, students to repeat AP courses raises concerns about pedagogy and equity. First, from a pedagogical standpoint, it can be challenging for instructors to teach students who already have credit for an AP course alongside students with much more limited prior knowledge of the material (National Research Council, 2002). In such situations, instructors have to choose between moving too quickly for students who have yet to encounter the course content, or moving too slowly to keep AP repeaters engaged and interested.

Second, as for equity concerns, students may lose confidence in their abilities or hesitate to ask questions if they perceive that other students in the same course have already mastered the material. And particularly in courses with norm-referenced grading policies, which assign grades based on performance relative to other students enrolled in the course, is it fair for students taking a course for the first time to have to compete with students who earned a 5 on the AP exam that covered the same content? Research on college student success typically defines success as academic achievement, which is nearly always measured with grades and GPA (York et al., 2015). College grades are a major factor influencing future college outcomes (Kuh et al., 2006), and grades in first-year mathematics courses have been shown to be especially important (Herzog, 2005). Given that grades have both far-reaching and immediate consequences for students (such

as admission to competitive majors in the sophomore year as at this institution), and the results of this study show there is no difference in subsequent course grades based on the use of AP credit, it is troubling that institutional policy would encourage students to repeat introductory courses for which they already have college credit. Institutions should examine such policies, and consider how they might provide incentives to students for using AP credit, thereby freeing up space in introductory courses for true beginners, and avoiding the pedagogy and equity concerns that arise when AP students repeat their credit.

Scholarly contributions of the study to AP literature

This study contributed to the literature on the AP program in three primary ways: 1) including students who repeat AP courses in order to understand the outcomes associated with using AP credit, 2) systematically investigating the variation of AP effects across subjects, and 3) strengthening the ability to draw causal inferences about AP effects on college student success. First, while the application of multilevel modeling and propensity score models addressed a call for the use of more sophisticated methodologies in studying the AP program (Warne, 2017), to some extent the most important scholarly contribution of this study was in going beyond questions about the effects of having AP credit and asking about the effects of actually using that credit. Student use of AP credit in place of taking equivalent courses at the college level is the source of most faculty concerns about the AP program described in Chapter 2, either that students would not have learned content required for future coursework (Conley, 2007; Eykamp, 2006), or that they would have learned to focus on procedures rather than on deep engagement (Hansen et al., 2006; Wade et al., 2016). If the controversy around AP at the college level is focused on the use of AP credit, it is surprising that so few studies have compared outcomes of students who used their credit with students who did not. This study explicitly focused on the effects of using AP credit, and

unlike prior studies of repeaters that only included one subject (e.g., De Urquidi, et al., 2015; Hansen et al., 2006), it investigated the effect of using AP credit from four STEM-related subject areas on grades in target courses spanning 17 science and engineering departments. The results of this study showed that, to the extent that a student's mastery of the depth and breadth of introductory college STEM courses can be measured by grades in subsequent courses, there is no evidence that students who use AP credit to meet course pre-requisites are any less prepared to succeed than their classmates who repeated introductory courses in college.

A second contribution of the study was investigating the variation of AP effects across courses. Rather than assuming a common effect for all AP subjects (e.g., Ackerman et al., 2013), this study produced an average effect of using AP credit, but did so while allowing that effect to vary across courses, and investigating both the extent of and possible explanations for that variation. One benefit of this study's approach to investigating variation in AP effects was the inclusion of non-cognate courses, whereas most AP studies limit target courses to those in the cognate department that granted the AP credit (e.g., Godfrey & Beard, 2016). Students in this study used their AP Biology, Calculus, Chemistry, and Physics credit as pre-requisites for courses in those departments, and additionally in animal science, statistics, and ten engineering departments. Limiting the analysis of AP effects to only courses in the same cognate area does not reflect the reality of how AP credit is used by college students, and this study added to the limited prior research that included non-cognate departments (e.g., Patterson & Ewing, 2013).

The third way this study contributed to the AP literature was by accounting as much as possible for the non-random assignment to AP groups using propensity weights. One of the major challenges in any AP research is that it is extremely difficult to randomly assign students to participating in AP courses (e.g., Conger et al., 2021), succeeding on AP exams, or using AP credit

in college, so we are mostly limited to observational studies. There is extensive evidence of confounding factors that affect both AP status and student outcomes, such as prior academic achievement and demographic factors (Sadler, 2010a), and even if researchers attempt to control for these covariates, unobserved student characteristics (e.g., motivation) still contribute to selection bias, casting doubt on estimates of AP effects (Clark et al., 2012). Prior AP studies used some form of propensity score matching to address this selection bias (e.g., Warne et al., 2019); a downside to using matching, however, is the potential loss of significant numbers of unmatched cases, as seen in Patterson and Ewing's (2013) study in which the percentage of treatment students excluded varied from 35% to 79% depending on the subject. Inverse probability weighting, the alternate approach to propensity score matching used in this study, minimizes the loss of cases from both treatment and control groups. Additionally, the method used to create propensity weights in this study allowed for complex and nonlinear relationships among the variables predicting selection into treatment without overfitting the data (McCaffrey et al., 2013), which is an advantage over logistic regression propensity models used in one prior AP study that employed propensity weights (Sadler & Sonnert, 2010). The goal in using propensity models is to define all the variables that systematically determine assignment to treatment, such that any remaining differences between treatment and control groups are random and therefore ignorable (Morgan & Winship, 2015). While there could still be some meaningful, unobservable differences between students who choose to use AP credit and those who do not, the propensity model used in this study included factors related to student demographic and academic background, factors related to the target course, and factors associated with local policy and practice. After weighting, the treatment and control groups in this study were sufficiently homogeneous, in terms of the pre-treatment covariates, that we can conclude there is no causal effect of using AP credit on

subsequent course grades. This study's use of a sophisticated propensity score methodology will contribute to the limited but growing number of AP and other higher education studies (e.g., Routon & Walker, 2019) that provide examples for how to apply propensity modeling in student success research. In this study, the treatment of interest was choosing to use AP credit, but the same approach could be used if the treatment were any other independent variable that was endogenous to the outcome variable of interest.

Lessons learned: Methodological considerations for future AP research

I had to make various methodological choices in the design of this study, and in doing so learned a good deal that could inform future studies of AP effects. I will share some of the key decisions I made and their implications for the current study. I hope sharing the process of methodological decision making will provide some insights into the design of future AP research. AP studies that offer naïve comparisons of outcomes between AP and non-AP students are relatively simple methodologically, although even then there are choices to make: what is it that makes someone an AP student? Which outcomes do you measure, and how? Studies that attempt to address AP effects more rigorously can get complicated quickly, but most of the decisions I had to make were related to the question of what makes someone an AP student, or in this case, a user of AP credit.

The most complicated issue I encountered was that college courses often have more than one pre-requisite course. In the preceding study comparing outcomes of AP users and AP non-users (Hurt & Maeda, under review), all target courses included had only one course pre-requisite, so it was simple to identify to which AP group students belonged. Half of the courses in this study, however, required two or three AP courses as pre-requisites, so I had to account for this more complex data structure in two ways. First, I had to decide what would constitute the "treatment"

variable. I could limit the treatment to a binary choice (i.e., students use AP credit, or they do not), or I could model the effects of multiple treatment options with a multi-category propensity model (e.g., McCaffrey et al., 2013). Had I taken this second route, I could have defined one treatment option as the use of AP credit for one pre-requisite course, and a second treatment option as the use of AP credit for two or more pre-requisite courses. I did not choose to use multiple treatment options for both theoretical and practical reasons. From a practical standpoint, even without using multiple treatment options, I had already eliminated ten possible target courses due to low numbers of students who did not use AP credit (only one or two students per course). When I investigated the possibility of dividing students into three AP groups rather than two, there were simply not enough students in each group to be able to estimate an effect. Even more importantly, from a theoretical perspective, only half the courses required more than one pre-requisite; students in courses with only one pre-requisite would have no chance of selecting into one of the treatment options, which is a violation of one of the assumptions of the potential outcomes framework on which propensity models are based (Arpino & Mealli, 2011). I think it would be interesting for future researchers to consider the use of multiple treatment options when studying the use of AP credit, but they would need to limit their target courses to only those with multiple pre-requisites, and ensure adequate numbers of students in each treatment group.

The issue of multiple course pre-requisites prompted a second decision as well. Which patterns of AP credit use (and non-use) should count as the treatment? When I considered all the possible combinations of how many pre-requisites a course required, how many pre-requisite courses students had AP credit for, and how many of their AP credits they decided to use, there were sixteen separate categories. Combining these categories into two groups of AP Users and AP Non-Users required two choices. First, I had to decide how to categorize students who had AP

credit for some, but not all, of the course pre-requisites. If they had AP credit for one of two course pre-requisites and used it, but had to take the other pre-requisite in college, should they be combined with students who used AP credit for both course pre-requisites? Second, I had to decide how to classify students who had AP credit for multiple course pre-requisites but did not use it all. They could be included with AP Users (treatment students) because they did use AP credit for at least one course pre-requisite, or they could be included with AP Non-Users because they repeated at least one AP course. Because there were more concerns about possible negative effects of using AP credit on target course grades than concerns about not using AP credit, I decided to make the control group as "pure" as possible in that control students did not use any AP credit. This meant a student could be in the AP User category despite repeating (or taking for the first time) two out of three pre-requisite courses.

One consequence of this decision was that in courses with multiple pre-requisites, students would be considered an AP User by using AP credit for any single course; in order to be considered an AP Non-User, students would have to repeat all pre-requisites for which they had AP credit. The inverse is that in courses with only one pre-requisite, students only had to repeat one course to be considered an AP Non-User. This was the case for all three Calculus 2 courses, and it is possible that the decision about how to classify treatment and control students might have contributed to the lower propensity to use AP credit observed in all three Calculus 2 courses. Those designing future studies on the use of AP credit could make different decisions about how to categorize students who use AP credit for some but not all pre-requisites, or they could create multiple treatment options, so the "partial users" form a separate treatment group. If they do so, they will need to decide if "partial use" is a matter of the extent to which students used the AP credit they had earned, or a matter of whether they used AP credit for one pre-requisite or more

than one pre-requisite. In either case, future AP researchers should be aware that, unless the sample of courses is restricted such that all target courses have the same number of pre-requisites, students will have to be combined in one way or another when deciding on what patterns of AP credit use constitute the treatment.

Another independent variable that was challenging to define was students' AP exam scores. It was not obvious how to capture a single value for this variable given that students took multiple exams that could count as target course pre-requisites (e.g., AP Biology and AP Chemistry). If a student had credit for two or more relevant AP courses, should I use the average score? What if the student only used credit for one of the two courses? If a student earned a 4 on one exam and a 5 on another, and earned course credit based on both exams, but only used the credit for one course, should I only use the exam score associated with the course the student used? If I defined the AP exam score variable in that way for students who used at least some AP credit (i.e., the treatment group), I would need to create a separate variable for exam scores associated with credit students did not use, because students in the AP Non-User control group by definition did not use any of their AP credit.

Alternately, I could use the average exam score of all the relevant subjects a student earned credit for, whether the student used the credit or not. The downside to this option is that exam score was a factor in a student's decision to use AP credit. So if the example student I described above used the credit from the exam on which they scored a 5, and did not use the credit from the exam on which they scored a 4, they were included in the study as an AP User with an average AP exam score of 4.5. Neither option is perfect, but ultimately I decided that my priority was to have one common way to operationalize the variable rather than defining AP exam score differently for treatment and control students. Once again, future researchers could make other choices, but will

need to balance the desire to model precise distinctions among student choices with the desire for a parsimonious model.

The last methodological consideration I would like to highlight was the question of whether I was interested in the effect of using AP credit from all four subjects combined, or if I wanted to investigate the effect of using AP credit in each subject separately. Pursuing the latter option would be a way to compare how well each AP course prepared students for subsequent college courses, e.g., to see if the effect of using AP Chemistry credit differed from the effect of using AP Biology credit. Each of the AP courses I included serves as a pre-requisite for multiple courses in the same and related disciplines. For example, AP Chemistry can be a pre-requisite for other chemistry classes, but also for courses in animal science, biology, physics, and multiple engineering disciplines. Therefore, the effect of using AP credit for just one subject could still vary across courses. I was able to investigate this possibility with the AP Calculus case study, and found that the effect of using AP Calculus credit did vary significantly across courses. I was not able to construct similar subject-specific models for AP Biology, AP Chemistry, or AP Physics, however, because there were not enough courses requiring any one of those three subjects to estimate the variation in effect across courses with multilevel modeling. I considered using a single-level regression model with dummy variables for each of the subjects, but as there were no courses in the sample that did not require any of those subjects as pre-requisites, there would have been no reference group. Future scholars with access to larger samples of courses requiring biology, chemistry, or physics as course pre-requisites may choose to investigate the effect of using AP credit in any of those subject areas. In this study, the results of the AP Calculus case study were nearly identical to the results of the primary study that combined all four subjects, suggesting that

the effect of using AP credit from any one subject is likely to be similar to the effect of using any AP STEM credit.

Clearly, there were numerous methodological choices I had to make in designing this study, balancing competing interests. Future studies will need to consider these and other complex variables associated with the use of AP credit; studies that include more than one institution will need to contemplate even greater complexity. The results of this study showed that when we talk about the use of AP credit, and possible AP effects on academic performance, we cannot ignore institutional context. Student academic college and specific course enrollment were among the most important factors predicting a student's choice to use their AP credit, and these factors are likely to vary depending on the policies and practices of the institution, or how specific courses are perceived by students and their advisors. I tried to account for differences in college-level policies and practices by including the indicators of a student's home college within the university within the propensity models, and these differences would be especially important to consider across institutions.

Limitations and directions for future research

There are a few limitations to the conclusions that can be drawn from the results of this study. Future research may be able to address these limitations, or investigate new questions related to the use of AP credit that were not within the bounds of the current study. First, this study included students and courses from only one institution. The nature of the institution (large, public, selective), the students in the sample (predominantly STEM majors with strong academic backgrounds), and institutional policies related to AP (e.g., requiring a 5 for credit in physics) should all be kept in mind when interpreting the results. This may limit the generalizability of the findings (Sadler & Tai, 2007), but given that the courses for which students earn credit, and the

policies for granting AP credit, will vary across institutions (Ackerman et al., 2013), it may be that there is no common effect of using AP credit that could be meaningfully interpreted without institutional context. Any multi-institutional study would likely have to simplify distinctions even more in order to be able to operationalize variables consistently across universities. Perhaps the ideal outcome is that future studies are conducted at other individual institutions in order to explore whether the nuances of local policies lead to different or similar outcomes as found in this study, and other future studies are conducted across multiple institutions to identify possible common effects of using AP credit.

Another potential limitation of the conclusions that can be drawn from the results of this study is the extent to which results apply to students who do not currently use AP credit (or who are not very similar to those students who do use AP credit). My goal in using ATE weights was to estimate an average treatment effect that would apply to all students with AP credit. However, some AP Non-Users with very low propensity scores are so dissimilar to AP Users in terms of SAT Math scores and average AP exam scores that they were excluded from the estimation of the average treatment effect. I am confident in concluding that more students could use AP credit than currently do, without concern that their grades in subsequent courses would be different from what they would be if the students repeated AP credit first, but I cannot say that all students could expect to see no effect of using AP credit. If future studies employed similar approaches to estimating propensity to use AP credit with other samples, that would provide additional information about the profile of students who fall within the range of propensity scores in which current AP Non-Users overlap with current AP Users.

Additional avenues for future research include considering other possible outcomes related to the choice to use AP credit beyond grades in target courses. Longer term outcomes such as

grades in advanced courses, STEM major persistence, and time to degree would provide an indication if there are significant effects of using AP credit that take longer to emerge than the next subsequent course. It would also be interesting to consider how far along a student was in their college career when they enrolled in the target course, and therefore made the decision to use AP credit or not. It could be that students are more or less likely to use AP credit depending on whether they are new beginners or upper-division students; newer students might be more cautious and want to repeat courses to start college with higher grades, but students could also be less likely to use AP credit as more time passes between when they took the AP exam and enrolling in the relevant target course. Finally, given that DFW rate was the only course-level variable that I identified as a significant predictor of target course grades, and it did not explain any of the variation in AP effects across courses, future research could explore other potential course-level predictors.

Conclusion

Most studies of the AP program at the college level attempt to determine if students who participated in the AP program, or performed well on AP exams, have better college outcomes than non-AP students. The primary question about AP at the college level, however, is not about whether students should participate in AP, but about whether students who have AP credit should use it. This study showed that students who use their AP credit to meet course pre-requisites earn the same grades in subsequent STEM courses as they would have earned if they had repeated those pre-requisite courses in college. In short, there was no effect of using AP credit on subsequent course grades. These results support College Board assertions that on average, students who earn 4s and 5s on AP exams are prepared to succeed in subsequent college courses. The study also found that along with indicators of prior academic achievement (AP exam scores, SAT Math

scores, and high school GPA), student propensity to use credit was also strongly influenced by local context (their home College or School) and which course they were taking. Except for International student status, however, no demographic variables (including gender, underrepresented minority status, first generation status, or Pell grant recipient status) were predictors of the use of AP credit, once other variables were included. Based on the results of this study, students who use their AP credit should feel confident that they are not likely to earn lower grades as a result of that decision, and more students whose academic backgrounds are similar to those who currently use AP credit could decide to do so without expecting any negative effects on their subsequent course grades.

APPENDIX: TABLES AND FIGURES

Table A1: Primary Analysis: ATT Model Summary

	Treatment Students (AP Users)	Control Students (AP Non-Users)
N in full sample	7,019	3,133
N included in propensity model (estimate)	7,019	1,445.35
Percent included in propensity model	100%	46.1%

Table A2: Primary Analysis: ATE Model Summary

	Treatment Students (AP Users)	Control Students (AP Non-Users)
N in full sample	7,019	3,133
N included in propensity model (estimate)	6375.466	2059.546
Percent included in propensity model	90.8%	65.7%

Table A3: Primary Analysis: ATT Balance Table Before and After Weighting (all variables)

	Before Weighting						After Weighting					
	Treatment Group (AP Users)		Control Group (AP Non-Users)		Std ES	*p*	Treatment Group (AP Users)		Control Group (AP Non-Users)		Std ES	*p*
Variable	*M*	*SD*	*M*	*SD*			*M*	*SD*	*M*	*SD*		
Gender	0.276	0.447	0.294	0.456	-0.041	0.058	0.276	0.447	0.269	0.444	0.014	0.607
URM	0.054	0.225	0.055	0.228	-0.007	0.736	0.054	0.225	0.053	0.224	0.003	0.913
Int'l	0.095	0.293	0.063	0.243	0.110	<0.001	0.095	0.296	0.101	0.302	-0.021	0.516
FirstGen	0.116	0.321	0.118	0.322	-0.004	0.842	0.116	0.321	0.117	0.321	0.000	0.991
Pell	0.146	0.353	0.145	0.352	0.003	0.882	0.146	0.353	0.141	0.348	0.015	0.620
HSGPA	3.702	0.279	3.660	0.280	0.150	<0.001	3.702	0.279	3.701	0.276	0.004	0.894
SATM	739.66	48.46	718.12	50.95	0.445	<0.001	739.66	48.46	737.38	48.53	0.047	0.090
FYES	0.358	0.479	0.387	0.487	-0.062	0.004	0.358	0.479	0.361	0.480	-0.007	0.796
EngrS	0.295	0.456	0.264	0.441	0.067	0.001	0.295	0.456	0.310	0.463	-0.034	0.223
SciS	0.194	0.396	0.170	0.376	0.062	0.003	0.194	0.396	0.178	0.383	0.042	0.211
AvgScore	4.660	0.468	4.454	0.514	0.439	<0.001	4.660	0.468	4.649	0.478	0.023	0.384
C1	0.014	0.117	0.006	0.076	0.069	<0.001	0.014	0.117	0.010	0.100	0.032	0.205
C2	0.001	0.032	0.002	0.044	-0.029	0.290	0.001	0.032	0.002	0.044	-0.031	0.316
C4	0.003	0.051	0.009	0.094	-0.126	<0.001	0.003	0.051	0.003	0.051	0.007	0.995
C7	0.025	0.155	0.030	0.172	-0.036	0.122	0.025	0.155	0.029	0.167	-0.025	0.287
C8	0.015	0.123	0.003	0.056	0.098	<0.001	0.015	0.123	0.012	0.111	0.023	0.524
C9	0.010	0.101	0.012	0.111	-0.022	0.345	0.010	0.101	0.010	0.101	0.000	0.994
C10	0.008	0.087	0.002	0.047	0.061	<0.001	0.008	0.087	0.005	0.072	0.027	0.341
C12	0.002	0.041	0.002	0.044	-0.005	0.824	0.002	0.041	0.002	0.050	-0.019	0.537
C13	0.005	0.068	0.003	0.056	0.022	0.244	0.005	0.068	0.004	0.064	0.009	0.725

Table A.3 continued

C14	0.017	0.129	0.011	0.105	0.044	0.020	0.017	0.129	0.016	0.127	0.003	0.916
C15	0.005	0.072	0.001	0.025	0.064	<0.001	0.005	0.072	0.002	0.039	0.052	0.007
C16	0.013	0.113	0.006	0.076	0.063	<0.001	0.013	0.113	0.009	0.095	0.033	0.227
C17	0.005	0.067	0.004	0.067	0.001	0.950	0.005	0.067	0.007	0.084	-0.038	0.307
C18	0.005	0.072	0.013	0.115	-0.112	<0.001	0.005	0.072	0.005	0.074	-0.002	0.901
C19	0.005	0.073	0.003	0.050	0.039	0.023	0.005	0.073	0.006	0.077	-0.007	0.840
C20	0.034	0.182	0.021	0.144	0.072	<0.001	0.034	0.182	0.035	0.183	-0.002	0.949
C21	0.002	0.041	0.003	0.054	-0.028	0.280	0.002	0.041	0.002	0.045	-0.009	0.694
C26	0.004	0.064	0.004	0.059	0.010	0.634	0.004	0.064	0.005	0.068	-0.010	0.709
C27	0.016	0.124	0.015	0.120	0.007	0.745	0.016	0.124	0.017	0.129	-0.011	0.679
C28	0.058	0.233	0.106	0.307	-0.206	<0.001	0.058	0.233	0.066	0.248	-0.034	0.107
C29	0.166	0.372	0.203	0.402	-0.099	<0.001	0.166	0.372	0.184	0.388	-0.049	0.060
C30	0.082	0.274	0.276	0.447	-0.710	<0.001	0.082	0.274	0.090	0.287	-0.031	0.093
C31	0.148	0.356	0.078	0.268	0.199	<0.001	0.148	0.356	0.167	0.373	-0.051	0.138
C32	0.070	0.254	0.027	0.163	0.165	<0.001	0.070	0.254	0.056	0.230	0.053	0.206
C33	0.047	0.212	0.023	0.151	0.113	<0.001	0.047	0.212	0.044	0.205	0.016	0.591
C34	0.026	0.158	0.016	0.125	0.061	0.001	0.026	0.158	0.024	0.154	0.008	0.755
C35	0.079	0.270	0.069	0.253	0.037	0.072	0.079	0.270	0.077	0.267	0.006	0.808
C37	0.004	0.063	0.002	0.047	0.028	0.121	0.004	0.063	0.004	0.062	0.002	0.939
C39	0.005	0.067	0.001	0.036	0.049	0.001	0.005	0.067	0.004	0.066	0.003	0.941
C40	0.036	0.186	0.020	0.141	0.083	<0.001	0.063	0.186	0.036	0.187	-0.004	0.902
C41	0.029	0.169	0.016	0.127	0.077	<0.001	0.029	0.169	0.030	0.172	-0.007	0.832
C42	0.002	0.040	0.001	0.036	0.007	0.714	0.002	0.040	0.003	0.051	-0.026	0.479
C43	0.004	0.061	0.001	0.036	0.040	0.012	0.004	0.061	0.006	0.078	-0.039	0.458

Table A4: Primary Analysis: ATE Balance Table Before and After Weighting (all variables)

	Before Weighting						After Weighting					
	Treatment Group (AP Users)		Control Group (AP Non-Users)		Std ES	*p*	Treatment Group (AP Users)		Control Group (AP Non-Users)		Std ES	*p*
Variable	*M*	*SD*	*M*	*SD*			*M*	*SD*	*M*	*SD*		
Gender	0.276	0.447	0.294	0.456	-0.041	0.058	0.282	0.450	0.278	0.448	0.010	0.685
URM	0.054	0.225	0.055	0.228	-0.007	0.736	0.055	0.229	0.054	0.225	0.007	0.758
Int'l	0.095	0.293	0.063	0.243	0.116	<0.001	0.087	0.281	0.089	0.284	-0.007	0.805
FirstGen	0.116	0.321	0.118	0.322	-0.004	0.842	0.119	0.324	0.117	0.321	0.008	0.755
Pell	0.146	0.353	0.145	0.352	0.003	0.882	0.147	0.354	0.142	0.349	0.013	0.624
HSGPA	3.702	0.279	3.660	0.280	0.149	<0.001	3.690	0.281	3.687	0.278	0.009	0.703
SATM	739.66	48. 46	718.12	50.95	0.429	<0.001	733.46	50.11	730.99	50.17	0.049	0.051
FYES	0.358	0.479	0.387	0.487	-0.062	0.004	0.366	0.482	0.370	0.483	-0.008	0.751
EngrS	0.295	0.456	0.264	0.441	0.068	0.001	0.284	0.451	0.295	0.456	-0.025	0.316
SciS	0.194	0.396	0.170	0.376	0.063	0.003	0.187	0.390	0.175	0.380	0.031	0.262
Avg Score	4.660	0.468	4.454	0.514	0.418	<0.001	4.601	0.488	4.584	0.498	0.033	0.178
C1	0.014	0.117	0.006	0.076	0.076	<0.001	0.012	0.108	0.009	0.093	0.029	0.211
C2	0.001	0.032	0.002	0.044	-0.026	0.290	0.001	0.035	0.002	0.044	-0.020	0.459
C4	0.003	0.051	0.009	0.094	-0.095	<0.001	0.004	0.059	0.005	0.068	-0.017	0.370
C7	0.025	0.155	0.030	0.172	-0.034	0.122	0.026	0.160	0.029	0.168	-0.019	0.400
C8	0.015	0.123	0.003	0.056	0.113	<0.001	0.012	0.109	0.009	0.096	0.024	0.433
C9	0.010	0.101	0.012	0.111	-0.021	0.345	0.010	0.101	0.011	0.104	-0.007	0.749
C10	0.008	0.087	0.002	0.047	0.069	<0.001	0.006	0.078	0.004	0.065	0.026	0.311
C12	0.002	0.041	0.002	0.044	-0.005	0.824	0.002	0.042	0.002	0.048	-0.013	0.623
C13	0.005	0.068	0.003	0.056	0.023	0.244	0.004	0.066	0.004	0.062	0.008	0.729
C14	0.017	0.129	0.011	0.105	0.046	0.020	0.016	0.124	0.015	0.120	0.007	0.779

Table A.4 continued

C15	0.005	0.072	0.001	0.025	0.075	<0.001	0.004	0.064	0.001	0.035	0.046	0.010
C16	0.013	0.113	0.006	0.076	0.069	<0.001	0.011	0.103	0.008	0.089	0.028	0.256
C17	0.005	0.067	0.004	0.067	0.001	0.950	0.005	0.067	0.006	0.079	-0.025	0.416
C18	0.005	0.072	0.013	0.115	-0.093	<0.001	0.007	0.082	0.008	0.090	-0.014	0.475
C19	0.005	0.073	0.003	0.050	0.043	0.023	0.005	0.068	0.005	0.069	-0.003	0.920
C20	0.034	0.182	0.021	0.144	0.077	<0.001	0.031	0.172	0.030	0.171	0.003	0.907
C21	0.002	0.041	0.003	0.054	-0.026	0.280	0.002	0.043	0.002	0.048	-0.010	0.647
C26	0.004	0.064	0.004	0.059	0.010	0.634	0.004	0.063	0.004	0.066	-0.007	0.788
C27	0.016	0.124	0.015	0.120	0.007	0.745	0.015	0.121	0.016	0.126	-0.011	0.653
C28	0.058	0.233	0.106	0.307	-0.185	<0.001	0.072	0.258	0.079	0.270	-0.028	0.199
C29	0.166	0.372	0.203	0.402	-0.096	<0.001	0.177	0.382	0.191	0.393	-0.035	0.139
C30	0.082	0.274	0.276	0.447	-0.558	<0.001	0.138	0.345	0.152	0.359	-0.039	0.076
C31	0.148	0.356	0.078	0.268	0.213	<0.001	0.126	0.332	0.137	0.344	-0.032	0.281
C32	0.070	0.254	0.027	0.163	0.182	<0.001	0.057	0.232	0.047	0.211	0.046	0.178
C33	0.047	0.212	0.023	0.151	0.123	<0.001	0.041	0.198	0.037	0.189	0.020	0.453
C34	0.026	0.158	0.016	0.125	0.065	0.001	0.023	0.149	0.022	0.145	0.009	0.717
C35	0.079	0.270	0.069	0.253	0.038	0.072	0.076	0.265	0.075	0.263	0.006	0.789
C37	0.004	0.063	0.002	0.047	0.030	0.121	0.004	0.060	0.003	0.058	0.005	0.840
C39	0.005	0.067	0.001	0.036	0.055	0.001	0.004	0.061	0.003	0.058	0.008	0.815
C40	0.036	0.186	0.020	0.141	0.088	<0.001	0.031	0.174	0.031	0.174	<0.001	0.990
C41	0.029	0.169	0.016	0.127	0.083	<0.001	0.025	0.157	0.026	0.158	-0.004	0.889
C42	0.002	0.040	0.001	0.036	0.008	0.714	0.001	0.037	0.002	0.046	-0.020	0.515
C43	0.004	0.061	0.001	0.036	0.045	0.012	0.003	0.054	0.004	0.067	-0.028	0.511

Table A5: Case Study: ATT Model Summary

	Treatment Students (AP Users)	Control Students (AP Non-Users)
N in full sample	6,283	3,128
N included in propensity model (estimate)	6,283	1450.736
Percent included in propensity model	100%	46.4%

Table A6: Case Study: ATE Model Summary

	Treatment Students (AP Users)	Control Students (AP Non-Users)
N in full sample	6,283	3,128
N included in propensity model (estimate)	5707.195	2107.265
Percent included in propensity model	90.8%	67.4%

Table A7: Case Study: ATT Balance Table Before and After Weighting (all variables)

	Before Weighting						After Weighting					
	Treatment Group (AP Users)		Control Group (AP Non-Users)		Std ES	*p*	Treatment Group (AP Users)		Control Group (AP Non-Users)		Std ES	*p*
Variable	*M*	*SD*	*M*	*SD*			*M*	*SD*	*M*	*SD*		
Gender	0.264	0.441	0.269	0.444	-0.012	0.585	0.264	0.441	0.260	0.438	0.010	0.733
URM	0.053	0.223	0.057	0.232	-0.019	0.399	0.053	0.223	0.056	0.230	-0.014	0.621
Int'l	0.098	0.298	0.066	0.249	0.108	<0.001	0.098	0.298	0.108	0.310	-0.031	0.395
FirstGen	0.115	0.320	0.116	0.320	-0.001	0.961	0.115	0.320	0.126	0.332	-0.033	0.287
Pell	0.137	0.344	0.140	0.347	-0.007	0.756	0.137	0.344	0.141	0.348	-0.009	0.757
HSGPA	3.701	0.279	3.661	0.278	0.145	<0.001	3.701	0.279	3.702	0.277	-0.002	0.958
SATM	742.10	47.04	720.63	50.07	0.456	<0.001	742.10	47.04	740.04	46.96	0.044	0.117
FYES	0.394	0.489	0.410	0.492	-0.033	0.138	0.394	0.489	0.399	0.490	-0.009	0.756
EngrS	0.278	0.448	0.266	0.442	0.027	0.214	0.278	0.448	0.294	0.456	-0.035	0.197
SciS	0.198	0.399	0.174	0.379	0.060	0.005	0.198	0.399	0.182	0.386	0.041	0.199
APMScore	4.715	0.451	4.511	0.500	0.453	<0.001	4.715	0.451	4.704	0.457	0.026	0.311
C1	0.012	0.108	0.007	0.085	0.041	0.031	0.012	0.108	0.012	0.107	0.002	0.944
C9	0.011	0.103	0.008	0.089	0.027	0.169	0.011	0.103	0.012	0.108	-0.009	0.759
C13	0.004	0.059	0.004	0.067	-0.016	0.489	0.004	0.059	0.003	0.055	0.008	0.681
C14	0.021	0.142	0.020	0.141	0.004	0.859	0.021	0.142	0.022	0.148	-0.011	0.670
C15	0.009	0.093	0.002	0.040	0.077	<0.001	0.009	0.093	0.005	0.072	0.037	0.245
C20	0.036	0.186	0.027	0.163	0.046	0.021	0.036	0.186	0.038	0.191	-0.010	0.710
C21	0.001	0.033	0.001	0.036	-0.005	0.830	0.001	0.033	0.001	0.033	0.001	0.955
C26	0.005	0.070	0.004	0.064	0.011	0.592	0.005	0.070	0.006	0.080	-0.021	0.499
C27	0.018	0.133	0.014	0.113	0.032	0.113	0.018	0.133	0.017	0.131	0.005	0.849

Table A.7 continued

C28	0.070	0.254	0.119	0.324	-0.195	<0.001	0.070	0.254	0.076	0.266	-0.027	0.218
C29	0.186	0.389	0.207	0.406	-0.055	0.014	0.186	0.389	0.204	0.403	-0.046	0.087
C30	0.100	0.301	0.287	0.452	-0.621	<0.001	0.100	0.301	0.109	0.312	-0.029	0.144
C31	0.168	0.374	0.091	0.287	0.206	<0.001	0.168	0.374	0.181	0.385	-0.035	0.292
C32	0.073	0.260	0.031	0.173	0.161	<0.001	0.073	0.260	0.054	0.226	0.072	0.098
C33	0.034	0.181	0.027	0.162	0.039	0.056	0.034	0.181	0.034	0.180	0.002	0.952
C34	0.025	0.156	0.020	0.139	0.033	0.104	0.025	0.156	0.026	0.158	-0.004	0.875
C35	0.085	0.279	0.067	0.251	0.064	0.002	0.085	0.279	0.083	0.275	0.009	0.723
C37	0.004	0.059	0.003	0.056	0.005	0.808	0.004	0.059	0.004	0.062	-0.006	0.822
C39	0.007	0.086	0.001	0.036	0.0072	<0.001	0.007	0.086	0.006	0.075	0.022	0.562
C40	0.031	0.172	0.025	0.157	0.031	0.135	0.031	0.172	0.035	0.184	-0.026	0.392
C41	0.032	0.176	0.018	0.134	0.078	<0.001	0.032	0.176	0.033	0.178	-0.006	0.842
C42	0.004	0.065	0.001	0.031	0.051	0.001	0.004	0.065	0.002	0.042	0.038	0.072

Table A8: Case Study: ATE Balance Table Before and After Weighting (all variables)

	Before Weighting						After Weighting					
	Treatment Group (AP Users)		Control Group (AP Non-Users)		Std ES	*p*	Treatment Group (AP Users)		Control Group (AP Non-Users)		Std ES	*p*
Variable	*M*	*SD*	*M*	*SD*			*M*	*SD*	*M*	*SD*		
Gender	0.264	0.441	0.269	0.444	-0.012	0.585	0.267	0.443	0.263	0.440	0.010	0.706
URM	0.053	0.223	0.057	0.232	-0.019	0.399	0.056	0.229	0.056	0.230	-0.003	0.912
Int'l	0.098	0.298	0.066	0.249	0.114	<0.001	0.090	0.286	0.093	0.290	-0.012	0.693
FirstGen	0.115	0.320	0.116	0.320	-0.001	0.961	0.118	0.322	0.122	0.328	-0.014	0.590
Pell	0.137	0.344	0.140	0.347	-0.007	0.756	0.140	0.347	0.140	0.347	-0.002	0.938
HSGPA	3.701	0.279	3.661	0.278	0.145	<0.001	3.689	0.279	3.687	0.278	0.007	0.798
SATM	742.10	47.04	720.63	50.07	0.437	<0.001	735.31	48.94	733.20	48.97	0.043	0.090
FYES	0.394	0.489	0.410	0.492	-0.033	0.138	0.398	0.489	0.403	0.491	-0.011	0.679
EngrS	0.278	0.448	0.266	0.442	0.027	0.214	0.273	0.446	0.284	0.451	-0.025	0.312
SciS	0.198	0.399	0.174	0.379	0.061	0.005	0.190	0.392	0.179	0.384	0.027	0.319
APMScore	4.715	0.451	4.511	0.500	0.427	<0.001	4.652	0.476	4.636	0.481	0.035	0.153
C1	0.012	0.108	0.007	0.085	0.044	0.031	0.011	0.104	0.010	0.100	0.008	0.762
C9	0.011	0.103	0.008	0.089	0.029	0.169	0.010	0.098	0.010	0.102	-0.007	0.802
C13	0.004	0.059	0.004	0.067	-0.016	0.489	0.004	0.059	0.004	0.059	<0.001	0.988
C14	0.021	0.142	0.020	0.141	0.004	0.859	0.020	0.139	0.022	0.145	-0.012	0.620
C15	0.009	0.093	0.002	0.040	0.090	<0.001	0.007	0.081	0.004	0.063	0.033	0.227
C20	0.036	0.186	0.027	0.163	0.048	0.021	0.033	0.179	0.034	0.181	-0.004	0.868
C21	0.001	0.033	0.001	0.036	-0.005	0.830	0.001	0.032	0.001	0.034	-0.003	0.867
C26	0.005	0.070	0.004	0.064	0.011	0.592	0.005	0.069	0.006	0.075	-0.013	0.646
C27	0.018	0.133	0.014	0.116	0.033	0.113	0.017	0.130	0.016	0.126	0.008	0.746
C28	0.070	0.254	0.119	0.324	-0.177	<0.001	0.085	0.278	0.091	0.288	-0.024	0.272

Table A.8 continued

C29	0.186	0.389	0.207	0.406	-0.055	0.014	0.192	0.394	0.205	0.404	-0.033	0.184
C30	0.100	0.301	0.287	0.452	-0.056	<0.001	0.159	0.0366	0.172	0.377	-0.035	0.128
C31	0.168	0.374	0.091	0.287	0.220	<0.001	0.142	0.349	0.149	0.356	-0.021	0.475
C32	0.073	0.260	0.031	0.173	0.177	<0.001	0.059	0.235	0.046	0.209	0.055	0.111
C33	0.034	0.181	0.027	0.162	0.040	0.056	0.031	0.175	0.031	0.174	0.001	0.954
C34	0.025	0.156	0.020	0.139	0.034	0.104	0.023	0.151	0.024	0.152	-0.002	0.922
C35	0.085	0.279	0.067	0.251	0.066	0.002	0.080	0.271	0.077	0.267	0.010	0.688
C37	0.004	0.059	0.003	0.056	0.005	0.808	0.003	0.057	0.004	0.060	-0.006	0.788
C39	0.007	0.086	0.001	0.036	0.084	<0.001	0.006	0.075	0.004	0.064	0.022	0.495
C40	0.031	0.172	0.025	0.157	0.032	0.135	0.028	0.166	0.032	0.175	-0.020	0.460
C41	0.032	0.176	0.018	0.134	0.083	<0.001	0.028	0.166	0.028	0.164	0.004	0.899
C42	0.004	0.065	0.001	0.031	0.059	0.001	0.003	0.058	0.002	0.039	0.033	0.098

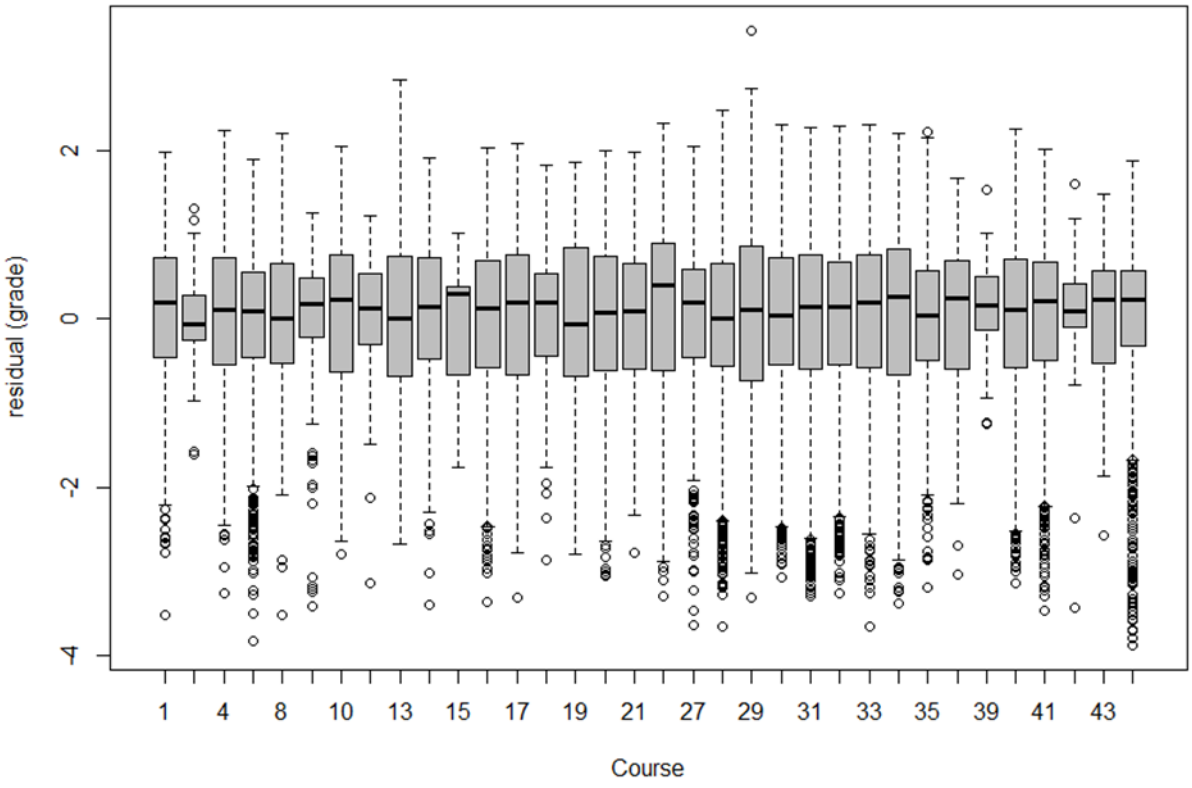

Figure A1: Preliminary Analysis: Box plots of level one residuals by course

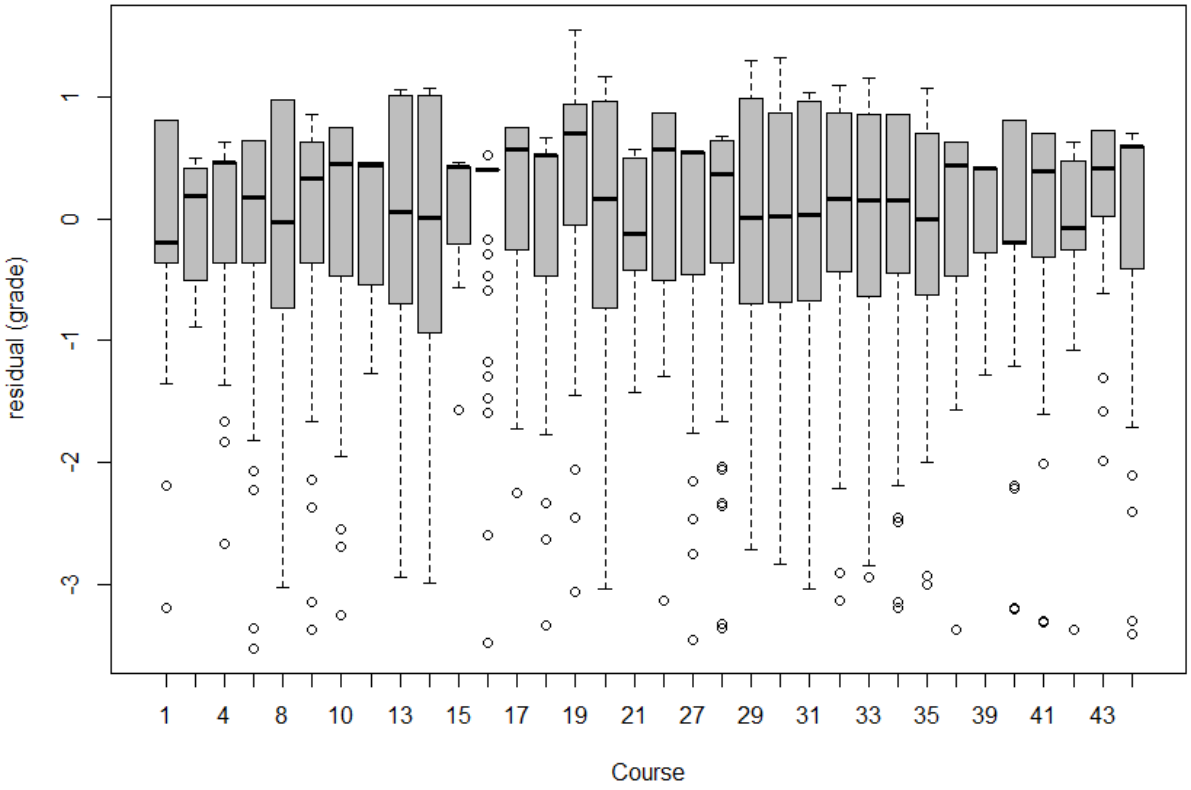

Figure A2: Primary Analysis: Box plots of level one residuals by course

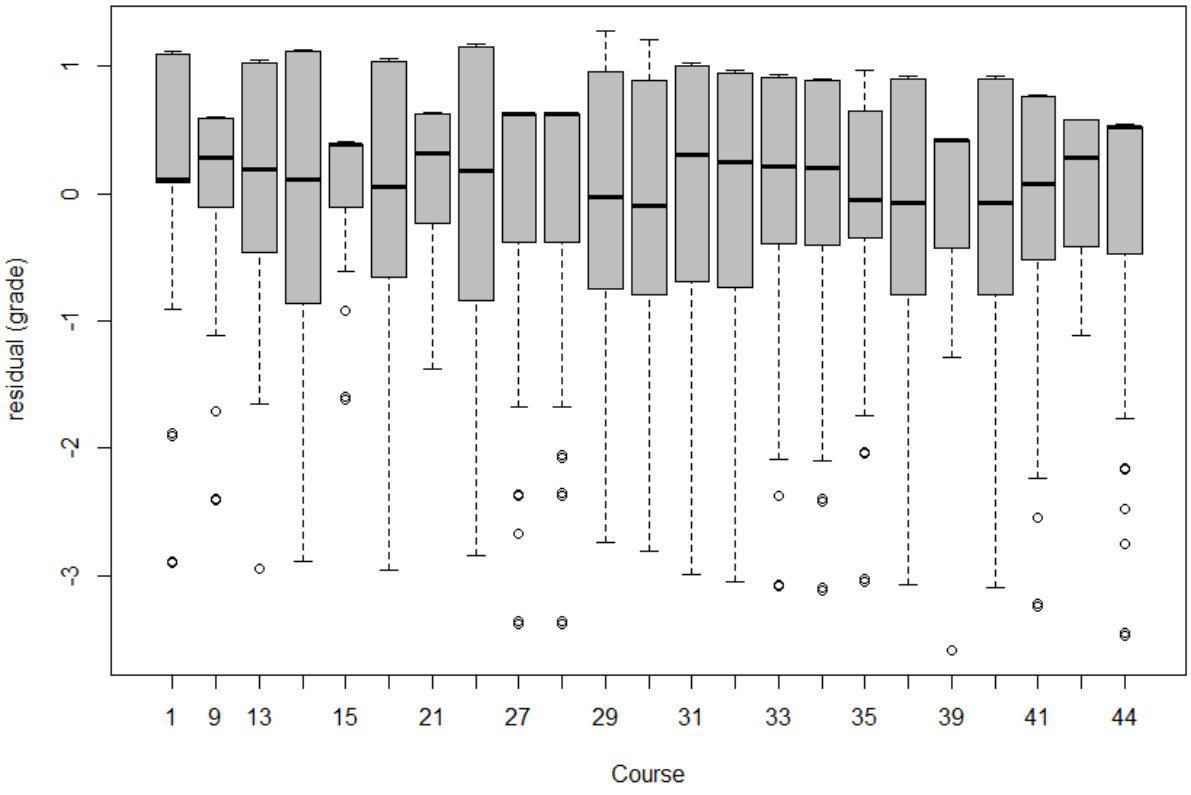

Figure A3: Case Study: Box plots of level one residuals by course

REFERENCES

Ackerman, P. L., Kanfer, R., & Calderwood, C. (2013). High school advanced placement and student performance in college: STEM majors, non-STEM majors, and gender differences. *Teachers College Record, 115*(10), 1-43.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis, 55*(4), 1770-1780.

Bates D, Mächler M, Bolker B, Walker, S. (2015). Fitting linear mixed-effects models using lme4." *Journal of Statistical Software*, 67(1), 1–48. DOI: 10.18637/jss.v067.i01

Bergeson, J. B. (1967). The academic performance of college students granted advanced standing as a result of participation in the Advanced Placement program. *The Journal of Educational Research, 61*(4), 151-152. https://doi.org/10.1080/00220671.1967.10883623

Breland, H. M., & Oltman, P. K. (2001). An analysis of Advanced Placement (AP®) examinations in economics and comparative government and politics. *ETS Research Report Series, 2001*(2), i-31.

Bressoud, D. M., Carlson, M. P., Mesa, V., & Rasmussen, C. (2013). The calculus student: insights from the Mathematical Association of America national study. *International Journal of Mathematical Education in Science and Technology*, *44*(5), 685-698.

Burkholder, E. W., & Wieman, C. E. (2019). What do AP physics courses teach and the AP physics exam measure?. *Physical Review Physics Education Research*, *15*(2), 020117.

Burnham, P. S., & Hewitt, B. A. (1971). Advance Placement scores: Their predictive validity. *Educational and Psychological Measurement, 31*(4), 939-945.

Burns, K., Ellegood, W. A., Bernard Bracy, J. M., Duncan, M., & Sweeney, D. C. (2019). Early college credit programs positively impact student success. *Journal of Advanced Academics, 30*(1), 27-49.

Casserly, P. L. (1986). Advanced placement revisited. *ETS Research Report Series*, *1986*(2), i-14.

Chajewski, M., Mattern, K. D., & Shaw, E. J. (2011) Examining the role of Advanced Placement exam participation in 4-year college enrollment. *Educational Measurement: Issues & Practice*, *30*(4), 16-27.

Clark, C., Scafidi, B., & Swinton, J. R. (2012). Does AP Economics improve student achievement?. *The American Economist*, *57*(1), 1-20.

College Board. (n.d.) *AP program participation and performance data 2019*. Retrieved from https://research.collegeboard.org/programs/ap/data/archived/ap-2019.

College Board. (n.d.) *AP program results: Class of 2019*. Retrieved from https://reports.collegeboard.org/ap-program-results/class-2019-data.

College Board. (n.d.) *Benefits of AP*. Retrieved from https://apcentral.collegeboard.org/launch-grow-ap-program/ap-a-glance/discover-benefits.

College Board. (n.d.) *Concordance*. Retrieved from https://collegereadiness.collegeboard.org/educators/higher-ed/scoring/concordance.

College Board. (n.d.) *State and systemwide AP credit and placement policies*. Retrieved from https://aphighered.collegeboard.org/setting-credit-placement-policy/state-credit-placement-policy.

Conger, D., Kennedy, A. I., Long, M. C., & McGhee, R. (2021). The effect of Advanced Placement science on students' skills, confidence, and stress. *Journal of Human Resources*, *56*(1), 93-124.

Conger, D., Long, M. C., & Iatarola, P. (2009). Explaining race, poverty, and gender disparities in advanced course-taking. *Journal of Policy Analysis and Management, 28*(4), 555-576.

Conley, D. T. (2007). *Redefining college readiness*. Educational Policy Improvement.

De Urquidi, K., Verdin, D., Hoffmann, S., & Ohland, M. W. (2015). Outcomes of accepting or declining advanced placement calculus credit. In *2015 IEEE Frontiers in Education Conference (FIE)* (pp. 1-6). IEEE.

Dodd, B. G., Fitzpatrick, S. J., De Ayala, R. J., & Jennings, J. A. (2002). An investigation of the validity of AP® grades of 3 and a comparison of AP and non-AP student groups. Research Report No. 2002-9. *College Board*.

Dougherty, C., Mellor, L., & Jian, S. (2006). The relationship between Advanced Placement and college graduation. 2005 AP Study Series, Report 1. *National Center for Educational Accountability*.

Drew, C. (2011, January 10). Rethinking advanced placement. *The New York Times*, p. ED24. Retrieved from https://www.nytimes.com/2011/01/09/education/edlife/09ap-t.html.

Duffet, A., & Farkas, S. (2009). Growing pains in the Advanced Placement program: Do tough trade-offs lie ahead?. *Thomas B. Fordham Institute.*

Duffy, W. R. (2010). Persistence and performance at a four-year university: The relationship with advanced coursework during high school. In P. M. Sadler, G. Sonnert, R. H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program.* (pp. 136-163). Harvard Education Press.

Ellis, J., Fosdick, B. K., & Rasmussen, C. (2016). Women 1.5 times more likely to leave STEM pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit. *PLOS ONE, 11*(7), e0157447. https://doi.org/10.1371/journal.pone.0157447.

Evans, B. J. (2019). How College Students use Advanced Placement Credit. *American Educational Research Journal, 56*(3), pp. 925-954. https://doi-org./10.3102/0002831218807428

Ewing, M. (2006). The AP Program and student outcomes: A summary of research. Research Notes. RN-29. *College Board*.

Ewing, M., Huff, K., & Kaliski, P. (2010). Validating AP exam scores: Current research and new directions. In P. M. Sadler, G. Sonnert, R. H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program* (pp. 85-105). Harvard Education Press.

Eykamp, P. W. (2006). Using data mining to explore which students use Advanced Placement to reduce time to degree. *New Directions for Institutional Research, 131,* 83-99.

Flowers, L. A. (2008). Racial differences in the impact of participating in Advanced Placement programs on educational and labor market outcomes. *Educational Foundations, 22,* 121-132.

Geiser, S., & Santelices, V. (2004). *The role of Advanced Placement and honors courses in college admissions*. Berkeley, CA: Center for Studies in Higher Education.

Godfrey, K. E., & Beard, J. J. (2016). Advanced Placement validity research at four university System of Georgia institutions: Placement validity study results. Statistical Report. *College Board*.

Godfrey, K., Matos-Elefonte, H., Ewing, M., & Patel, P. (2014). College completion: Comparing AP, dual-enrolled, and nonadvanced students. Research Report 2014-3. *College Board.*

Hamaker, E. L., van Hattum, P., Kuiper, R. M., & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 231-255). Psychology Press.

Hansen, K., Reeve, S., Gonzalez, J., Sudweeks, R. R., Hatch, G. L., Esplin, P., & Bradshaw, W. S. (2006). Are Advanced Placement English and first-year college composition equivalent? A comparison of outcomes in the writing of three groups of sophomore college students. *Research in the Teaching of English, 20*(4), pp. 461-501.

Hargrove, L., Godin, D., & Dodd, B. (2008). College outcomes comparisons by AP® and non-AP high school experiences. Research Report No. 2008-3. *College Board*.

Hedges, L.V., & Hedberg, E.C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87. https://doi.org/10.3102/0162373707299706

Herzog, S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in higher education, 46*(8), 883-928.

Hurt, S. F., & Maeda, Y. (2019). *Should students with AP credit repeat coursework in college? A multilevel analysis.* Manuscript submitted for publication.

Huang, F. L. (2018). Multilevel modeling myths. *School Psychology Quarterly*, *33*(3), 492.

Johnstone, D.B., & Del Genio, B. (2001). *College-level learning in high school: Purposes, policies, and practical implications.* Association of American Colleges & Universities.

Judson, E. (2017). Science and mathematics Advanced Placement exams: Growth and achievement over time. *The Journal of Educational Research*, *110*(2), 209-217.

Klopfenstein, K. (2010). Does the Advanced Placement program save taxpayers money? The effect of AP participation on time to college graduation. In P. M. Sadler, G. Sonnert, R. H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program* (pp. 189-218). Harvard Education Press.

Klopfenstein, K., & Thomas, M. K. (2009). The link between advanced placement experience and early college success. *Southern Economic Journal, 75*, 873-891.

Klopfenstein, K., & Thomas, M. K. (2010). Advanced Placement participation: Evaluating the policies of states and colleges. In P.M. Sadler, G. Sonnert, R.H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program.* (pp. 167-188). Harvard Education Press.

Kuh, G. D., Kinzie, J. L., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006). *What matters to student success: A review of the literature.* National Postsecondary Education Cooperative.

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2020). lmerTest: Tests in linear mixed effects models. R package version 3.1-3.

Lacy, T. (2010). Access, rigor, and revenue in the history of the Advanced Placement program. In P. M. Sadler, G. Sonnert, R. H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program* (pp. 17-48). Harvard Education Press.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research, 50*(3), 265-284. https://doi.org/10.1080/00273171.2014.991018

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine, 32*(19), 3373-3387.

Lichten, W. (2000). Whither advanced placement? *Education Policy Analysis Archives, 8*(29). Retrieved October 6, 2018 from https://epaa.asu.edu/ojs/article/viewFile/420/543.

Long, M. C., Conger, D., & Iatarola, P. (2012). Effects of high school course-taking on secondary and postsecondary success. *American Educational Research Journal*, *49*(2), 285-322.

Mattern, K. D., Shaw, E. J., & Xiong, X. (2009). The relationship between AP exam performance and college outcomes. Research Report No. 2009-4. *College Board.*

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, *32*(19), 3388-3414.

McCoy, A., Jurist Levy, A., Frumin, K., Lawrenz, F., Dede, C., Eisenkraft, A., Fischer, C., Fishman, B., & Foster, B. (2020). From the inside out: Teacher responses to the AP curriculum redesign. *Journal of Science Teacher Education*, *31*(2), 208-225. https://doi.org/10.1080/1046560X.2019.1685630.

McKillip, M. E., & Rawls, A. (2013). A closer examination of the academic benefits of AP. *The Journal of Educational Research*, *106*(4), 305-218.

Morgan, R., & Klaric, J. (2007). AP students in college: An analysis of five-year academic careers. Research Report No. 2007-4. *College Board.*

Morgan, S. L., & Winship, C. (2015). Counterfactuals and causal inference. Cambridge University Press.

Murphy, D., & Dodd, B. (2009). A comparison of college performance of matched AP and non-AP student groups. Research Report No. 2009-6. *College Board.*

National Research Council. (2002). *Learning and understanding: Improving advanced study of mathematics and science in US high schools*. National Academies Press.

Patterson, B. F., & Ewing, M. (2013). Validating the use of AP® exam scores for college course placement. Research Report 2013-2. *College Board*.

Patterson, B. F., Packman, S., & Kobrin, J. L. (2011). Advanced Placement® exam-taking and performance: Relationships with first-year subject area college grades. Research Report No. 2011-4. *College Board*.

Raudenbush, S. W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., Burgette, L. & Cefalu, M. (2020). twang: Toolkit for weighting and analysis of nonequivalent groups. R package version 1.6.

Routon, P.W., & Walker, J. K. (2019). College internships, tenure gaps, and student outcomes: a multiple-treatment matching approach. *Education Economics*, *27*(4), 383-400. https://doi.org/10.1080/09645292.2019.1598336.

Sadler, P. (2010a). Advanced Placement in a changing educational landscape. In P. M. Sadler, G. Sonnert, R. H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program* (pp. 3-16). Harvard Education Press.

Sadler, P. (2010b). How are AP courses different?. In P. M. Sadler, G. Sonnert, R. H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program* (pp. 51-62). Harvard Education Press.

Sadler, P. & Sonnert, G. (2010). High school Advanced Placement and success in college coursework in the sciences. In P.M. Sadler, G. Sonnert, R.H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program* (pp. 119-138). Harvard Education Press.

Sadler, P., & Sonnert, G. (2018). The path to college calculus: The impact of high school mathematics coursework. *Journal for Research in Mathematics Education*, *49*(3), 292-329.

Sadler, P. M., & Tai, R. H. (2007). Advanced Placement exam scores as a predictor of performance in introductory college biology, chemistry and physics courses. *Science Educator*, *16*(2), 1-19.

Scott, T.P., Tolson, H., & Lee, Y.H. (2010). Assessment of Advanced Placement participation and university academic success in the first semester: Controlling for selected high school academic abilities. *Journal of College Admission, 208*, 26-30.

Shaw, E.J., & Barbuti, S. (2010). Patterns of persistence in intended college major with a focus on STEM majors. *NACADA Journal*, *30*(2), 19-34.

Shaw, E. J., Marini, J. P., & Mattern, K. D. (2013). Exploring the utility of Advanced Placement participation and performance in college admission decisions. *Educational and Psychological Measurement*, *73*(2), 229-253.

Smith, J., Hurwitz, M., & Avery, C. (2017). Giving college credit where it is due: Advanced Placement exam scores and college outcomes. *Journal of Labor Economics*, *35*(1), 67-147.

Tai, R. H., Liu, C. Q., Almarode, J. T., and Fan, X. (2010). Advanced Placement course enrollment and long-range educational outcomes. In P.M. Sadler, G. Sonnert, R.H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program* (pp. 109-118). Harvard Education Press.

U.S. Department of Education. (2006). "Advanced Placement test fee program." Guide to U.S. Department of Education Programs. Retrieved from http://www.ed.gov/programs/apfee/index.html.

Wade, C., Sonnert, G., Sadler, P., Hazari, Z., & Watson, C. (2016). A comparison of mathematics teachers' and professors' views on secondary preparation for tertiary calculus. *Journal of Mathematics Education at Teachers College*, *7*(1).

Warne, R. T. (2017). Research on the academic benefits of the advanced placement program: Taking stock and looking forward. *SAGE Open*, *7*(1), 1-16. https://doi.org/10.1177/2158244016682996

Warne, R. T., Larsen, R., Anderson, B., & Odasso, A. J. (2015). The impact of participation in the Advanced Placement program on students' college admissions test scores. *The Journal of Educational Research*, *108*(5), 400-416.

Warne, R. T., Sonnert, G., & Sadler, P. M. (2019). The relationship between Advanced Placement mathematics courses and students' STEM career interest. *Educational Researcher*, *48*(2), 101-111.

Wiley, J. F. (2020). multilevelTools: Multilevel and mixed effects model diagnostics and effect sizes. R package version 0.1.1.

Wladis, C., Conway, K., & Hacheym, A. C. (2017). Using course-level factors as predictors of online course outcomes: A multi-level analysis at a US urban community college. *Studies in Higher Education, 42*(1), 184-200. https://doi.org/10.1080/03075079.2015.1045478

Wyatt, J., Jagesic, S., & Godfrey, K. (2018). Postsecondary course performance of AP® exam takers in subsequent coursework. *College Board*.

Wyatt, J. N., Patterson, B. F., & Di Giacomo, F. T., (2015). A comparison of the college outcomes of AP and dual enrollment students. Research Report 2015-3. *College Board.*

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research, and Evaluation*, *20*(1), 5.