THE INFLUENCE OF SCHOOL ACCOUNTABILITY INCENTIVES ON STATE-LEVEL ADVANCED PLACEMENT OUTCOMES

by

PAUL TIMOTHY BEACH

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2020

DISSERTATION APPROVAL

Student: Paul Timothy Beach

Title: The Influence of School Accountability Incentives on State-Level Advanced Placement Outcomes

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy, and Leadership by:

Keith Zvoch	Chair
Gina Biancarosa	Core Member
David D. Liebowitz	Core Member
Joanne Goode	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded December 2020.

DISSERTATION ABSTRACT

Paul Timothy Beach

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

December 2020

Title: The Influence of School Accountability Incentives on State-Level Advanced Placement Outcomes

Research examining the effects of school accountability on student outcomes has focused almost exclusively on numeracy and literacy proficiency. However, states now use a variety of indicators for school accountability purposes, including Advanced Placement (AP) exam data. This study examined the influence of school accountability incentives on average state-level AP exam performance in four individual courses. Outcome data was collected from the College Board's publicly available national and state annual AP reports from 1997 to 2019. Event study, difference-in-differences, and comparative interrupted time series analyses compared outcomes between states that did and did not introduce AP exam accountability indicators in the timeframe analyzed.

Results were interpreted using coherent pattern matching to describe and evaluate the empirical coherence of policy effects across all analyses and outcomes. In general, inconsistent findings emerged with respect to statistical significance with the comparative interrupted time series models producing the vast majority of statistically significant policy effects. However, results were moderately consistent in direction and magnitude across analyses, models, and outcomes. Taken together, differences in findings across outcomes suggests state policymakers may see a range of effects from introducing AP exam accountability indicators. Policymakers may see a small, positive increase in the

trend of average state-level AP exam performance, an immediate decrease in average AP exam performance, an immediate short-term decrease in AP exam performance that is offset by a larger long-term increase, or no effects at all.

Supporting contextual information and exploratory analyses revealed that several factors may contribute to the mixed findings, including but not limited to variation in the strength of AP accountability incentives, different indicator designs, and the presence of other types of state AP policies. Future research on AP accountability incentives should employ a rigorous mixed methods approach to further isolate the effects of AP accountability incentives and to provide needed evidence on how school leaders and teachers behave in response to the introduction of AP exam accountability indicators. Additionally, future research should examine the benefits and costs associated with introducing AP exam accountability indicators to determine how well public resources are being spent in the pursuit of school improvement.

CURRICULUM VITAE

NAME OF AUTHOR: Paul Timothy Beach

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of Wisconsin Oshkosh

DEGREES AWARDED:

Doctor of Philosophy, Quantitative Research Methods in Education, 2020, University of Oregon
Master of Public Administration, 2013, University of Oregon
Bachelor of Science, 2011, Economics, University of Wisconsin Oshkosh
Bachelor of Science, 2011, Political Science, University of Wisconsin Oshkosh

AREAS OF SPECIAL INTEREST:

School accountability policy
College and career readiness
Educational equity
Policy analysis and program evaluation
Advanced research design

PROFESSIONAL EXPERIENCE:

Research Associate, Inflexion, Eugene, Oregon, 2012-present

Graduate Employee, Center for Equity Promotion, University of Oregon, 2015-2018

Graduate Teaching Fellow, Department of Planning, Public Policy, and Management, University of Oregon

Research Assistant, School of Architecture and Allied Arts, University of Oregon

Student Titan Employment Program Research Assistant, Department of Political Science, University of Wisconsin Oshkosh

GRANTS, AWARDS, AND HONORS:

Travel Grant, College of Education, University of Oregon, 2018

Education Policy Academy Class of 2017, American Enterprise Institute, 2017

Amelia and Albert Lee DeFerris Scholarship, College of Education, University of Oregon, 2015

Travel Grant, College of Education, University of Oregon, 2015

Travel Grant, Department of Educational Methodology, Policy, and Leadership, University of Oregon, 2015

Alumni Scholarship, College of Education, University of Oregon, 2015

Outstanding Research Award, University of Wisconsin Oshkosh Economics Department, 2009

Max Sparger Scholar-Athlete Award, Wisconsin Intercollegiate Athletic Conference, 2009

North Central Region Scholar Team, National Soccer Coaches Association of America, 2008; 2009

PUBLICATIONS:

Thier, M., & Beach, P. (in press). Still where, not if, you're poor: International Baccalaureate opportunities to learn international-mindedness and proximity to U.S. cities. *Journal of Advanced Academics*.

Thier, M., & Beach, P., Martinez Jr., C. R., & Hollenbeck, K. (2020). Take care when cutting: Five approaches to disaggregating school data as rural and remote. *Theory & Practice in Rural Education*, *10*, 63–84.

Beach, P., Zvoch, K., & Thier, M. (2019). The influence of school accountability incentives on Advanced Placement access: Evidence from Pennsylvania. *Education Policy Analysis Archives*, *27*(138), 1–29.

Thier, M., & Beach, P. (2019). Stories we don't tell: Research's limited accounting of rural schools. *School Leadership Review*, *14*(2), 1–14.

Beach, P., Thier, M., Fitzgerald, J., & Pitts, C. (2018). Cutting-edge (and dull) paths forward: Accountability and Career and Technical Education under the Every Student Succeeds Act. In G. Lauzon (Ed.), *Educating a working society: Vocationalism, the Smith-Hughes Act, and modern America*. Charlotte, NC: Information Age.

Thier, M., Beach, P., & Fitzgerald, J. (2018). Accountability policies that segregated high school students by track and the innovations that can reunite them. In G. Lauzon (Ed.), *Educating a working society: Vocationalism, the Smith-Hughes Act, and modern America*. Charlotte, NC: Information Age.

Beach, P. (2017) Multiple spheres of influence: Who influenced Australia's education revolution? In Y. Zhao & B. Gearin (Eds), *Imagining the future of global education: Dreams and Nightmares*. New York, NY: Routledge.

Nollenberger, K., Maher, C., Beach, P., & McGee, M. K. (2012). Budget priorities and community perceptions of service quality and importance. *Journal of Public Budgeting, Accounting & Financial Management*, *24*, 255–277.

Siemers, D. J., & Beach, P. (2012). Presidential coequality: The evolution of a concept. *Congress & the Presidency*, *39*, 316–339.

ACKNOWLEDGEMENTS

Thank you first and foremost to my advisor, Dr. Keith Zvoch, for his guidance, support, and encouragement throughout my entire doctoral program, but especially during the dissertation process. I feel privileged to have had an advisor that supported my ideas, provided timely and constructive feedback, and shared my enthusiasm for rigorous quasi-experimental research. I will be forever grateful for the mentorship Dr. Zvoch provided me on educational research and life in general.

I also wish to thank the other members of my committee, including Dr. Gina Biancarosa and Dr. Joanne Goode, for their willingness to serve on my committee and for improving my understanding of the analyses I conducted and the content I studied. Thank you to Dr. David Liebowitz for serving on my committee and for his patient support and guidance throughout the analysis process. I feel honored to have worked with such a talented and committed group of scholars.

I also wish to thank my current and past colleagues at Inflexion (formerly the Educational Policy Improvement Center) and the Center for Equity Promotion. Several of my colleagues, far too many to name here, initially encouraged me to pursue a PhD and then provided extraordinary support throughout my doctoral program. These colleagues have also served as incredible role models and are remarkable researchers.

Finally, I want to thank all of the educators that have helped shape me into the person I am today. A special thanks goes to Dr. David Siemers who a long time ago challenged me to be a better student, taught me the value of research, and encouraged me to pursue graduate studies. Thank you to all of the other educators that helped me see a future for myself I never thought was possible.

To my grandparents,
for teaching me the value of
hard work, curiosity, and respect.

To my parents,
for encouraging my curiosity
and forever being proud of me.

To my brother,
for never letting me give up,
and for always having fun in life.

To my wife Caitlin,
for inspiring me every day
and for always being by my side.

To my son Levi,
for bringing me constant joy,
and for showing me the future is bright.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Since the 1990s, test-based accountability has been the preferred approach among state and federal policymakers for improving public schools and promoting educational equity (Ladd, 1996). To date, the most visible and controversial accountability policy was the federal No Child Left Behind Act of 2001 (NCLB). Initially, NCLB's most notable mandates required states to administer annual standardized tests in mathematics and reading in grades 3 through 8 and once between grades 10 and 12. NCLB required states to use standardized tests scores to reward or sanction schools based on aggregated student performance and the performance of eligible subgroups of students. NCLB's design, often referred to as test-based accountability (Ryan & Shepard, 2008), mirrored the systems created by states during the 1990s (Goertz & Duffy, 2001).

A common conclusion among educational researchers is that the evidence on test-based accountability is "decidedly mixed" (Figlio & Ladd, 2015, p. 205; Jacob, 2005, p. 763; Springer, 2008, p. 556). This comes as no surprise given research on test-based accountability has been "broadly focused, methodologically diverse, and theoretically varied" (Gunzenhauser & Hyde, 2007, p. 489). Studies on school accountability vary widely in which outcome variables they model, what samples they employ, and what time periods they examine (Lee, 2008). Regardless, several literature reviews do suggest test-based accountability has had a modest, positive effect on student achievement, but little to no effect on closing achievement gaps (Figlio & Ladd, 2015; Figlio & Loeb, 2011; Lee, 2008). For example, in a meta-analysis of 14 studies that examined the effect of school accountability on state National Assessment of Educational Progress (NAEP)

exam scores, Lee (2008) found a "modestly positive policy effect on average but no significant effect on narrowing the racial achievement gap" (p. 628-629).

Any positive effects stemming from test-based accountability have been tempered by a host of unintended consequences (Figlio & Loeb, 2011; Jacob, 2017; McEachin, ND). These unintended consequences can be attributed to the narrow set of outcomes targeted in test-based accountability policy and NCLB's ambitious goal of 100% proficiency in numeracy and literacy by 2014. As Koretz (2017) argues, the narrow set of ambitious goals set by test-based accountability policy presented teachers with three options: "fail, cut corners, or cheat—and many chose not to fail" (p. 244).

Common tactics included systematically excluding students from taking standardized tests by reclassifying low-performing students into student subgroups not subjected to NCLB's testing requirements (Booher-Jennings, 2005; Cullen & Reback, 2006; Figlio & Getzler, 2007; Jacob, 2005) and disciplinary practices that disproportionately punished students with low academic achievement (Figlio, 2006). Some schools also performed "educational triage" by focusing instruction on so-called "bubble students," those students near the cutoff for proficiency on standardized tests (e.g., Lauen & Gaddis, 2016), though other researchers have found no evidence of educational triage (Ballou & Springer, 2017). Yet another critique is that NCLB created incentives for schools to narrow the curriculum by focusing more instructional time on tested subjects and thereby less time on non-tested subjects (Au, 2007; Berliner, 2011; Ladd & Zelli, 2002; Koretz & Hamilton, 2003; McMurrer, 2007; Pederson, 2007). A narrower curriculum, coupled with instructional practices that emphasized "teaching to the test," resulted in schools and teachers focusing more on discrete content knowledge

than broad conceptual understanding in the tested subjects (Au, 2007; Jacob, 2005; Jennings & Bearak, 2014; Jennings & Lauen, 2016). Perhaps most problematically, the pressure generated from NCLB led to outright cheating (e.g., changing student answers) in some cases (Severson, 2011).

These unintended consequences all point to a common critique of NCLB—the law focused too narrowly on standardized test scores. Research shows that many educational stakeholders' value critical thinking, civic engagement, social and emotional skills, the arts, and other, more difficult to measure outcomes (e.g., Brighouse et al., 2018; Labaree, 1997; Rothstein et al., 2008). Subsequently, one of the most common policy recommendations for improving the design of NCLB was to expand the outcomes states used to measure school quality (e.g., Darling-Hammond et al., 2014). Calls to hold schools accountable to a more holistic set of outcomes and the increasingly apparent unintended consequences of NCLB coincided with a law more than seven years overdue for reauthorization. As a result, the Obama administration initiated the Elementary and Secondary Education (ESEA) Act flexibility waiver process, referred to as ESEA waivers. ESEA waivers allowed state policymakers to request flexibility from the U.S. Department of Education on certain NCLB mandates. Most notably, states could expand the outcomes used to hold schools accountable.

Polikoff et al. (2014) argue that expanding the outcomes employed in school accountability is a significant improvement from NCLB, though this view is not universally shared (Goldhaber & Özek, 2019). Polikoff et al. refer to NCLB's narrow focus on mathematics and ELA as a construct validity problem. The authors argue school performance ratings "under an accountability policy have construct validity if the

performance measures adequately cover the latent set of desired student outcomes" (p. 46). In their analysis of 42 ESEA waivers (eventually 43 states, the District of Columbia, and Puerto Rico operated under approved ESEA waivers), Polikoff et al. found that many states, but not all, improved the construct validity within its school accountability system by adding more outcomes. Common additions included graduation rates, attendance rates, and indicators of college and career readiness (Polikoff et al., 2014).

A common college and career readiness indicator employed by several states, and the focus of this study, are Advanced Placement (AP) outcomes. Within the timeframe analyzed, state policymakers in 19 states incorporated AP exam indicators into their school accountability systems. AP outcomes are a popular college and career readiness accountability indicator for several reasons. First, the AP program, administered by the College Board, represents the most common way students can earn college credit during high school (Kolluri, 2018). Students can earn college credit at most postsecondary institutions with eligible AP exam scores on one or more of the College Board's 38 courses. Second, the AP program is well known to the general public with more than 16,000 public schools enrolling at least one student who completed an AP exam (College Board, 2019). Third, AP exams are viewed as technically adequate for the purposes of school accountability (e.g., Conley et al., 2014). Perhaps most importantly, AP courses and exams have the potential to provide students with academic (e.g., Warne, 2017), economic (e.g., Kolluri, 2018), and social benefits (e.g., Klopfenstein, 2010). Thus, there are several potential reasons policymakers in 19 states have introduced AP exam indicators into their school accountability system.

The purpose of this study is to examine the effects of introducing AP exam

indicators in school accountability systems. In the sections below I discuss a theoretical rationale for improving the construct validity of state accountability systems by measuring student achievement and evaluating school performance with a broader set of indicators, including AP exam data. Next, I describe how ESEA waivers and the Every Student Succeeds Act (ESSA) of 2015 created space for state policymakers to experiment with different approaches to improving the construct validity of their school accountability systems. I end by situating the current research in the scholarly literature on school accountability and the AP program.

A Theoretical Rationale for School Accountability

The historical account of school accountability is varied and multifaceted, with authors pointing to different eras and justifications for why school accountability is so entrenched in public education and how the core policy design remains relatively unchanged from era to era (e.g., Cuban, 2004; Dorn, 2007; Groen, 2012; Ravitch, 2002; West & Peterson, 2003). West and Peterson (2003) argue that declining SAT scores and mediocre performance on international assessments ultimately led to the publication of *A National at Risk* (West & Peterson, 2003). According to these authors, the influential *A National at Risk* report issued by the U.S. Department of Education sounded the alarm on declining performance of American students on standardized tests. The calls for reform issued in *A National at Risk* are credited with generating political pressure in states to address educational issues. During the next two decades, several states began experimenting with school accountability (Cornett & Gaines, 1997; Goertz & Duffy, 2001). From the very beginning, accountability proved to be a policy approach with bipartisan support from both major U.S. political parties (West & Peterson, 2003).

The bipartisanship generated at the state-level eventually made its way to the U.S. Congress in 2001 with NCLB. NCLB, just like the state reforms that preceded it, arose from a "grand bargain" between Democrats focused on improving educational equity by exposing long-standing achievement gaps and Republicans who championed strict accountability measures to ensure federal funds produced the desired outcomes (McGuinn, 2005). The bargain included more federal dollars targeted to low-performing schools in exchange for increased accountability. The next two waves of accountability reform, the Race to the Top and ESEA waivers, although federally imposed and slightly different in design, maintained these same general principles of equity-focused spending and test-based accountability measures. ESSA, the reauthorization of ESEA, also produced a rare moment of congressional bipartisanship (Hess & McShane, 2018). ESSA maintained the focus on accountability for equity, but loosened the Federal government's grip on states by providing much greater flexibility in the identification of low-performing schools as well as how states approached improving these schools.

While accountability policy has changed names, each of these reforms federally mandated that states use school performance metrics to identify and intervene in low-performing schools. Although political considerations helped lead to the *adoption* of accountability policies across the states and at the federal-level, economic theory that emerged in the 1970s helps explain why accountability systems are *designed* the way they are. A growing body of scholars point to principal-agent theory as one theoretical rationale for the design of school accountability systems (Bae, 2018; Figlio & Ladd, 2015; Figlio & Loeb, 2011; Jacob, 2017; Jacobsen et al., 2014; Ladd & Zelli, 2002; McEachin, ND; Moe, 2003; Polikoff et al., 2014; West, 2009). The following sections

provide a brief overview of principal-agent theory and its application to the design and effects of test-based school accountability policy.

A Brief Overview of Principal-Agent Theory

Principal-agent theory emerged in the economics literature as a way to conceptualize contractual relations between businesses and employees (Holmstrom & Costa, 1986; Jensen & Meckling, 1976; Milgrom & Roberts, 1990). In the basic model, principals (e.g., business owners, managers) have certain goals they want agents (e.g., employees) to pursue. For example, in the business world, the principal's primary goal is almost always to make profit. Understanding that they cannot do all the work themselves and that it would be inefficient to try, principals enter into contractual relations with agents for labor. Creating an effective contract begins with identifying who the principals and agents are, defining what both are capable of doing, and specifying how to appropriately evaluate agent performance (Gailmard, 2012). However, many potential problems complicate this basic contractual setup. For one, principals and agents rarely have interests and goals that perfectly align. Agents also may possess information that principals are not privy to, have difficulty accessing, or simply do not know exists. Thus, principals may have to create new measures for the information they seek.

Economists use the term moral hazard to describe situations where principals and agents have divergent interests, principals cannot directly observe agent performance or control their actions, and agents have an incentive to act counter to the interests of principals in order to maximize their own well-being (Dixit, 2002; Ferris, 1992; Gailmard, 2012). In moral hazard situations, principal-agent theory recommends "the principal observes some information affected by or correlated with the agent's action, and

administers a reward or punishment … based on that information" (Gailmard, 2012, p. 4). The effectiveness of a contract attempting to address the moral hazard problem rests on the costs of enforcement as well as incentive compatibility. In most circumstances, it is either prohibitively costly or impossible to measure all the pertinent actions of agents or the outcomes of their work. Moreover, the incentives contained within contracts, whether they are rewards (e.g., raise, bonus pay) or punishments (e.g., probation, termination), may not effectively motivate agents to pursue principals' goals at the expense of their own self-interest. In these cases, principal-agent theory predicts noncompliance on the part of agents, also referred to as agency loss (Gailmard, 2012; Moe, 2003).

Several additional issues can further complicate the basic principal-agent relationship, especially in the public sector. First, there is often a hierarchy of principals and agents rather than a single, isolated relationship (Laffont, 1990; Moe, 1984). Second, agents are often required to perform multiple tasks and therefore have multiple outcomes for which principals could evaluate performance (Holmstrom & Milgrom, 1991). Problems occur when multi-tasking is required of agents, but principals only hold agents accountable to a single outcome or a narrow set of outcomes. In these cases, agents may focus a disproportionate amount of energy on the outcomes they are held accountable to at the expense of their other tasks. Agents may also have an incentive to use counterproductive tactics designed to show the appearance of high performance (Moe, 2003). A key to addressing the multi-tasking problem is ensuring enough of the relevant inputs, processes, and outcomes can be measured in a valid and reliable manner. Agents also should have the capacity and be provided with the necessary support to improve on the measured outcomes (McEachin, ND). However, the principal-agent relationship

becomes increasingly complex when hierarchical structures and multi-tasking are both present. In these scenarios, the incentives applied by multiple principals are weaker than if there were just one principal-agent relationship (Dixit, 2002). Public sector principal-agent relationships are also characterized by a lack of competition among organizations (e.g., schools) as well as agents who are more motivated by moral and ethical purposes than their private-sector peers (Dixit, 2002).

Principal-Agent Theory and Test-Based School Accountability. All of the features of the principal-agent relationship, as well as those specific to the public sector, apply to public education and school accountability. As Dixit (2002) argued,

> "the system of public education is a multi-task, multiprincipal, multiperiod, near-monopoly organization with vague and poorly observable goals (p. 719). This goes a long way toward explaining the disagreements over its problems and the inefficacy of the single-minded solutions that have been tried" (p. 719).

Despite the existence of multiple principals, state officials and federal policymakers largely control billions ($649 billion in 2015/14; Cornman et al., 2018) in public revenues spent on primary and secondary education. As O'Day (2002) argues, "[i]t is reasonable that the public and its representatives want to know where the money is going and what it is producing" (p. 293). It is clear that the general public, or perhaps more accurately, the legislators that represent them, are principals who hold considerable power in public education. Federal and state legislators, however, obviously cannot hope to educate the more than 50 million public school students in the United States by themselves (National Center for Education Statistics, 2018). Therefore, state legislators "hire administrators and teachers to educate children," creating a potential moral hazard situation (Moe, 2003,

p. 82). For one, legislators and educators have goals and interests that do not perfectly align (Rothstein et al., 2018). Legislators also cannot directly control the actions of educators and educators clearly have access to information on student achievement and school performance not easily accessible to legislators (Moe, 2003). The self-interests of educators also likely went beyond the narrow set of test-based goals NCLB required schools to pursue, whether those interests were professional in nature or related to pursing student outcomes not included in accountability systems.

The design of NCLB seems logical, if not necessary, when viewed through the lens of principal-agent theory. Federal and state legislators, representing the general public, have an obligation to hold schools accountable for achieving results with taxpayer dollars. Results in the case of NCLB meant improving student achievement, or more accurately, mathematics and ELA proficiency, as well as closing achievement gaps between low- and high-performing students.[1] And since legislators also "have difficulty monitoring the activities of schools, then educators might behave in a manner contrary to the interests" of said legislators, "it would follow that more effective monitoring of educators could result in improved student outcomes" (Figlio & Loeb, 2011, p. 386). As a result, NCLB mandated that states create accountability systems that publicly reported data on school performance and allocated rewards or sanctions based on performance.

Despite appearing logical, some scholars argue that NCLB's theory of change was "fatally simple" (Lee & Reeves, 2012, p. 210), while others maintain school accountability can have transformative effects on schools (Hess, 2002). For example, Hess argues high-stakes, test-based accountability systems generate a "coercive force of

[1] It should also be noted that starting in 2007/08, states were required to administer standardized tests in science once in grades 3-5, 6-9, and 10-12 (National Center for Education Evaluation and Regional Assistance, 2009). Also, many accountability systems prior to and during NCLB included measures other

self-interest" where "students and teachers are compelled to cooperate" or risk not graduating or being fired (p. 70). Those in the former camp maintain the complex web of principal-agent relationships in public education and the growing list of demands placed on educators produce an array of self-interests that students and teachers act on in addition to accountability incentives. There are also several other issues that threaten the effectiveness of school accountability policy as a mechanism for improving student achievement and school performance. These issues include but are not limited to a lack of school and state capacity, misaligned incentive structures, and the multi-tasking problem.

Multiple, Hierarchical Principal-Agent Relationships. There is a vast hierarchy of principal-agent relationships throughout public education (Ferris, 1992; Ladd & Zelli, 2002; Moe, 1984; West, 2009). Several different groups of stakeholders could be considered principals, including the general public, the Federal Government, states, counties, cities, school districts, schools, school boards, and of course, parents and students (e.g., Conley, 2003; Epstein, 2004). Defining school administrators and teachers as the agents seems somewhat more straightforward (Moe, 2003). The important takeaway is that agents may be obligated to follow the directives of multiple principals. At any given time, a teacher may be asked to pursue goals at the direction of their school principal, their district superintendent, their state's school accountability system, and in some cases, a second, separate state accountability system designed specifically to meet federal regulations (Polikoff et al., 2014). Situations of multiple, sometimes conflicting goals among all these different principals is a common occurrence in public education (Conley, 2003). Problematically, multiple principals with different goals will likely dilute the strength of incentives in school accountability systems (Moe, 2003). Thus, one partial

explanation for the relatively inconclusive, or at best, modestly positive effects of test-based accountability may be the influence of an inherently hierarchical structure in the public education system and the multiple principals with different, sometimes conflicting goals this system produces (Moe, 2003).

Lack of School and State Capacity. Another potential explanation for the ineffectiveness of school accountability incentives is a lack of school and state capacity. Figlio and Ladd (2015) argue "schools may have insufficient resources to effect serious change in student outcomes, and others may lack the leadership required for significant change" (p. 203). The authors further argue a key assumption embedded in the design of school accountability policy is that low-performing schools are failing simply because they are not effectively monitored. It is assumed effective monitoring coupled with explicit rewards and sanctions will motivate educators to improve student achievement and close achievement gaps. Several authors challenge this assumption and point to a lack of capacity and inadequate resources as an alternative explanation for low-performance (Figlio & Ladd, 2015; Figlio & Loeb, 2011; McEachin, ND).

Importantly, schools serving high proportions of students who are economically disadvantaged are more likely to lack capacity and face resource constraints. For example, high-poverty schools are more likely to have less qualified teachers (Clotfelter et al., 2007). High-poverty schools can also experience difficulty in retaining teachers due to a lack of adequate resources and poor working conditions (Simon & Johnson, 2015). School accountability is often presented as a magic bullet that can overcome the challenges posed by inadequate resources and a lack of capacity and support.

Furthermore, officials in some state departments of education report lacking the

internal capacity to design, facilitate, or implement effective interventions for low-performing schools (Forte, 2010). Inadequate funding, insufficient technical expertise, and modest technological capabilities plague most state departments of education (Brewer et al., 2013; Center on Education Policy, 2007). As Forte argues, these challenges exist at both the state and school levels and can lead to a situation where

> "publicly reporting [of] a school's poor academic performance may garner staff attention but may lead to incoherent stabs at change and to demoralization rather than to actual improvements in leadership, management, programming, or instruction in schools with limited capacity and no coherent support for improving that capacity" (p. 84).

Stated differently, test-based accountability relied more heavily on using rewards, sanctions, and public pressure as incentives to motivate change within schools rather than building the capacity of schools and states (Lee, 2006). Regardless, it appears even if these incentives generated the necessary motivation, schools and states may not have had the capacity to improve. In fact, recent research suggests school accountability systems that include strong capacity building supports may be most likely to be effective. Of five state-level studies on the effects of ESEA waivers, four produced either null (Dee & Dizon-Ross, 2019; Dougherty & Weiner, 2019; Hemelt & Jacob, 2017) or negative effects (Atchison, 2020), with several authors noting weak interventions coupled with poor fidelity of implementation in low-performing schools. However, one study that found positive effects credited school-level gains of 17% in math and 9% in reading to Kentucky's "reputation as an enthusiastic and energetic adopter and implementer of school reforms" as well as evidence that "comprehensive school planning" and "high-

quality teacher professional development" was implemented with high fidelity (Bonilla & Dee, 2020, p. 101). Thus, it does appear some states do have the capacity to adequately support low-performing schools in ways leads to positive outcomes.

Misaligned Incentives and the Multi-Tasking Problem. Another issue with test-based accountability centered on the perverse incentives created by reward and sanction schemes that relied primarily on standardized test scores to evaluate student achievement and school performance. As described briefly above, principal-agent theory holds that when agents are required to multi-task but are only held accountable to a single outcome or narrow set of outcomes they will focus a disproportionate amount of energy on improving performance on the measured outcome(s) (Dixit, 2002). This phenomenon is also related to Campbell's Law, the idea that "the more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social process it was intended to monitor" (Campbell, 1976, p. 54). Moreover, both large rewards and severe sanctions based on narrowly defined performance metrics may cause agents to chase short-term gains rather than invest in long-term improvements (Gibbons, 1998; Holmstrom & Milgrom 1991; 1994; Milgrom, 1988; Milgrom & Roberts, 1990). Indeed, research shows some schools reacted to NCLB's rewards and sanctions by chasing short-term gains on school-level standardized test score performance (Figlio & Loeb, 2011).

The most obvious and perhaps easiest to implement short-term strategy is referred to as curriculum narrowing (Au, 2007; Berliner, 2011; McMurrer, 2007; Pederson, 2007). As Berliner (2011) argues, the "most rational of the many responses" to the consequences attached to high-stakes standardizing testing is to increase the amount of instructional

time spent on mathematics and ELA (p. 299). Curriculum narrowing results in less instructional time being spent on subjects such as social studies, the arts, and physical education (McMurrer, 2007). Closely related, evidence suggests educators employ instructional tactics designed to help students perform well on standardized tests, commonly referred to as "teaching to the test" (Au, 2007; Jacob, 2005; Jennings & Bearak, 2014; Jennings & Lauen, 2016). Teaching to the test can include presenting content only in the context of the test, isolating fragmented pieces of information relevant to test items, and teaching skills specific to certain item formats (e.g., multiple choice). Teaching to the test can be harmful when it takes the place of helping students build their conceptual understanding within a subject area. The consequences of narrowly defined, performance-based accountability goals may go beyond lost instructional time for nontested subjects. Research suggests instructional practices geared toward mastery-oriented goals (i.e., conceptual understanding) are more likely to build students' higher order thinking skills than performance-based goals (Gafoor & Kurukkan, 2016).

The final set of responses associated with misaligned incentive structures and the multi-tasking problem are less rational and more exclusionary. One of the more problematic strategies resulted in schools systematically reclassifying low-performing students as students with disabilities so they did not contribute to the school's aggregated mathematics and ELA proficiency rate (Booher-Jennings, 2005; Cullen & Reback, 2006; Figlio & Getzler, 2007; Jacob, 2005). Schools that engaged in this process created a more high-performing testing pool in an effort to produce a better accountability rating. Other exclusionary tactics included encouraging absences (Cullen & Reback, 2006) and suspending low-performing students so they were unable to sit for standardized tests

(Figlio, 2006). It is also worth noting this practice directly opposes the intent of test-based accountability under NCLB, which was to close achievement gaps on standardized tests by focusing instruction on low-performing students. Instead, it appears some schools intentionally excluded low-performing students from taking standardized tests (Cullen & Reback, 2006; Figlio & Getzler, 2007).

Another exclusionary response occurred when schools focused instruction on the students near the cutoff for proficiency on standardized tests, also referred to as "educational triage" (Booher-Jennings, 2005; Lauen & Gaddis, 2016; Krieg, 2008; Neal & Schanzenbach, 2010; Reback, 2008). Focusing on students near the cutoff for proficiency on standardized tests represents an efficient way for a school to improve its accountability rating. When schools make this exclusionary calculation, they deem high-achieving students as safe bets and low-performing students as lost causes. This response is especially likely in accountability systems that rely on status rather than growth indicators. Status indicators measure aggregated proficiency in a single year whereas growth indicators measure a school's progress in improving aggregated proficiency from year to year. Accountability systems where schools are rewarded for growth across all students may be less susceptible to educational triage. With status indicators, schools are incentivized to focus on those students near the cutoff for proficiency since these students have the largest impact on a school's accountability rating. Conversely, growth indicators reward schools for improving the performance of all students rather than a few, although it is worth noting that students who reach a ceiling and cannot technically improve their performance may garner less attention from educators. That said, growth indicators provide schools with more of an incentive to focus attention on low- and mid-performing

students than status measures.

Despite the array of unintended consequences and the number of potential issues discussed above, test-based accountability has been credited with two main positive effects (Jacob, 2017). Perhaps the most notable is the fact that test-based accountability "has illuminated large and persistent inequities in student performance, many of which had been ignored in the past" (Jacob, 2017, p. 470). For this reason, some advocacy groups believe school accountability systems are critical to keeping the light shined on inequities and continuing to force policymakers, schools, and communities to devote time and resources to closing opportunity and achievement gaps (e.g., The Education Trust, 2020). School accountability has also been credited with spurring new research programs dedicated to better understanding the factors that influence student achievement and the development of school improvement interventions (Jacob, 2017).

Additionally, although research on the effects of school accountability are generally mixed overall, there appears to be an emerging consensus backed by literature reviews that show test-based accountability has had positive effects on student achievement (Figlio & Ladd, 2015; Figlio & Loeb, 2011; Lee, 2008; National Research Council, 2011), but little to no effect on closing achievement gaps between low- and high-performing students (Gaddis & Lauen, 2014; Lee, 2008; Figlio & Loeb, 2011). However, research that isolates the specific mechanisms that produce these positive effects are rare (Figlio & Ladd, 2015). Some evidence suggests a strong commitment to reform at the state-level and strong fidelity of implementation for interventions focused on low-performing schools may strengthen the effectiveness of accountability incentives (Bonilla & Dee, 2020). Other evidence suggests teachers may improve their effort and

effectiveness when they feel more pressure from having a larger proportion of their students close to the cutscore for proficiency on state tests used for school accountability purposes (Macartney et al., 2018). Taken together, longstanding bipartisan political support, the continued exposure of achievement inequities, and the existence of research showing accountability can produce positive effects helps explain why the intent, design, and implementation of school accountability policy has remained relative consistent during the past two decades. Though, as the following sections describe, there has been at least one notable change in the design of school accountability policy.

Expanding the Outcomes Employed in School Accountability Policy

In 2011, the Obama administration initiated the ESEA waiver process partially in response to the steady stream of criticisms leveled at NCLB and to provide states with relief from some of the law's statutory requirements, including reaching 100% proficiency by 2014 (Duncan, 2011; Polikoff et al., 2014). The ESEA waiver process provided state policymakers with a limited opportunity to submit proposals to redesign their school accountability systems. The first state ESEA waivers were approved in February of 2012 and all waivers expired in August of 2016 upon passage of ESSA. Where ESEA waivers provided states with an opportunity to employ new indicators in school accountability, ESSA required it. Importantly, as described below, ESSA allowed states to maintain the design of the school accountability systems created under ESEA waivers. It also created an opportunity for state policymakers to expand the indicators used in school accountability beyond what was proposed in ESEA waivers, an opportunity many states took advantage of. The following sections provide a brief overview of the ESEA waiver process and ESSA as well as the available research on the

various accountability system design decisions made by states during these periods.

Elementary and Secondary Education Act Flexibility Waivers

All State Education Agencies had an opportunity to submit an ESEA waiver to the U.S. Department of Education, with 45 eventually receiving approval. Approved ESEA waivers had to adhere to a set of principles laid out by the U.S. Department of Education. Many of these principles dealt with how states measured student achievement and school performance.[2] To be approved, state policymakers had to describe what subjects they would administer annual standardized tests in, how they would set annual goals for increasing student proficiency, how they would generate student subgroups for accountability purposes, and how they would measure student proficiency. A final principle allowed states to propose new "systems of differentiated recognition, accountability, and support" (Duncan, 2011, para. 3). States still had to identify the highest and lowest performing schools as they did during NCLB. Using at least achievement and graduation rate gap data, states had to identify "reward" schools as the highest performing schools and "focus" and "priority" schools as the lowest performing 10% and 5% of schools in the state, respectively. ESEA waivers also provided states with an opportunity to use new indicators in addition to standardized test scores to identify reward, priority, and focus schools. Many state policymakers seized this opportunity by adding new indicators to their school accountability system (Polikoff et al., 2014).

Research on the design of ESEA waivers suggests state policymakers improved the construct validity of their school accountability systems (Polikoff et al., 2014; Wrabel et al., 2018). To these researchers, improving construct validity simply meant using

[2] The Federal government also required states to respond to principles relating to creating systems of teacher evaluation. The limitations associated with this policy requirements as well as the threats and opportunities posed by AP and other policies are described in the Discussion section.

additional indicators other than state standardized test scores to measure school performance. For example, in their study of 42 ESEA waivers, Polikoff et al. found 23 states used a composite index to identify focus and priority schools, with 15 states employing indicators not calculated from state standardized test scores. Though many states used new indicators in school accountability, some did not use them to identify focus or priority schools. Additionally, despite appearing to improve construct validity, these new indicators "rarely account[ed] for a substantial proportion of the total" composite score used to measure school performance (Polikoff et al., 2014, p. 49).

The Every Student Succeeds Act

ESSA officially ushered in the era of expanded outcome accountability by mandating states use indicators other than standardized test scores and graduation rates to identify low-performing schools. For example, states now have to identify the bottom 5% of high schools using at least four indicators: (a) standardized test scores in mathematics and ELA, (b) graduation rates, (c) English language learner progress, and at least (d) one indicator of school quality. States also have to identify low performing high schools that graduate less than two-thirds of their students.

All states now operate under approved ESSA plans that were submitted in 2017. However, given their recent implementation there has been very little analysis of ESSA state plans. One study found it difficult to determine the purpose specific indicators served within accountability systems (Aldeman et al., 2017a; 2017b), an issue that also emerged during the ESEA waiver period (Polikoff et al., 2014). Many states also use one set of accountability indicators to assign low-stakes school accountability ratings and another set of indicators to calculate separate high-stakes ratings used specifically to

identify low-performing schools subjected to state mandated interventions. In other words, as states have expanded the type and number of indicators used to hold schools accountable under ESSA, the process used and indicators employed in identifying low-performing schools has also become more complex. AP exam accountability indicators help illustrate this complexity. As the following sections describe, states use AP exam accountability indicators in the calculation of both low- and high-stakes school ratings.

The Advanced Placement Program

Administered by the College Board, the AP program provides students with access to college-level content and the opportunity to acquire economic benefits by earning college credit via AP exams. Currently, the College Board offers 38 AP courses and associated end-of-course AP exams. AP exams represent "the most common means through which high school students can earn college credit in high school" (Kolluri, 2018, p. 5). The vast majority of four-year postsecondary institutions award college credit to students that score a 3 or higher on an AP exam (all AP exams are scored on a 1-5 point scale). For example, a College Board study found only eight postsecondary institutions out of 1,380 did not award college credit to students for qualifying AP exam scores (Adams, 2014). As discussed below, what constitutes a "qualifying score" varies by institution. Recent research also suggests earning AP credits while in high school is associated with reduced time to degree, doubling majoring, and participation in more advanced college coursework (Evans, 2019). In addition to the potential economic benefits associated with AP exams, students can derive academic benefits from AP course participation by being exposed to college-level content and developing time management and study skills (College Board, 2014; 2020b, Warne, 2017). Students can

also earn social benefits from AP participation. For example, students can use AP credits on their high school transcripts as an indicator of college readiness in an effort to improve their chances of earning acceptance at colleges and universities (Klopfenstein, 2010).

The history of AP, though, demonstrates these economic, academic, and social benefits have not been equally available to all students. The AP program originated in the 1950s with the intention of providing "superior" or "gifted" students with access to college-level content while in high school (Lacy, 2010). AP continued to cater to high-achieving, mostly privileged White students for several decades until the 1980s and 1990s when the increasing popularity of the program and federal and state financial support led to substantial increases in access for all students (Schneider, 2009). These efforts have resulted in increased AP course and exam participation for low-income, Black, and Latinx students overtime, although substantial disparities in AP participation and exam performance remain (Kolluri, 2018; Judson & Hobson, 2015; Malkus, 2016). For example, the most recent data from the College Board demonstrates the impressive reach of the AP program. In 2019, more than 16,000 public schools in the U.S. administered 4.4 million AP exams to 2.4 million students (College Board, 2019). The increased access, however, has been coupled with a decline in AP exam performance across time (Kolluri, 2018) and the AP program has continued to face persistent equity challenges. For instance, American Indian/Alaska Native, Black, Latinx, and Native Hawaiian/Pacific Islander students continue to pass AP exams at lower rates than White and Asian students (College Board, 2019). This equity challenge is important to note when considering the design of AP accountability incentives. States with AP exam accountability indicators largely employ status measures that assess average performance

across all students without requiring schools to disaggregate AP exam data by student subgroups. Only four states (District of Columbia, Kentucky, Massachusetts, and Mississippi) report average AP accountability exam data by subgroups.

Advanced Placement Indicators in School Accountability Policy

Based on an online search process conducted in May of 2020, 25 states employ AP indicators in school accountability. Of these 25 states, 21 states use school-level AP exam performance as an accountability incentive (the other four states use indicators associated with AP course-taking rather than AP exam outcomes). Of the 21 states, 19 have introduced AP exam accountability incentives in the timeframe that can be analyzed empirically (two states plan to introduce AP exam accountability incentives in the coming years but have not yet done so).

Descriptively, states that have introduced AP exam accountability indicators (hereby referred to as "policy-adopting states") differ from comparison states politically, geographically, and in size (i.e., the number of public school students in a state). For example, the Republican party carried slightly more than half (58%) of the policy-adopting states in at least three of the past four presidential elections (2004, 2008, 2012, 2016), compared to only six policy-adopting states (32%) for the Democratic party, with the remaining two policy-adopting states (11%) splitting two presidential elections a piece for both parties. When examining policy-adopting states by the geographic regions used by the US Census Bureau, more than half (53%) are located in the South, followed by the Midwest (21%), West (16%), and Northeast (11%). Policy-adopting states are also larger. The number of Grade 11 and 12 AP exam participants is highly correlated with the number of public school students within states ($r = 0.99$ in 2019). Policy-adopting states

accounted for 13 of the top 25 states when ranked by AP exam participation in 2019, with six (California, Texas, Florida, Illinois, Georgia, and Pennsylvania) of these states accounting for nearly half (49.16%) of all public school participants in 2019. By contrast, only one policy-adopting state is in the bottom ten of all states. Finally, as examined in more detail in the Discussion section, policy-adopting states also appear to have more AP policy activity overall than comparison states, on average.

One significant commonality among policy-adopting states is using the percentage of students who scored a 3 or higher on an AP exam as the performance metric schools are held accountable to. This is not necessarily surprising given the majority of postsecondary institutions that award students college credit for AP exam scores do so for scores of 3 or better (Adams, 2014). It therefore appears that most postsecondary institutions have historically followed the College Board and American Council on Education's (2020) recommendations for awarding college credit for AP exam scores of 3 or higher. These recommendations are supported by research conducted by the College Board that suggest students who score a 3 or better on an AP exam outperform their non-AP peers in subsequent college coursework (Dodd et al., 2002; Patterson & Ewing, 2013; Patterson et al., 2011). Yet, postsecondary institutions are trending toward awarding students college credit for only AP exam scores of 4 or 5, if at all. For example, in a study of 1,380 postsecondary institutions, Adams (2014) found that approximately 68% accepted credit for an AP exam score of 3 or higher, 30% for a 4 or higher, 2% for a 5 or higher, and less than 1% accepted no AP credit. Recent College Board research examining a representative group of 93 postsecondary institutions suggests the value of an AP exam score of 3 may be diminishing even further (Wyatt et

al., 2018). When looking specifically at the institutions that awarded credit for one or more of the four AP exams analyzed in this study, exactly half accepted an AP exam score of 3 or higher for college credit, 43% accepted only a 4 or higher, and 4% accepted only a 5 or higher. This trend aligns with recent research suggesting the postsecondary academic benefits students accrue from AP exam scores increases as students score higher, with a score of 3 providing the minimum level of benefits (Warne, 2017). The implications of using the AP exam score of 3 or higher as a threshold for success in school accountability systems is explored in greater detail in the Discussion section.

Outside of this one major commonality, the 19 states that have introduced AP exam accountability incentives vary in the design and use of AP exam indicators. In general, states fall into three categories. The first category includes nine states that use AP exam performance indicators as an explicit accountability incentive tied both to the calculation of school accountability ratings and the identification of schools in need of improvement. For example, Florida began using AP indicators for accountability purposes in 2010/11. A school's aggregated percentage of students who took and were successful on an exam associated with an accelerated course (e.g., AP, International Baccalaureate, industry certification) accounted for 18.75% of a school's A-F high-stakes accountability rating in 2010/11. As is the case in most states, "success" on an AP exam meant a student scored a 3 or higher on a 1-5 scale. In Florida's case, AP exam performance is an explicit incentive tied directly to rewards and sanctions within the state's school accountability system. Schools that received "D" or "F" grades were classified as focus and priority schools, respectively. Even within states with explicit incentives such as Florida, the strength of the incentive varies. For example, starting in

2013/14, Missouri assigned A-F school grades based on a 100-point scale. Missouri's college and career indicator made up 15% of a school's high-stakes rating. Theoretically, a school could use AP exam performance for all 15% of the college and career indicator. However, a school could use student performance on 10 other indicators, potentially weakening the incentive for schools to improve AP exam performance.

The second and third categories of schools use AP exam performance as an implicit accountability incentive. The second category includes seven states that use AP exam performance to calculate school ratings but not to identify low-performing schools that receive state mandated interventions. For example, since 2011/12, Georgia has used AP exam performance as one of several indicators in calculating low-stakes school ratings based on a 100-point scale. However, the stakes are implicit in Georgia's case because the state does not use AP exam performance as an indicator for identifying schools in need of improvement. Instead, it appears Georgia primarily relies on the public pressure generated from low-stakes school ratings as a mechanism for improving performance (Figlio & Loeb, 2011).

Kentucky provides an example of a state in the third category, one of three states that simply report AP exam performance data but do not use these data to calculate low- or high-stakes school ratings or to identify schools in need of improvement. Beginning in 2014/15, Kentucky reported the aggregated percentage of all students in every high school who took an AP exam and the percentage who scored a 3 or higher. AP exam performance data was also reported for student subgroups with more than 10 students. The third category of schools is similar to the second in that both rely on public pressure to generate improvements in performance. The major difference is that AP exam

performance did not factor into low- or high-stakes school ratings in these states, potentially further weakening the overall strength of the accountability incentive.

These three categories demonstrate different ways of conceptualizing the strength of the policy dosage associated with AP exam accountability incentives. An exploratory aspect of this study examines whether the strength of policy dosage is associated with state-level AP exam performance. A dummy coded variable is used to place the 19 states that have introduced AP exam accountability incentives into one of two categories: (a) states that simply report AP exam data or use AP exam data in the calculation of low-stakes school ratings and (b) states that use AP exam data in high-stakes ratings that lead to the identification of schools that receive state mandated interventions. Past research suggests schools in the first category respond similarly to accountability pressure regardless if AP data are factored into low-stakes school performance ratings for some schools but not others (Beach et al., 2019).

Past Research in the Context of AP Accountability Incentives

The policy reforms examined in this study are situated in literature on school accountability and the AP program, two areas of research with almost no overlap. The vast majority of quantitative research on the effects of school accountability has focused on either standardized tests directly aligned to state accountability incentives or separate low-stakes standardized tests, such as NAEP. The majority of policy research on the AP program focuses on resource-based initiatives (e.g., AP exam fee waivers, scholarships for qualifying AP exam scores). To my knowledge, there is no quantitative research that directly explores the influence of school accountability incentives on outcomes pertaining to AP exam performance. Therefore, before presenting the research literature on test-

based school accountability and the AP program, it is necessary to note the similarities and differences between AP exams, state standardized tests, and NAEP.

AP exams share some notable similarities with the standardized tests generally used by researchers to examine the effects of school accountability, however, several features of AP exams make them a distinctly different type of outcome. Similarities include the standardized nature of AP exams and the timing of AP exams and state standardized tests, both usually administered in the late spring of an academic year. With the former, the fact that AP exams are standardized across all schools and students makes them particularly useful for studies with a nationwide scope. With the latter, the fact that AP exams are administered at the same time throughout the U.S. helps with one aspect of analyzing policy effects (timing of the outcome). However, as discussed below, the precise timing of when states introduced AP accountability incentives varies widely.

A major difference between AP exams and the standardized tests used most often in accountability research relates to the nature of the stakes for schools, teachers, and students. State standardized tests used for accountability purposes are almost always high-stakes for schools, sometimes for teachers, and less often for students. Depending on the state, the consequences stemming from state tests can be as severe as school closures. Other states tie the standardized tests used for accountability purposes to teacher evaluation systems, which can result in teachers being denied promotion or being outright dismissed from their position (Kraft et al., 2018). For students, however, state tests are inconsequential and provide little to no educational currency (Conley et al., 2014), although in the past many states required that students pass an exit exam before graduating high school (Education Commission of the States, 2007). Fewer states require

such tests now, however (Education Commission of the States, 2017). National standardized tests, such as NAEP, are low-stakes for schools, teachers, and students, but can be considered relatively high-stakes for states due to the media attention generated by the annual publication of the Nation's Report Card.

By contrast, the stakes are arguably the highest for students when AP exams are considered. Students front the $95 cost associated with almost all AP exams in the absence of an exam fee-waiver (College Board, 2020d). In 2016, 29 states offered some sort of AP exam fee waiver for low-income students according to the Education Commission of the States. For some students, the cost of taking an AP exam is a bargain if they score well enough to earn college credits at the postsecondary institution they end up attending. AP exam scores can also be a useful indicator of college readiness for applications to postsecondary institutions (Klopfenstein, 2010). The stakes associated with AP exams for teachers are not directly observable. Unlike state standardized test results, AP exams do not appear to be tied to any state mandated teacher evaluation system (Education Commission of the States, 2016). However, AP exams could potentially create stakes for teachers if they are included in district- or school-level evaluation systems. Finally, the stakes associated with AP exams for schools relate to the publication of school rankings (e.g., *U.S. News and World Report Best High School Rankings*), and in at least 19 states, school accountability incentives.

Another notable difference between AP exams and other standardized tests pertains to the College Board's requirements for schools and teachers who intend on offering AP courses. Most often, the content on state standardized tests is directly aligned to state-mandated academic standards. For example, as of April 2018, 20 states

administered either the Smarter Balanced or Partnership for Assessment of Readiness for College and Careers testing systems, both of which are aligned to the Common Core State Standards (Education Commission of the States, 2018). How districts and schools align curriculum and instruction to state standards and assessments varies widely (Wong et al., 2015). Conversely, the College Board exerts significant control over AP course content. For example, every single AP teacher in the world must have their course syllabus officially approved by the Advanced Placement Course Audit (College Board, 2020a). Whereas a common tactic used by schools and teachers in response to accountability incentives aligned to state standardized tests was to narrow the curriculum and teach to the test (Berliner, 2011; Jennings & Bearak, 2014), AP courses are explicitly designed to prepare students for the AP exam associated with the course.

A final difference between state standardized tests and AP exams relates to student participation. Depending on each state's opt out policies and other participation requirements, nearly all students are required to sit for state standardized tests, whereas AP exams are voluntary and completed by fewer students. For example, in 2019 approximately 4.4 million students nationwide completed an AP exam out of nearly 7.3 million Grade 11 and 12 public school students in the U.S.—roughly 60% of all students.

Lessons for Test-Based Accountability Research

Lessons from past research on school accountability incentives helped inform the methodological approach employed in this study. Researchers have typically taken two main approaches to studying the effects of accountability policy. One set of studies examines how accountability has influenced NAEP scores (e.g., Lee, 2008; Wong et al., 2015) and other outcomes not directly connected to school accountability systems (e.g.,

Cronin et al., 2005; Jennings & Lauen, 2016). These studies typically include a national dataset comprised of all or several states. These studies also focus on the effects of the broader accountability system rather than attempting to isolate the incentives associated with individual indicators, such as AP exam indicators.

Other studies are state- or local-specific. These studies either examine changes pre- and post-accountability across an entire system or take advantage of differing levels of accountability incentive strength within states. The former set of studies (see Lee, 2006) are largely limited by a lack of a comparison group. The latter takes advantage of discontinuities in accountability incentives to create rigorous quasi-experimental designs (e.g., regression discontinuity). This study includes elements from both types of studies while still being unique in its design. Similar to the first group of studies, the current research examines a common outcome measure across states, AP exam scores. However, in contrast to these studies, AP exams are explicit measures in accountability systems rather than separate low-stakes outcomes. Similar to the second group of studies, this study takes advantage of differing levels of incentive strength. However, the exploratory policy dosage analyses conducted in this study focus on the incentives associated with individual indicators rather than pressures that differentially impact specific schools.

In terms of expected effects, past literature reviews suggest test-based accountability has had modest, positive effects on aggregated state-level standardized test score performance (Figlio & Ladd, 2015; Figlio & Loeb, 2011; Lee, 2008; National Research Council, 2011). Effects have appeared "far more clearly and frequently for mathematics than for reading" (Figlio & Ladd, 2015, p. 206). Some research suggests students in the middle of the achievement distribution may benefit from accountability

more than students on the low- or high-end of the distribution (Neal & Shanzenbach, 2010; Lauen & Gadis, 2012). The conclusion of a modest, positive effect of school accountability on student achievement is an evaluation based on averaging the net effect across all relevant studies. However, not all studies on school accountability incentives have generated positive, significant effects. In fact, the variation across individual studies has caused some scholars to assert the research on test-based accountability is "decidedly mixed" (Figlio & Ladd, 2015, p. 205; Jacob, 2005, p. 763; Springer, 2008, p. 556). Recent research on ESEA waivers helps illustrate the decidedly mixed nature of school accountability studies. That is, the handful of state-specific studies focused on ESEA waivers have produced positive (Bonilla & Dee, 2020), null (e.g., Dee & Dizon-Ross, 2020), and negative effects (e.g., Atchison, 2020).

One factor that may explain variation across studies relates to the accountability consequences attached to school performance (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Hanushek & Raymond, 2005; Wong et al., 2015). For example, Hanushek and Raymond (2005) compared states with consequential accountability systems (i.e., publicly reported test score data tied to explicit consequences) to states without such systems prior to NCLB, finding a positive association between consequential accountability and increases in NAEP math and reading scores. Hanushek and Raymond's conclusions are consistent with Carnoy and Loeb (2002), although the two studies used a slightly different system for classifying states with strong, moderate, or weak accountability systems. On the other hand, Lee and Wong (2004) found no dosage-effects for states with consequential accountability systems prior to NCLB, which Dee and Jacob (2011) attribute to differences in how the authors determined the strength of

each state's accountability systems.

In more recent studies, Dee and Jacob (2011) and Wong et al. (2015) use two different coding schemas and methodological approaches to analyze the degree to which policy dosage relates to school accountability outcomes. Dee and Jacob adopted Hanushek and Raymond's (2005) consequential accountability classifications with slight modifications to determine if states that did not have such systems saw a policy shock from NCLB. The authors found that the consequential accountability requirements associated with NCLB led to improvements in NAEP fourth and eighth grade math scores but not fourth grade reading. The authors attribute these outcomes to the strong accountability requirements mandated by NCLB. Using a slightly different approach, Wong et al. (2015) classified states as either high-, mid-, or low-proficiency based on the percentage of students meeting proficiency standards on state standardized tests, aggregated across fourth and eighth grade math and fourth grade reading. States with fewer than 50% of students meeting proficiency standards were classified as high-proficiency states. Wong et al. argued high proficiency states had more difficult tests, more rigorous academic standards, and more ambitious school improvement goals. As a result, these high-proficiency states identified more low-performing schools than other states and "had to undertake more frequent and more stringent school changes" (p. 252). High-proficiency states also initially had lower average NAEP scores than mid- and low-proficiency states. Wong and colleagues found high-proficiency states saw greater gains on NAEP and closed the gap with low- and mid-proficiency states post-NCLB, with stronger effects in fourth and eighth grade math than in fourth grade reading.

Recent research focused on the ESEA waiver period also addresses dosage

strength, albeit from a different perspective. This growing body of literature focuses on the varying treatment strength associated with focus and priority schools, the designations the Federal government required waiver states to apply to the lowest performing 10 and 5 percent of schools in state, respectively. The Federal government provided states with considerable latitude in designing interventions for focus schools whereas states had to institute school turnaround interventions for priority schools that closely resembled the models required by NCLB (Bonilla & Dee, 2020). Studies examining focus schools are mixed, with large school-level math (17%) and reading (9%) gains in Kentucky (Bonilla & Dee, 2020) and null results in Louisiana (Dee & Dizon-Ross, 2019) and Rhode Island (Dougherty & Weiner, 2019). Hemelt and Jacob (2017) reported a small reduction in the within-school math achievement gap in Michigan focus schools, a finding that was explained by stagnant performance from low-achievers and declining performance among high-achievers. The few studies to examine priority schools show null results in Michigan (Hemelt & Jacob, 2017) and negative effects in New York elementary and middle schools (Atchison, 2020). One explanation for the mixed findings described above may result from the differing levels of capacity at the state-level and the varying design and implementation of interventions (Bonilla & Dee, 2020; Polikoff et al, 2014).

The literature above suggesting stronger accountability consequences are associated with positive effects on standardized test scores indicate that the strength of AP accountability incentives is a factor worth exploring. This collection of literature also shows analysts take varying approaches to modeling policy dosage, a decision partly dependent on policy context. Notably, all of the above studies either focused on a low-stakes exam (i.e., NAEP) or state standardized tests that factored into high-stakes school

ratings used to identify low-performing schools that received state mandated interventions. Conversely, AP exam data factors into high-stakes ratings in only nine of the 19 policy-adopting states. The other 10 policy-adopting states rely on weaker accountability mechanisms (incorporating AP exam data in low-stakes school ratings or simply reporting these data publicly). Therefore, I created a policy dosage variable that is conceptually distinct from past approaches to modeling the strength of school accountability incentives. This dosage variable will be used to examine associations between the strength of accountability incentives attached to AP exam indicators and average state-level AP exam scores. The exploratory policy dosage analyses will contribute to the research base presented above and provide insights for future research on the effects of individual accountability indicators.

Other research on test-based accountability points to potential unintended consequences of AP exam accountability incentives. Of all the potential unintended consequences, perhaps the most problematic would be the systematic exclusion of low-performing students from participating in AP courses or sitting for AP exams. On the one hand, the likely intent of using AP indicators in accountability policy is to incentivize schools to provide more access to AP courses and encourage more students to complete AP exams. However, if schools are rewarded or sanctioned based on *average* AP exam performance, schools may also have an incentive to encourage only high achieving students to sit for AP exams. This unintended consequence manifested itself during NCLB through the reclassification of low-performing students into student subgroups not subjected to state testing requirements (Booher-Jennings, 2005; Cullen & Reback, 2006; Figlio & Getzler, 2007; Jacob, 2005).

The only study that explored the tension between expanding AP access and improving average exam performance suggests this unintended consequence may occur (Rowland & Shircliffe, 2016). The authors studied one school's response to the confluence of two separate policies, the Florida Partnership for Minority and Underrepresented Student Achievement Act and the state's introduction of an AP exam accountability indicator. The former was designed specifically to increase participation in AP courses and the latter focused on improving school-level average AP exam performance. Interviews with teachers exposed "tensions between efforts to expand AP access while confronting accountability measures that often encourage teacher resistance to such expansion" (p. 417). Of particular concern among teachers was that accountability policies, which first focused on expanding participation and later exam performance, did not address "other factors related to college readiness, such as student preparedness, teacher professional development, and socioeconomic context" (p. 417).

Lessons for AP Research

Past research on the strength of accountability interventions and the study from Rowland and Shircliffe (2016) point to two critical questions: (a) what type of resources and training are needed to improve school-level AP exam performance and (b) how well do the states that introduced AP exam accountability indicators provide this support to schools? First, Kolluri (2018) puts forth three potential explanations for the "dual challenge" of AP equity and excellence. The challenge Kolluri refers to centers on efforts to expand AP access for more and more students while also ensuring the program has positive effects on students' college readiness and success. One explanation is that students do not perform well on AP exams when they are not prepared to do so. Scholars

that ascribe to this line of reasoning suggest students are being placed into AP courses without the prerequisite knowledge and skills necessary to be successful (Klopfenstein, 2009; Lichten, 2010; Loveless, 2009). These scholars go on to argue the content and instruction in AP courses is slowly declining in rigor due to the wide range of ability within AP courses. Kolluri argues that under this "logic, the underrepresentation of marginalized students is merely an unfortunate consequence of their underachievement" (p. 30). Although this explanation seems oblivious to the structural, social, and political barriers preventing access to and success in AP (Kolluri, 2018), it suggests one potential response by schools to the use of AP outcomes as an accountability indicator is to limit access to AP courses and exams (Rowland & Shircliffe, 2016).

Another explanation is that "the nature of AP curriculum may be poorly suited to students from historically marginalized backgrounds" (Kolluri, 2018, p. 30). Kolluri draws on several lines of research (e.g., culturally relevant pedagogy; Ladson-Billings, 1995) to argue the expansive nature of AP curriculum makes it less likely for students from historically marginalized communities to connect to and succeed in AP courses and on AP exams. In one study, Baker-Bell (2013) found that critical language pedagogies were an effective method for engaging Black students in AP English. The findings from Baker-Bell's study suggests additional training on culturally relevant practices for teachers in schools serving high proportions of historically marginalized students may be one method that improves AP outcomes (Kolluri, 2018).

Kolluri's (2018) final explanation suggests the pervasive, though declining inequalities in AP access, and the persistent gaps in AP exam performance, are both a product of social reproduction. For example, Klugman (2013) used the theory of

effectively maintained inequality (Lucas, 2001) to explain why a suite of California policies designed to improve AP access widened the gap between the least and most advantaged schools. The theory of effectively maintained inequality suggests parents with students in the most advantaged schools will push administrators to continually increase AP course offerings to maintain their relative advantage over less advantaged schools. Beach et al. (2019) found similar results when examining the inclusion of an AP access indicator in Pennsylvania's school accountability system. Similar to Klugman's findings, Beach et al. reported an initial increase in AP course offerings across all schools immediately after the policy intervention, but that the gap between schools with the most and least AP course offerings widened across time. This finding was especially problematic given that Pennsylvania's AP accountability incentive was specifically designed to improve AP access in schools with the fewest AP course offerings.

Within school structural barriers and dynamics, such as academic tracking (Oakes, 1985; 2005) and gatekeeping (Rowland & Shircliffe, 2016) may also play a role in social reproduction by ensuring low-income and students of color do not participate in AP. Academic tracking occurs when school administrators and teachers place students into different courses based on perceived ability level. Oakes (1985) found considerable gaps between different race/ethnicities and with regard to socioeconomic status, with non-White and low-income students persistently underrepresented in AP courses. A mix of structural, social, and political barriers make it difficult to dismantle tracking policies, whether explicit or implicit, though a small number of schools have done so with success (Oakes, 2018). Closely related, gatekeeping occurs when school leaders and teachers deny access to students they perceive will not benefit from AP participation or who are

excluded based on discretionary placement policies (Dougherty, et al., 2017). In a study conducted at one Florida high school, Rowland and Shircliffe found policies designed to expand AP course access for all students conflicted with accountability incentives for improving aggregated school-level AP exam performance. The conflict generated by the policy was related to the idea that expanding access to students who may not perform well on AP exams may punish the school with a lower accountability rating. Thus, the school as a whole had an incentive to ensure only high-performing students participated on AP exams, one form of gatekeeping.

Unfortunately, as Kolluri (2018) notes, "existing research does not provide satisfactory answers" for overcoming the dual challenge of achieving AP equity and excellence or sufficient evidence for any of the potential explanations above (p. 32). Similar to school accountability, much of the research on AP has "been largely atheoretical" and "insights on how schools and districts might interrupt patterns of AP inequality are few" (Kolluri, 2018, p. 32-33). Yet, despite a lack of direction from research, states have instituted numerous policies designed to expand AP access and improve student performance. For example, the Education Commission of the States (2016) reports that all but four states have at least one explicit AP policy.

Research on these policies is limited and has produced mixed results, with most focused on resource-based incentives designed to improve AP equity and excellence. For example, Jackson (2010) found that the Texas AP Incentive Program, which provided cash to teachers and students for passing AP exam scores, improved AP course participation and exam performance. Studying a different policy and using a different design, Kramer (2016) found states that weighted AP course-taking in the calculation of

grade-point averages used to award scholarships to students experienced higher increases in AP course participation and exam performance than comparison states. Jeong (2009), however, found students in the five states that awarded scholarships to students and provided cash to teachers for AP exam performance were not more likely to take AP exams than students in states without merit-based AP policies. Similar null findings were found in a study of AP exam fee waivers in Texas (Klopfenstein, 2004).

In addition to school accountability and AP research, insights from principal-agent theory provide further context for AP accountability incentives. First, multiple, hierarchical principal-agent relationships are inescapable when it comes to decisions surrounding AP. Teachers are undoubtedly responsible for instructing students on AP content. However, a school's decision to offer an AP course, who will teach it, and what students will enroll, involves a web of interactions between district officials, school administrators, teachers, parents, and students—all with potentially different interests and goals with regards to AP. This complex collaboration could influence the strength of state accountability incentives, especially in states that use implicit incentives such as relying only on public pressure to improve AP exam performance. Other accountability goals may take precedence in situations such as these.

Second, state policymakers partially addressed the multi-tasking problem inherent to test-based accountability by adding new indicators to their school accountability systems. However, it does not appear states created corresponding capacity-building mechanisms that might aid schools in improving AP exam performance. In other words, the school improvement mechanisms states must employ to improve low-performing schools does not necessarily include resources, training, or support specific to AP. An

initial review of ESEA waivers and ESSA state plans shows policymakers have not addressed any of potential barriers to AP equity and excellence, such as tracking, gatekeeping, cultural relevant pedagogy, or low prior achievement (Kolluri, 2018). For example, a review of approved ESEA waiver applications and ESSA state plans in the 19 policy-adopting states demonstrated relatively few mentions of "Advanced Placement" or "AP" and almost no mention of specific capacity building or other accountability mechanisms designed to improve AP exam scores. The results of this brief document analysis are discussed in more detail later in the Discussion.

Taken together, research on the effects of school accountability suggests states may see modest, positive effects on state-level AP exam performance from the introduction of AP exam accountability incentives. The school accountability literature also suggests states with strong systems of consequential accountability may be more likely to see positive effects. However, the strong interventions studied by these scholars were tied directly to state test scores. As described in the Discussion, little evidence suggests states are explicitly targeting AP performance in interventions for low-performing schools. At the very least, none of the ESEA waivers or ESSA plans analyzed addressed factors that may influence AP exam performance (Kolluri, 2018). Consequently, it seems just as likely that AP exam accountability incentives will not lead to improvements in the capacity of schools and teachers to improve AP outcomes.

Finally, studying the influence of AP exam accountability incentives is more than an academic exercise. Practically speaking, states have limited resources to allocate toward improving student AP outcomes (Education Commission of the States, 2016). If the effect of introducing AP accountability incentives is null or negative it would imply

states should consider alternative policies for improving AP outcomes. The cost of introducing AP indicators is not trivial. For example, the state of California commissioned at least one study (Conley et al., 2014) and spent nearly three years selecting, designing, and implementing individual measures for its new college and career accountability indicator. The findings from this study will provide one of the first pieces of evidence necessary for a more expansive analysis that explores whether the potential benefits of introducing AP accountability indicators outweigh the costs.

From a principal-agent theory perspective, all states that have introduced AP exam accountability incentives rely on the pressure generated from publicly reporting school data to induce schools and teachers to pursue the goal of improving students' AP exam performance. Some states also use AP exam data in the calculations of low- or high-stakes school ratings as an additional incentive. Policymakers spent the time and resources to create AP exam accountability indicators presumably because they assumed the incentives attached to these indicators would induce behavioral change that improved students' AP exam performance. However, to my knowledge, no empirical research has tested this assumption. This study is designed to fill this gap. With these considerations in mind, the following section describes the methodological approach taken to address the following overarching research question: *Did the introduction of AP exam accountability indicators impact state-level average AP exam performance?*

CHAPTER II

METHOD

Before conducting analyses, I complied 23 years of data from the College Board's state and national reports to create a dataset with individual observations for public school students in policy-adopting and comparison states as well as private students across all states. College Board data was merged with state demographic information from the National Center for Education Statistics. After compiling these data, the analyses performed were sequential, ranging from simple visual inspection to more complex hierarchical linear modeling. All analyses were conducted using base packages in R as well as a specialized package for hierarchical linear models (Bates et al., 2014).

The identification of causal effects in this study relied on the differential timing of implementation in policy-adopting states. Policymakers in policy-adopting states introduced AP exam data into school accountability either on their own accord or in reaction to one of two federal policy shocks, ESEA waivers or ESSA. The core feature of the research design employed in this study assumes the academic year policy-adopting states ultimately introduced AP exam data into school accountability varied as a function of a political process, legislative bargaining, and administrative procedures. I argue the combination of these factors led to near-random policy adoption timing among states.

There are several assumptions that underpin the estimates presented in this study. First, comparison states must be a valid counterfactual for policy-adopting states. That is, comparison states must not differ from policy-adopting states with respect to trends in AP exam scores, changes in observable state characteristics across time, or other observable or unobservable factors correlated with AP exam scores. This first assumption was tested

and addressed using two methods. First, event study models were constructed to both test for policy effects and examine pre-policy trends in AP exam scores. Different AP exam score trajectories would suggest that policy-adopting and comparison states do not have parallel trends. Second, I used time-varying state characteristic data to control for demographic differences between policy-adopting and comparison states.

A second assumption is that policy-adopting and comparison states must be exposed to the same internal validity threats, whether they are history, instrumentation, or other validity threats (Hallberg et al., 2018). Two nonequivalent comparison groups (public school students in comparison states in the main analytical models and private school students in the sensitivity analyses) were used to probe issues associated with internal validity threats. As Hallberg et al. (2018) note, "potential threats must operate differentially across groups to threaten inference" (p. 297). For example, if the College Board adopted a new policy providing additional time for students to complete AP exams, and that new policy resulted in an increase in average AP exam scores, we would expect to see a positive effect in both policy-adopting and comparison states. The use of a nonequivalent comparison group allows for this potential explanation to be tested empirically. For example, if both policy-adopting and comparison states responded similarly to the College Board policy change above, it would provide evidence that can be used to rule out this alterative explanation.

A second research design element—switching replications—has the potential to provide additional information for testing potential internal validity threats. In basic difference-in-differences and comparative interrupted time series (CITS) designs, the timing of the intervention is the same for both treatment and comparison groups.

However, with switching replications, introduction of a policy intervention is staggered across groups, which happened across states in the case of AP exam accountability incentives. As Shadish et al. (2002) argue, switching replications control "most threats to internal validity" and enhance external validity "because an effect can be demonstrated with two populations at two different moments in history, sometimes in two settings" (p. 192). Observing an effect in different settings and different years helps rule out alternative theories that rely on single event explanations. Outside of single event explanations, the only other events likely to influence the policy effect must happen within individual settings (e.g., state-specific AP policies) across time in "a pattern that mimics the time sequence of the treatment introductions" (Shadish et al., 2002, p. 147). The switching replication design is strong in cases where a policy effect is observed across settings and time since the likelihood of multiple events happening in different places at the same time as the policy is introduced is relatively implausible.

The third assumption relates to modeling pre-policy trends in AP exam performance. In difference-in-differences designs, trends in AP exam scores must be parallel across time between policy-adopting and comparison states. As described above, the parallel trend assumption was tested using event study estimates. Violation of the parallel trend assumption suggests parameter estimates in difference-in-differences may be biased and that a CITS design might be more appropriate. In CITS designs, the pre-policy intervention functional form (e.g., linear, quadratic) of pre-policy trends in AP exam scores must be correctly modeled. If the pre-policy functional form is modeled incorrectly it will lead to an inaccurate interpretation of policy effects.

Finally, in addition to addressing gaps in the literature on school accountability

and the AP program, the current research adds to methodological-focused literature in two areas: (a) the construction of school accountability policy dosage variables and (b) using difference-in-differences and CITS approaches to model the staggered implementation of a policy intervention. First, the existing research on the effects of policy dosage with respect to school accountability focuses on system-level distinctions (e.g., states with strong or weak sanctions) rather than parsing effects associated with individual indicators, as this study does. Second, a growing body of literature in econometrics is exploring methods for analyzing staggered policy implementation using difference-in-differences models (e.g., Goodman-Bacon, 2018). This approach contrasts with CITS designs that account for staggered implementation (e.g., Raudenbush & Bryk, 2002; Singer & Willet, 2003). As a result, this study will add to literature that compares and contrasts difference-in-differences and CITS techniques to modeling time series data in policy analyses (e.g., Gordon & Heinrich, 2004; Somers et al., 2013). In particular, given the staggered implementation design employed, this study will examine several notable distinctions between difference-in-differences and CITS techniques, such as approaches to modeling nested data structures, the coding of time and policy variables, and the construction and coding of comparison states.

The following sections describe the sample, decision-rules specific to constructing policy-specific variables, other measures and covariates, a description of the analyses performed, and approaches to probing validity threats and formulating causal inferences.

Data

Outcome data was collected from the College Board's publicly available national and state annual AP reports from 1997 to 2019. The College Board's annual reports

include data on average AP exam scores across all students as well as data disaggregated by public school students and individual courses. For example, it is currently possible to determine the number of public and private school students in Alabama that took an AP English Literature and Composition exam and what the average score was for each group in each year from 1997 to 2019. As will be described below, annual data from the National Center for Education Statistics' Common Core of Data and Private School Survey were combined with College Board's annual reports to generate descriptive information, create covariates, and conduct sensitivity analyses.

An online search process was conducted in May of 2020 to identify states with AP exam accountability incentives (i.e., policy-adopting states) and those without (i.e., comparison states). Table 1 below presents the results of the online search. States are assigned to the policy-adopting state group if school-level AP exam data are publicly reported on department of education websites for school accountability purposes, regardless if these data are simply reported or used to generate low-stakes or high-stakes school ratings. The academic year the AP exam accountability incentives took effect was also identified for each policy-adopting state.

The primary informational sources for the online search included data, resources, and reports on state department of education websites; ESEA waiver proposals; ESSA state plans; media reports; and archived webpages. The calendar year a state introduced its AP exam accountability indicator was used to determine the year of policy implementation. For example, Pennsylvania publicly released AP exam accountability data on October 13, 2013. Therefore, the calendar year of 2013 was used as the first academic year (i.e., 201*3*/14) in the state's academic year of policy implementation. This

Table 1

States with AP Exam Accountability Incentives

State Education Agency	Intervention Year	Accountability Indicator Design	Accountability Incentive Type
Alabama	2016/17	• A-F Grade • 1 of 11 sub-indicators • Qualifying AP exam score	• Sub-indicator Rating
Arizona	2017/18	• A-F Grade • 1 of 17 sub-indicators • Qualifying AP exam score	• Sub-indicator Rating • Identification
California	2017/18	• Dashboard • 1 of 9 sub-indicators • Qualifying score on two AP exams	• Sub-indicator Rating • Identification
District of Columbia	2018/19	• 5-Star Rating • 1 of 13 indicators • Qualifying AP exam score	• Rating • Identification
Florida	2010/11	• 0-1000 Rating • 2 of 6 indicators • Percent of total AP exam takers, Qualifying AP exam score	• Rating • Identification
Georgia	2011/12	• 0-100 Rating • 1 of 6 sub-sub-indicators • Qualifying score on two AP exams	• Individual Indicator Rating
Illinois	2015/16	• 4-point scale • 1 of 8 indicators • Taking an AP exam, Qualifying AP exam score	• Rating • Identification
Indiana	2005/06[e]	• A-F Grade • 1 of 4 sub-indicators • Qualifying AP exam score	• Sub-indicator Rating • Identification
Kentucky	2014/15	• 0-100 rating • Indicator with no rating • Taking an AP exam, Qualifying AP exam score	• Reported
Maryland	2008/09[d]	• 0-90 rating • Indicator with no rating • Taking an AP exam, Qualifying AP exam score	• Reported

Table 1 (continued)

State Education Agency	Intervention Year	Accountability Indicator Design	Accountability Incentive Type
Massachusetts	2018/19[a]	• 1-99 rating • Indicator with no ratings • Taking AP course, Qualifying AP exam score	• Reported
Mississippi	2015/16	• A-F grade[a] • 1 of 18 sub-indicators • Qualifying AP exam score	• Sub-indicator Rating • Identification
Missouri	2013/14	• 0-100 Rating • 1 of 6 sub-indicators • Qualifying AP exam score	• Sub-indicator Rating
New Mexico	2014/15	• A-F grade • 1 of 11 sub-indicators • Qualifying AP exam score	• Sub-indicator Rating • Identification
Ohio	2015/16	• A-F grade[b] • 1 of 6 sub-indicators • Qualifying AP exam score	• Sub-indicator Rating
Oklahoma	2014/15	• 0-100 Rating • 1 of 5 indicators (EC) • Qualifying AP exam score	• Rating
Pennsylvania	2013/14	• 0-100 Rating • 1 of 5 indicators (EC) • Qualifying AP exam score, AP course offerings in the core courses	• Rating
Texas	2013/14	• Met Standard or Improvement Required • 1 of 10/11 indicators (EC) • Qualifying ELA or math AP exam score at any point in high school[c]	• Rating
West Virginia	2018-19	• Dashboard • 1 of 3 sub-indicators • Qualifying score on two AP exams	• Sub-indicator Rating • Identification

Note. [a] Only academic year the policy intervention took effect; [b] Schools in Ohio did not receive A-F grades until 2017-18 despite reporting data on many of the individual indicators used to rate schools, including AP indicators in 2015/16 and 2016/17; [c] English Language and Composition, English Literature and Composition, Calculus AB, Calculus BC, and Statistics; [d] Maryland's school rating system went out of effect after 2016/17; [e] Indiana first began reporting AP exam data in 2005/06 and started using it for school ratings and identification in 2011/12 to the present day through the ESEA waiver process and ESSA

decision rule ensures schools have enough time, one entire semester at the very least, to respond to new accountability incentives before the administration of AP exams in the late spring of an academic year. Often, as was the case with Pennsylvania, which first announced the new accountability metrics in early 2013, schools have more than a full semester to react.

Measures

The data sources above were used to construct outcome, time, policy intervention, state characteristic, and policy dosage variables. These data sources were also used to collect information on AP exam participation rates and state demographics to inform this study's analytical approach and to further contextualize the findings.

Outcome Variables

Data collected from the College Board's annual state reports from 1997 to 2019 were used to create four outcome variables. These outcome variables include the state-level average exam score for four AP exams: English Language and Composition, Environmental Science, Psychology, and Statistics. Data on each outcome variable was collected for all public and private school students in all states.

State-level average exam performance was measured using four variables: AP English Language & Composition (*APExamELC*), AP Environmental Science (*APExamES*), AP Psychology (*APExamPsy*), and AP Statistics (*APExamStat*). I selected these four AP exams to ensure data was comparable across time and for sample size adequacy. First, in 2011/12, the College Board initiated a course and exam redesign process for 19 of its 38 courses. I excluded the 19 redesigned courses since AP exam data are not directly comparable from the pre- to post-redesign process, making any policy-

specific inferences highly susceptible to the internal validity threat of instrumentation. Next, of the remaining 19 non-redesigned courses, I used the same approach as Jeong (2009) by selecting the most popular English, mathematics, science, and social studies courses. Table 2 below presents the nationwide number of AP exam takers in all 19 non-redesigned courses. Employing AP courses with high numbers of test takers helps ensure data disaggregated by state generates adequate sample sizes.

Time and Policy Intervention Variables

The time and policy intervention variables capture the trajectory of change for each outcome variable as well as initial and long-term policy intervention effects. Three types of time variables were used, each specific to the event study, difference-in-differences, or CITS models. First, actual-year indicators were converted into relative-year indicators for the event study models. Relative-year indicators center time on the year of policy implementation in policy-adopting states. For example, if a state introduced AP accountability incentives in actual-year 2013, the relative-year was coded as 0. Relative-year policy indicators can improve statistical power given the staggered nature of implementation across states (no more than three states introduced AP exam accountability indicators in the same year). All comparison states were coded as -1 for all event study analyses since comparison states have no variance in relative-year values. As a result, the comparison states served as part of the reference group in addition to policy-adopting states in the year prior to implementation (Liebowitz et al., 2019). Next, I created dummy coded variables for the event study models to estimate year-to-year deviations in AP exam scores using the year immediately prior to policy implementation as the reference group for policy-adopting states.

Table 2

Nationwide Number of AP Exam Takers in Non-Redesigned Courses

AP Course Title	First Year Course Implemented	# of First Year AP Exam Takers	# of AP Exam Takers in 2018
Chinese Language and Culture	2007	2181	9489
Comparative Government & Politics	1997	5420	19535
Computer Science A	1997	5940	51645
Economics: Macro	1997	12702	119465
Economics: Micro	1997	9314	64532
English Language & Composition	***1997***	***126583***	***512676***
English Literature & Composition	1997	33367	343139
Environmental Science	***1998***	***3824***	***148053***
Human Geography	2001	2751	201977
Italian Language and Culture	1999	7	2390
Japanese Language and Culture	2007	1080	1962
Music Theory	1997	2760	16792
Physics C: Electricity & Magnetism	1997	4493	17863
Physics C: Mechanics	1997	9402	41971
Psychology	***1997***	***15370***	***273264***
Statistics	***1997***	***5986***	***188576***
Studio Art: 2D Design	2002	5967	30166
Studio Art: 3D Design	2002	1108	4846
Studio Art: Drawing	1997	4729	17395

Note. Figures represent the total number of U.S. public school students who took an AP exam in each of the respective subjects and academic years; ***AP courses selected for empirical analysis***
Source: College Board (1997, 1998, 1999, 2001, 2002, 2007, 2018)

A similar approach was taken in the difference-in-differences estimates, except that an interval variable was used to model trends in average AP exam scores in policy-adopting states. *TxTrend* was centered at zero in policy-adopted states on the year of policy implementation whereas all comparison states were still coded as -1. In the CITS models, *PreTxSlope* modeled the pre-policy intervention growth trajectory for all policy-adopting states and the entire 23-year timespan for comparison states. To model a linear functional form, the first year of data (1997) was coded as 0, the second year of data as 1, the third year of data as 2, and so on. Although visual inspection of state-level AP exam performance trends across time did not suggest the likelihood of higher-order growth trajectories, quadratic and cubic functional forms were also tested. These models failed to converge, and all CITS analyses therefore modeled changes in state-level average AP exam performance linearly.

Two policy-specific variables were used to model initial and long-term policy intervention effects and to account for switching replications (i.e., staggered implementation). Switching replications occur when two or more groups receive a treatment (i.e., policy intervention) at different times. In the example under consideration, states introduced AP exam accountability incentives in different years (see Table 1 on pages 48-49). In the difference-in-differences and CITS models, *PolicyImp* estimated changes in AP exam performance from the pre- to post-policy intervention time period within each policy-adopting state. In these states, every year preceding the initial year the policy intervention took effect was coded as 0. The first year the policy intervention took effect and every year after was coded as 1. Similarly, the variables used to model changes in growth from the pre- to post-policy intervention time periods within each state are

effectively the same in both the difference-in-differences and CITS models. In the difference-in-differences models, the post-policy slope was estimated using the interaction of *TxTrend* and *PolicyImp.* In the CITS models, *PostTxSlope* was coded 0 in the year preceding the policy intervention as well as the initial year of policy implementation, 1 in the second year, 2 in the third, 3 in the fourth, and so on. Both *TxTrend*PolicyImp* and *PostTxSlope* estimate deviations from the pre-policy slope between policy-adopting and comparison states.

The construction of time and the post-policy slope variables described above represent two different approaches, one specific to difference-in-differences models (*TxTrend)* and the other to CITS designs (*PreTxSlope* and *PostTxSlope*). The main difference between these two approaches has to do with the coding of comparison states. In difference-in-differences models, comparison states (i.e., public school students in states without AP exam accountability incentives) were coded as -1 in the pre-policy intervention time variable. As a result, comparison states do not contribute to pre-policy slope estimates (Liebowitz et al., 2019). The coding of the post-policy intercept variable (*PolicyImp*) is the same for both approaches.

The approach taken in this study to modeling the staggered policy implementation in CITS designs comes from Raudenbush and Bryk (2002) and Singer and Willet's (2003). Rather than centering policy implementation in policy-adopting states, the timing of when the coding begins for *PolicyImp* and *PostTxSlope* depends on what year each policy-adopting state first introduced AP exam accountability incentives. Also, in the CITS models, comparison states contribute to the pre-policy estimate, but are coded as all 0s for *PolicyImp* and *PostTxSlope*. That is, policy-adopting and comparison states stay on

the same 23-year time series. As such, the policy-specific coefficients in the CITS models are interpreted in reference to trends in average AP exam performance in comparison states. See Appendix A for an example of how the coding scheme described above can be applied to the time and policy intervention variables.

In the CITS models, a dummy coded variable was used to classify policy-adopting states into two groups. The first group, *TxIdentify*, includes policy-adopting states that employ AP outcome data in high-stakes ratings used to identify schools for state mandated interventions. *TxIdentify* includes nine policy-adopting states. The second group includes seven policy-adopting states that use AP outcome data to assign low-stakes school ratings (e.g., A-F grade, score from 0-100) and three policy-adopting states that use AP outcome data only as a publicly reported data point. The ten states in the second group served as the reference group when interpreting the policy-specific findings related to the strength of accountability incentives. Thus, the current research examines if using AP indicators to make high-stakes decisions about what schools are labeled low performing is associated with differential policy effects. Using AP exam data for high-stakes decisions constitutes a stronger form of accountability pressure than using AP outcome data only as a publicly reported data point or to assign low-stakes ratings. The rationale in this case is built partly on past research suggesting the accountability pressure associated with simply reporting AP data publicly and using these data in the calculation of low-stakes school quality ratings may be similar (Beach et al., 2019) and the hypothesis that using AP outcome data to identify schools for state mandated interventions via high-stakes ratings is a separate type of accountability pressure. That is, when AP outcome data are used for allocating state mandated interventions, it has

explicit consequences that go beyond the pressure to change that comes from publicly reporting school performance data (Figlio & Ladd; 2015).

State Characteristics

Data from the National Center for Education Statistics Common Core of Data and Private School Survey were used to create time-varying covariates for the event study, difference-in-differences, and CITS models. Specifically, variables for (a) the percentage of students eligible for free or reduced priced lunch, (b) the number of Grade 11 and 12 students, and the percentage of Grade 11 and 12 (c) Asian, (d) Black, (e) Hispanic, and (f) White students were created for public school students in each state in each year under analysis. These data came from the National Center for Education Statistics Common Core of Data. For the sensitivity analyses that included only private school students in policy-adopting states (described in more detail below), time-varying covariates were created for the number of (a) private school students and the percentage of (a) Asian, (d) Black, (e) Hispanic, and (f) White students in every other year (starting in 1998) in all states. These data came from the National Center for Education Statistics Private School Survey. The primary use for these covariates in the analytical models was to control for changes in state demographics across time. Additionally, these state characteristic data were also used to descriptively examine demographic differences between policy-adopting states, comparisons states, and private school students.

Analytic Procedures

The analyses began with simple visual inspection of the changes in average state AP exam scores centered on the year of policy implementation. From there, the analyses moved sequentially from event study estimates to difference-in-differences analyses to

CITS models that took full advantage of the 23-year time series.

Event Study

Event studies are a particularly useful research design for instances when units (in this case, states) implement a policy intervention at different time points (Borusyak & Jaravel, 2017). As with modeling staggered implementation in difference-in-differences models, event studies center time on the year of policy implementation in policy-adopting states. However, a central difference is the use of dummy coded variables for each year in event studies. As a result, event study estimates allow both for the examination of pre-policy intervention trends in average AP exam course offerings and post-policy intervention policy effects. Equation 1 below was used to fit the event study models for each AP exam, following a similar approach found in the school accountability literature (Liebowitz et al., 2019).

$$APExam_{st} = \sum_{r=-12}^{2+} 1\left(t = t_S^* + r\right)\beta_r + (\mathrm{x}_{st})\theta + \Delta_s + \Gamma_i + \varepsilon_s \quad (1)$$

In Equation 1, $APExam_{st}$ represents average AP exam scores for each state (s) and year (t) observation in each respective AP exam (i.e., English Language and Composition, Environmental Science, Psychology, Statistics). $APExam_{st}$ was regressed on a set of dummy coded variables, which represent individual years within the centered policy-adopting state time series (t_s^*), and the set of state characteristics (X) described above. Event study models also included state (Δ_s) and year fixed effects (Γ_i). The individual dummy variable coefficients β_r ($-11 \leq r \leq 2$) estimate the effect of introducing AP exam accountability indicators in the years before and after the policy intervention, compared to the year before implementation with all comparison states coded as -1. Only dummy coded variables that are balanced (include observations from all policy-adopting

states) are interpreted. The binned coefficients for 12 or more years prior to and 2 or more years post-policy intervention are not interpreted due an unbalance of policy-adopting states that resulted from the staggered implementation of AP accountability incentives across time.

Finally, three types of event study models were fit: (a) *baseline* models that included only state and year fixed effects, (b) *covariate-adjusted* models (Model A + covariates), and (c) and *weighted* models (Model B + AP exam participation weights). The inclusion of state and year fixed effects is a common econometric approach for estimating policy effects using within-state variation (Hallberg et al., 2018). Year fixed effects account for year-to-year deviations in AP exam performance across states whereas state fixed effects account for deviations in AP exam performance across states. As described above, the inclusion of time varying covariates was used to enhance the precision of estimates by controlling for state characteristics that may account for variance in average state AP exam scores. Finally, weighting estimates accounts for the wide variation in AP exam participation across states. Weighting also alters the interpretation of estimates. The coefficients for policy-specific variables are interpreted as the effect of accountability incentives on average state-level AP exam scores in average-sized policy-adopting states.

Difference-in-Differences

Following the event study analyses, I moved into a difference-in-differences framework that compares pooled estimates from the pre- to post-policy intervention periods. Additionally, the inclusion of a linear time trend variable allowed for the examination of any post-policy slope changes. Following Liebowitz et al.'s (2019)

approach, I estimated difference-in-differences models using Equation 2 below.

$$APExam_{st} = \beta_1 TxTrend_{st} + \beta_2 PolicyImp_{st} + \beta_3 TxTrend*PolicyImp_{st} + (x_{st})\theta + \Delta_S + \Gamma_s + \varepsilon_s \quad (2)$$

In Equation 2, *TxTrend* was coded as -1 in all comparison states and centered on the year of policy implementation in policy-adopting states. *PolicyImp* was coded as 0 for all comparison states in all years and policy-adopting states in the years prior to the introduction of AP exam accountability incentives and coded as 1 in the years of policy implementation in policy-adopting states. Thus, *TxTrend* estimates differences in pre-policy trends in average AP exam scores between policy-adopting states and comparison states and *PolicyImp* estimates changes in average AP exam scores immediately after the introduction of AP exam accountability incentives in policy-adopting states. In particular, the *TxTrend* coefficient allows for the examination of parallel trends between policy-adopting and comparisons states, a key assumption for difference-in-differences models. Finally, the interaction of *TxTrend*PolicyImp* measures any linear deviations in the post-policy slope among policy-adopting states. Similar to the event study analyses, baseline, covariate-adjusted, and weighted difference-in-differences models were fit. Finally, all three types of difference-in-differences models included standard errors clustered at the state-level to account for the nested data structure (time within states) that produces correlated errors across time within states (Abadie et al., 2017; Bertrand et al., 2004).

Comparative Interrupted Time Series

In the final set of analyses, I shift from modeling the nested nature of the state-level AP exam data using clustered standard errors in the difference-in-differences framework to using hierarchical linear modeling. Below, I describe the sequential model building and assumption testing process that was employed to build the 2-level piecewise

hierarchical growth models. I used Raudenbush and Bryk (2002) and Singer and Willet's (2003) approach to modeling the staggered implementation of AP exam accountability incentives across states using piecewise hierarchical growth models.

A two-step process was used to sequentially build the final analytical models. In step one, I determined if hierarchical linear modeling was an appropriate analytical approach given the nested nature of these data. Since *time* is nested within each individual *state*, it is likely that year-to-year observations within individual states share variance and that variance exists between states. Parameter estimates can be biased when this variance is not modeled correctly using a hierarchical structure. The unconditional means model in step one included only an outcome variable. The intraclass correlation coefficient described what proportion of the variance in each outcome variable was between states, which indicated if hierarchical linear modeling was an appropriate analytical approach for each individual outcome variable.

In step two, I used deviance tests to evaluate whether the time and policy-specific variables should include random effects. Specifically, I followed Singer and Willet's (2003) approach to compare the nested analytical models of interest. In total, eight different models were analyzed. The model that exhibited the best fit, as evidenced by statistically different deviance statistics, was chosen as the baseline model from which to conduct all other analyses. Estimates that rely on fixed effects assume states share the same trend in average AP exam scores whereas random effects allows these trends to vary across states, which adds more variance to the overall model (Hallberg et al., 2018). When random effects are used, states with less variance in AP exam scores across time will be weighted more in estimates of the relationship between time and AP exam scores

(Hallberg et al., 2018). Additionally, the use of random effects for policy-specific variables allows immediate (i.e., post-policy intercept) and long-term (i.e., post-policy slope) effects to vary in direction and magnitude across states, rather than constraining these effects to be equal across states. Equation 3 below represents the baseline model (including random effects), where Y_{it} represents the value for the AP outcome of interest for each policy-adopting and comparison state in each year.

$$APExam_{st} = \pi_{0t} + \pi_{1t} PreTxSlope_{st} + \pi_{2t} PolicyImp_{st} + \pi_{3t} PostTxSlope_{st} + \varepsilon_{s}$$

$$\begin{aligned} \pi_{0t} &= \beta_{00} + r_{0s} \\ \pi_{0t} &= \beta_{10} + r_{1s} \\ \pi_{0t} &= \beta_{20} + r_{2s} \\ \pi_{0t} &= \beta_{30} + r_{3s} \end{aligned} \tag{3}$$

The baseline CITS models were used for all additional analyses, including the construction of the covariate-adjusted models and those weighted by AP exam participation, similar to the process used for the event study and difference-in-differences models. I then added the policy dosage variable, *TxIdentify*, to the Level 2 equations for the baseline, covariate-adjusted, and weighted models in a separate set of analyses. The interactions between the policy intervention and policy dosage variables were designed to analyze if the strength of AP exam accountability incentives was associated with the size of the policy effect in policy-adopting states. In cases where these interactions are statistically significant, it indicates that policy dosage was positively or negatively associated with the individual AP outcome being analyzed.

To test assumptions in the CITS models, I used skewness and kurtosis statistics for each individual AP exam separated by year to examine the normality of the distribution. I tested the assumption that the Level 1 residuals are normally distributed

with a mean of 0 and variance of σ^2 using Levene's test for homogeneity of Level 1 variance and the visual inspection of Q-Q plots for each individual AP outcome. Initially, Level 1 residuals were only examined in the baseline model. However, I also used Levene's test and visual inspection of Q-Q plots for the weighted estimates to determine if the substantial increase in the number of statistically significant coefficients in these CITS models was potentially due to normality issues.

Sensitivity Analyses

Two types of sensitivity analyses were compared to the main analyses that included all policy-adopting and comparison states. First, a simple interrupted time series model that included only policy-adopting states was compared to the main models that included both policy-adopting and comparison states. Theoretically, any policy effects observed in the main models should remain consistent in the policy-adopting state only models, unless aspects of the comparison group composition bias estimates in the main models. Second, a separate model with policy-adopting states and a comparison group that included private school students within policy-adopting states was again compared to the main model results. Models that included private school instead of public school students were used to probe for unobservable state-level factors that may have influenced AP outcomes for public school students only. Observing consistent statistically significant effects in the main and private school student models would suggest that some event other than AP accountability incentives is responsible for the change in post-policy intercept or slope.

I follow the same approach as Wong et al. (2015) by constructing a separate model for the two separate nonequivalent comparison groups (i.e., comparison states used

in the main models; private school students in policy-adopting states used for sensitivity analyses). It is not possible to combine policy-adopting state, comparison state, and private school averages in the primary CITS models while also fully accounting for the nested nature of these data. Combining these three groups in one model would require coding private school students as a separate, nested group. This would fail to account for the shared variance between public and private school students within individual states, which could produce biased parameter estimates.

Coherent Pattern Matching

Finally, coherent pattern matching was used to evaluate the empirical coherence of the observed effects across multiple outcomes (Shadish et al., 2002; Wong et al., 2015). The primary purpose of coherent pattern matching is to strengthen the causal warrant. Coherent pattern matching builds off the work of Fisher (1935), Campbell (1966), Rosenbaum (2009, 2011), and Shadish et al. (2002)—all of whom advocated for using theory to predict complex patterns of outcomes as a way to improve causal inferences in quasi-experiments (Wong et al., 2015). Simple, straightforward research designs are preferred when random assignment is possible. But with quasi-experiments such as an interrupted time series, "causal inference is improved the more specific, the more numerous, and the more varied ... the causal implications of the treatment" (Shadish et al., 2002, p. 486). Coherent pattern matching essentially combines the information generated from analyzing multiple outcome variables with the different contrasts provided by nonequivalent comparison groups and switching replications.

Wong et al. (2015) provide a particularly strong example of coherent pattern matching. The authors analyzed how NCLB influenced state NAEP scores using "two

different ways of conceptualizing the treatment; three different kinds of comparison time series; two data sets; three combinations of academic grade and subject matter" (p. 250). One alternative explanation Wong et al. examined was that the observed improvement in NAEP scores after NCLB was in fact a statistical artifact resulting from students leaving Catholic schools for public schools. The authors showed that the only way this alternative explanation was plausible was if the following propositions were true:

> "(a) the students who left Catholic schools for public ones raised achievement there or (b) those who moved to non-Catholic private schools lowered means there, and (c) students leaving Catholic schools in [high-proficiency] states in 2002 were higher achievers than those exiting Catholic schools in [low-proficiency] states" (p. 271).

Each of the propositions above is associated with an internal validity threat that arises from potential selection issues. The authors used additional data to show only about 0.2% of students nationwide moved out of Catholic schools immediately after implementation of NCLB and that there were no noticeable shifts in the demographic composition of public and private schools nationwide. The authors deemed this alternative explanation implausible based on this additional information and the pattern of results that emerged from the nonequivalent comparison groups. The authors' methodological demonstration provides one approach to operationalizing coherent pattern matching in a way that renders most alternative theories implausible.

With respect to examining different causal implications from the introduction of AP exam accountability incentives, I addressed one overarching research question using four separate types of analyses, two different contrasts, and two different ways of

modeling accountability pressure (i.e., policy dosage). If results coalesce in statistical significance, direction, and relative magnitude across analyses and AP exams, it would provide strong evidence that the public pressure stemming from AP exam accountability incentives is the causal force producing the observed pattern of results. Consistent findings across the different analyses would also strengthen causal inferences. Ruling out internal validity threats and alternative explanations will take on more importance if a consistent pattern of results is not observed.

Finally, the main threats to external validity deal with generalizing the findings from this study across settings, outcomes, and treatments. The national scope of these data coupled with the novel use of additional research design elements has the potential to strengthen external validity with regard to both settings and outcomes. Stated differently, if a coherent pattern of results is found it provides strong evidence that the pressure generated from accountability incentives is the causal force influencing the state-level AP outcomes. Since the outcome domains being analyzed match the one being used for accountability purposes across policy-adopting states, and because all policy-adopting states are included in the analysis, coherent results would suggest the policy effect operates similarly across public schools in the U.S. If results do not cohere in statistical significance, direction, and relative magnitude it would provide weak evidence that the policy effect stemming from the introduction of AP exam accountability incentives is associated with changes in state AP exam scores. Weak evidence will suggest additional contextual information and other sources of evidence are needed to further probe the effects associated with AP exam accountability incentives.

The main caution with respect to external validity centers on variance in policy

strength across policy-adopting states. The common aspect of the policy intervention across the policy-adopting states being analyzed is the accountability pressure associated with publicly reporting school-level AP exam data. In each state, however, this pressure is either from the AP exam data being used (a) only as a publicly reported data point or to calculate low-stakes school ratings or (b) these data being used to identify schools in need of improvement using high-stakes ratings. The inclusion of a policy dosage variable is designed to model the strength of AP exam accountability incentives and explore this potential external validity threat.

CHAPTER III

RESULTS

The following section presents all relevant results, including a descriptive summary of the policy-adopting state, comparison state, and private school student groups; visual inspection of policy effects; and event study, difference-in-differences, and CITS models results. Additionally, this section presents findings from a series of exploratory and sensitivity analyses.

Descriptive Statistics

Table 3 presents descriptive statistics for public school students in policy-adopting and comparison states, private school students, and all students combined. State characteristic data was examined to compare policy-adopting and comparison states and to identify potential validity threats specific to changes in state characteristics at the point of policy intervention.

State Characteristics

As Table 3 shows, on average and when aggregated across years, policy-adopting states had a higher percentage of students eligible for free or reduced priced lunch, more Black and Hispanic students, fewer Asian and White students, and spent fewer dollars per student than comparison states. Using annual AP exam participation as a proxy for size (r = 0.99 between AP exam participation and public school student enrollment), policy-adopting states also were larger than comparison states, with an average of 145,679 and 49,891 AP exam participants in policy-adopting and comparison states in 2019, respectively. Trends in state characteristics across time were analyzed graphically, both comparing trends between policy-adopting and comparison states and examining the

Table 3

Descriptive Statistics

	Public			Private School Students
	All Students	*Policy Adopting States*	Comparison States	
Total States	51	19	32	51
State-Year Observations	1,173[a]	437	736	1,173[a]
AP Exam Performance				
All Exams	2.84 (0.29)	2.72 (0.34)	2.91 (0.24)	3.15 (0.23)
English Lang. and Comp.	2.87 (0.31)	2.72 (0.33)	2.95 (0.27)	3.31 (0.32)
Environmental Science	2.70 (0.50)	2.51 (0.46)	2.82 (0.48)	2.92 (0.60)
Psychology	3.01 (0.38)	3.05 (0.45)	3.13 (0.32)	3.29 (0.56)
Statistics	2.81 (0.42)	2.72 (0.44)	2.86 (0.40)	2.99 (0.53)
AP Exam Participation				
All Exams[b]	47.89 (85.98)	80.55 (124.32)	28.50 (40.09)	7.24 (11.01)
English Lang. and Comp.[b]	5.57 (10.46)	9.71 (15.35)	3.12 (4.29)	0.70 (1.18)
Environmental Science[b]	1.44 (3.13)	2.30 (4.36)	0.90 (1.84)	0.20 (0.35)
Psychology[b]	2.66 (4.79)	4.32 (6.80)	1.69 (2.61)	0.29 (0.53)
Statistics[b]	1.95 (3.43)	3.08 (4.89)	1.27 (1.78)	0.27 (0.41)
State Characteristics				
Grades 11-12 students[b]	129.61 (152.64)	203.17 (209.66)	85.59 (76.84)	97.59[d] (118.04)
% FRL	41.72 (12.22)	48.07 (12.40)	38.06 (10.50)	N/A
% Asian Grades 11-12	4.32 (8.66)	3.02 (2.71)	5.11 (10.69)	3.68 (5.82)
% Black Grades 11-12	14.44 (15.39)	20.55 (19.15)	10.76 (11.10)	5.88 (5.33)
% Hispanic Grades 11-12	10.86 (11.98)	14.68 (16.27)	8.55 (7.53)	5.38 (5.41)
% White Grades 11-12	66.76 (20.24)	58.48 (22.22)	72.75 (17.13)	63.11 (12.63)
Expenditures ($) per student[b]	9.51 (3.36)	9.10 (3.11)	9.75 (3.47)	N/A
% States with AP Exam School Accountability Indicator				
2006	2	5	0	0
2009	4	11	0	0
2011	6	16	0	0
2012	8	21	0	0
2014	14	37	0	0
2015	20	53	0	0
2016	25	68	0	0
2017	27	74	0	0
2018	31	84	0	0
2019	37	100	0	0

Note. [a] State-year observations are for individual AP exams. AP Environmental Science, Psychology, and Statistics had fewer than 1,173 state-year observations due to a lack of exam takers in some states in some years; [b] Figures are in the thousands (e.g., 10,562 = 10.56); [c] Average number of schools offering at least one student at least one AP exam; [d] Private school students across all grades; State characteristic data from the National Center for Education Statistics Common Core of Data and Private School Survey; FRL = Free or Reduced Lunch Eligible

trend for policy-adopting states centered on the point of policy intervention. Across all states, the percentage of students eligible for free and reduced price lunch rose steadily from 1997 to 2019 with trends for policy-adopting and comparison states strongly resembling one another. There was not a noticeable trend shift with respect to free and reduced price lunch for policy-adopting states when centered on the point of policy intervention. The same general linear slope emerged when examining average Grade 11 and 12 enrollment, again with no visible shifts at the point of intervention for policy-adopting states.

Similarly, the average percentage of Asian students increased at a slow, steady pace with one notable caveat. The percentage of Asian students in Hawaii dropped sharply from 75.09% in 2010 to 41.63% in 2011, which influenced the average in comparison states more severely than in policy-adopting states. In 2011, the U.S. Department of Education changed its racial/ethnic definitions by creating a new subgroup for Native Hawaiian or Other Pacific Islander students (among other changes). Therefore, to harmonize these data for Asian students, both the Asian and Native Hawaiian or Other Pacific Islander subgroups were added together for 2011 and beyond. As a result, there were no observable shifts at the point of policy intervention for either policy-adopting or comparison states. The percentage of Black students was relatively flat across time for both groups, although the trend within policy-adopting states exhibits a slight bell curve shape. The percentage of Hispanic students increased at a steady linear rate across all states with no noticeable shifts across time whereas the percentage of White students decreased a steady linear rate, again with no noticeable trend shifts.

The descriptive analyses suggest policy-adopting and comparison states are

different in terms of demographic composition. However, graphical displays do not show evidence of changing demographic trends at the point of policy intervention for policy-adopting states. The sensitivity analyses presented below statistically examined the internal validity threats associated with changes in state demographics across time.

AP Exam Performance and Participation

Among the four AP exams, students on average scored highest in AP Psychology, followed by AP English Language and Composition, AP Statistics, and AP Environmental Science. Comparison states performed better than policy-adopting states while private school students outperformed both groups across all exams (see Figures 1-4). Importantly, the trend in exam performance was relatively consistent and flat across time for all exams and all groups. As Figures 1-4 show, trend lines appeared to exhibit fewer and fewer noticeable trend deviations from the early 2000s until 2019, suggesting increased participation may have stabilized exam score trends across time.

In terms of participation (see Figures B1-B4 in Appendix B), AP English Language and Composition was the most popular exam, followed by AP Psychology, AP Statistics, and AP Environmental Science. Participation on each AP exam exhibited a nearly identical positive linear shape from 1997 to 2019, demonstrating the AP program's continued increase in popularity from year to year. Interestingly, despite only accounting for 37.25% of all states, policy-adopting states accounted for a majority of AP exam participants. For example, policy-adopting states accounted for 63.40% of all public school AP exam participants in 2019. The six policy-adopting states with the most AP exam participants in 2019 (California, Texas, Florida, Illinois, Georgia, and Pennsylvania) accounted for nearly half (49.16%) of all public school exam takers.

Figure 1

Average AP English Language and Composition Exam Scores

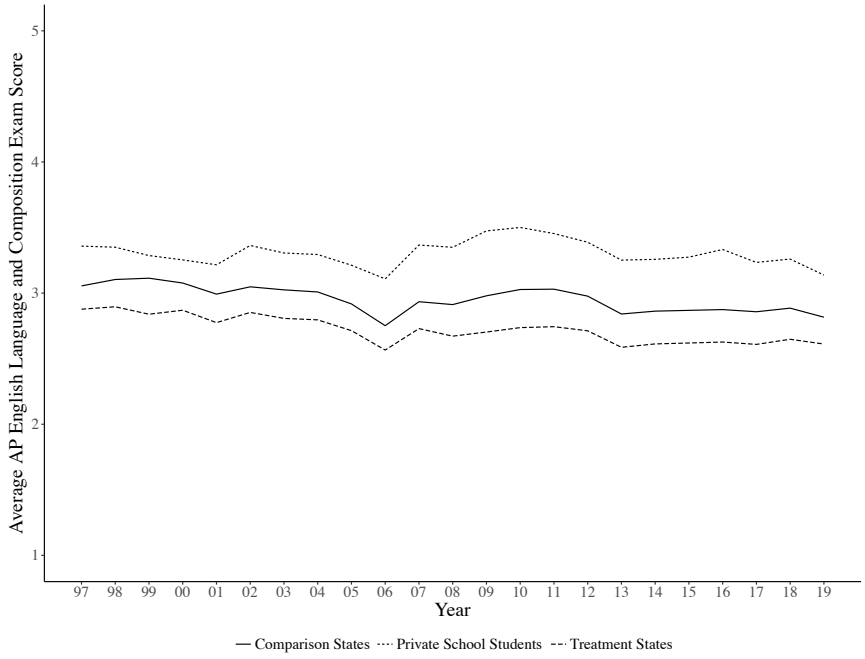


Figure 2

Average AP Environmental Science Exam Scores

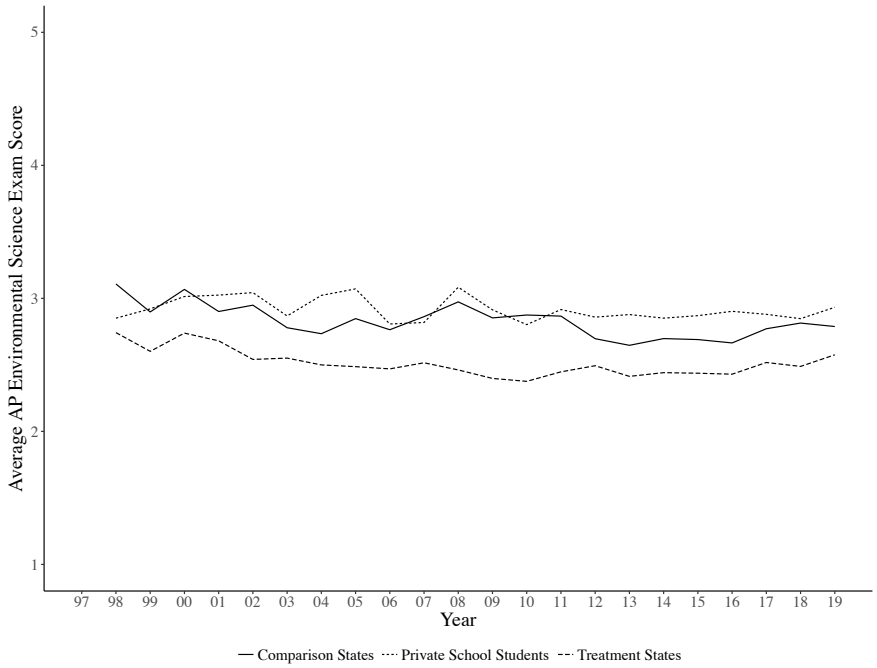


Figure 3

Average AP Psychology Exam Scores

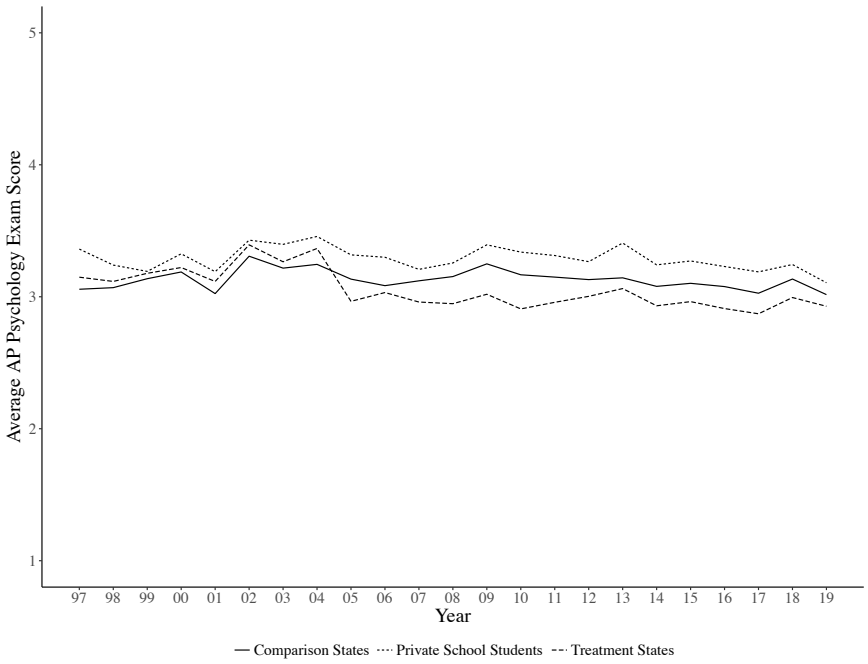
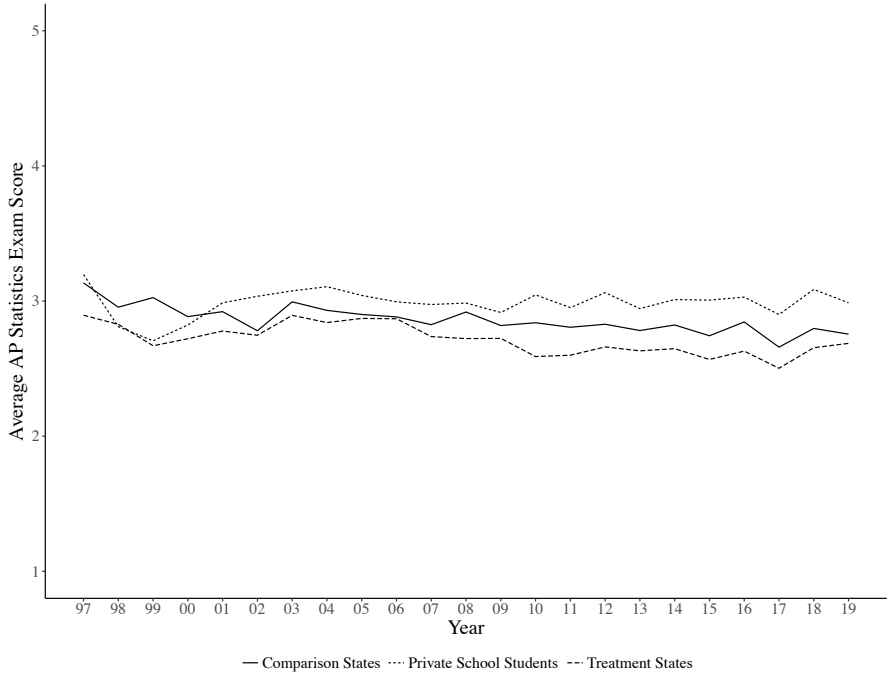


Figure 4

Average AP Statistics Exam Scores



Visual Inspection of Policy Effects

Visual inspection of the trend in AP exam performance centered on the point of intervention was used to preliminarily examine policy effects in policy-adopting states. Figures C1-C4 in Appendix C present the trend in AP exam performance in policy-adopting states centered on the point of intervention. In terms of AP exam performance, there appeared to be no noticeable shift in the trend immediately before or after the point of intervention in AP English Language and Composition and AP Psychology. There appears to be a very slight, delayed increase in AP Statistics after an initial decline in average exam scores immediately after policy introduction. There also appears to be a slight increase in exam performance immediately after the policy intervention in AP Environmental Science. AP Environmental Science exam performance continued to trend upward in the years after policy implementation, suggesting possible intercept and/or slope effects. Across every AP exam, participation in policy-adopting states remained steady with no noticeable shifts in the positive linear trend at or near the point of intervention (see Figures C5-C8 in Appendix C).

The trend in private school students in policy-adopting states was used as another nonequivalent comparison group. In AP English Language and Composition and AP Statistics, private school students outperformed public school students with the gap remaining relatively the same across time. Private school exam performance on AP Psychology was slightly better, but more volatile than public school students, though overall trends appeared largely similar. In AP Environmental Science, the trends between private and public school students remained parallel, except in the years following policy intervention where private school performance dipped slightly and public school

performance increased as described above.

Event Study

Baseline event study estimates, which include state and year fixed effects, showed no evidence of policy effects. Across all four AP exams, there was no evidence of policy effects in the academic years immediately after the introduction of AP accountability indicators. All estimates the year immediately after the policy intervention in each AP exam fall within the 95 percent confidence interval. In fact, there are only three statistically significant coefficients across all four AP exams. In AP Psychology, it appears average AP exam scores 11, 10, and 9 years prior to implementation were significantly greater than the year immediately before policy implementation, suggesting a slight decline in average scores across time. Results remained the same when covariates were added to the model (see Tables D1-D3 in Appendix D).

Figures 5-8 below graphically illustrate the event study estimates weighted by AP exam participation. The weighted models produced some notable differences in statistical significance across all AP exams. Scores for each AP exam appear to decline across time. Across all four AP exams, all coefficients six or more years prior to policy implementation were statistically different from the year prior to implementation. In AP Psychology, all coefficients five or more years prior to policy implementation were statistically different from the year prior to implementation. In AP Environmental Science, the coefficient for four or more years prior to policy implementation was also statistically different from the year prior to implementation. In AP Statistics, all coefficients three or more years prior to policy implementation were statistically different from the year prior to implementation. These event study findings suggest that pre-policy

Figure 5

Event Study Estimates for AP English Language and Composition: Weighted Model

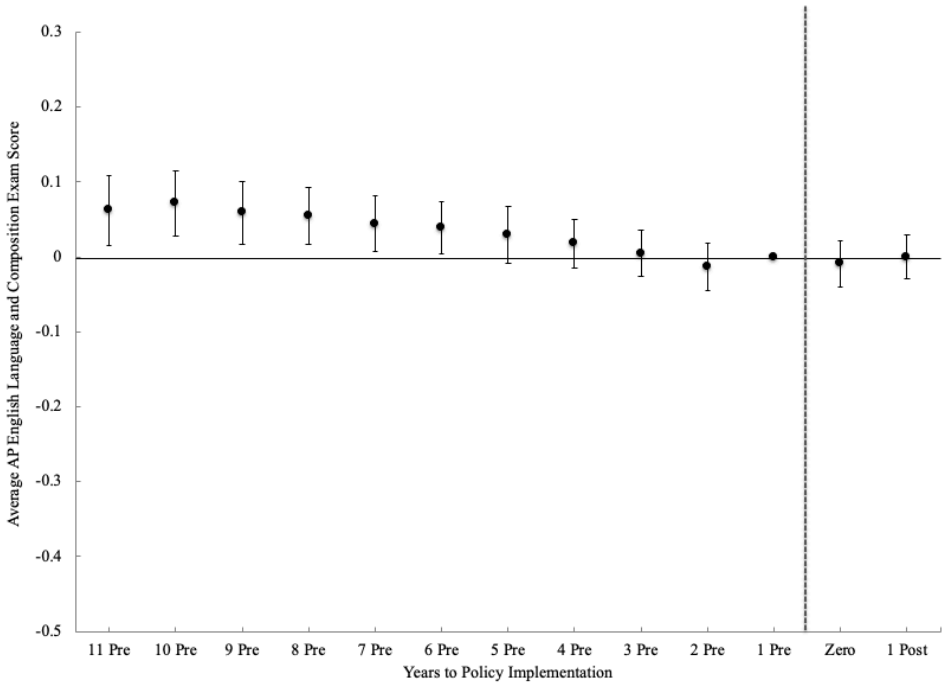


Figure 6

Event Study Estimates for AP Environmental Science: Weighted Model

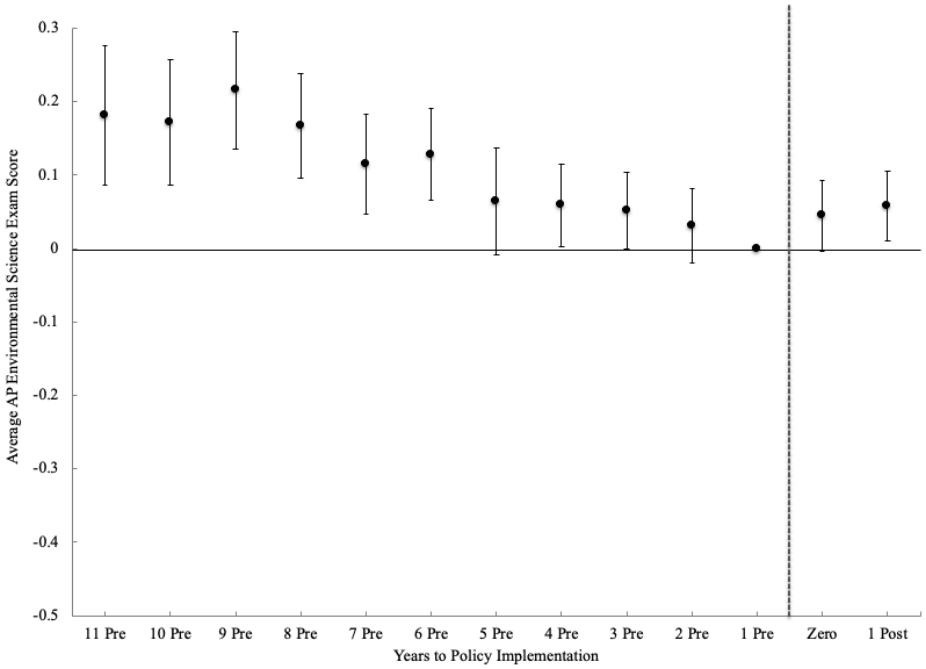


Figure 7

Event Study Estimates for AP Psychology: Weighted Model

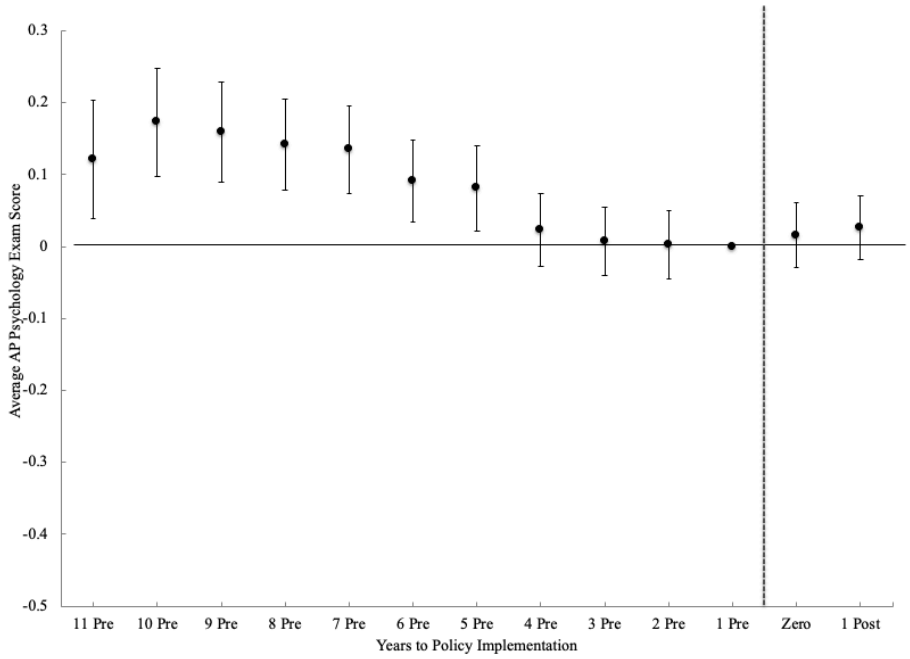
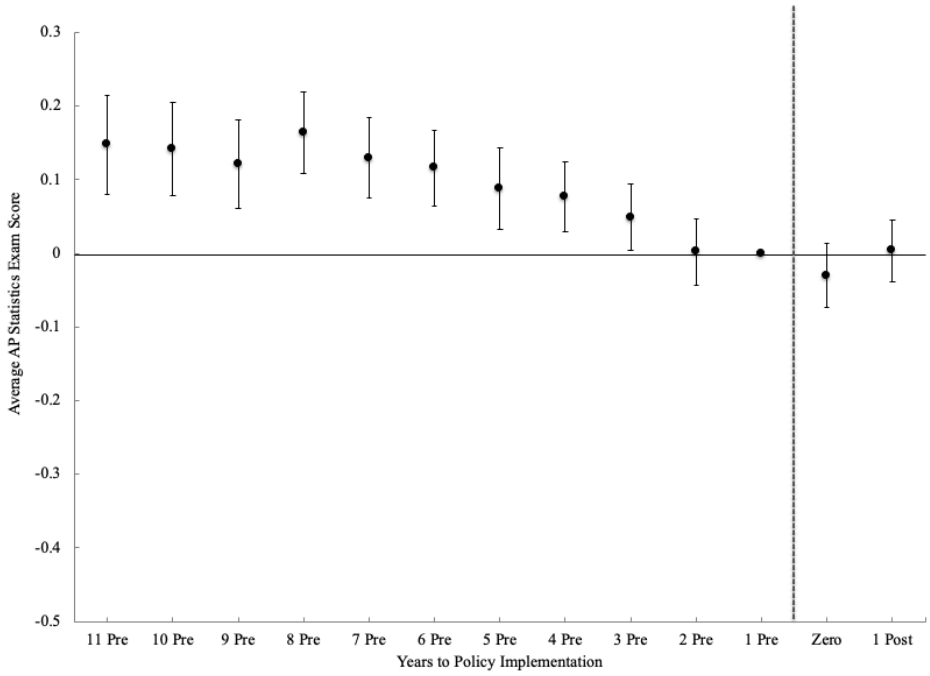


Figure 8

Event Study Estimates for AP Statistics: Weighted Model



trends in AP exam performance may bias weighted difference-in-differences estimates. When estimates are weighted, it appears comparison states that had not implemented AP exam accountability incentives (and policy-adopting states yet to do so) may not serve as a valid counterfactual due to the violation of the parallel trends assumption.

In terms of policy effects, weighted estimates produced no statistically significant changes in the years after policy implementation for AP English Language and Composition, AP Psychology, or AP Statistics. There is some evidence of delayed effects for AP Environmental Science. Weighted event study estimates showed a statistically significant increase in AP Environmental Science exam scores one year post-policy implementation ($b = 0.06$, $t = 2.42$, $SE = 0.02$, $p < 0.05$) in comparison to the year prior to the introduction of AP accountability metrics. These findings suggest policy-adopting states on average experienced a 0.06 increase in average AP Environmental Science exam scores one year post-policy implementation. In sum, event study estimates align well with the visual inspection of policy effects, showing possible delayed effects in AP Environmental Science, but not in any of the other AP exams.

Difference-in-Differences

The difference-in-differences models produced estimates that were largely consistent with findings from the event study analyses and the visual inspection of policy effects. No statistically significant policy effects were found across baseline difference-in-differences models for all four AP exams. Baseline difference-in-differences models included state and year fixed effects and standard errors clustered at the state level. These policy-specific findings held when covariates were added to the models and when models were weighted by AP exam participation. Below I discuss the results for each AP exam.

AP English Language and Composition (Difference-in-Differences)

As Table 4 below shows, difference-in-differences model results for AP English Language and Composition across all three models (baseline, covariate-adjusted, and weighted) provide no evidence of a change in intercept (i.e., PolicyImp) or slope (i.e., PolicyImp*TxTrend) after the introduction of AP exam accountability incentives. Additionally, it appears pre-policy intervention slopes (i.e., TxTrend) between the policy-adopting and comparison states did not differ statistically. None of the time-varying state characteristics in either the covariate-adjusted or weighted models produced statistically significant results. Finally, all AP English Language and Composition models explained

Table 4

Difference-in-Differences Estimates for AP English Language and Composition

Model	Baseline	Covariate-Adjusted	Weighted
Intercept	3.030 (0.048)*	3.361 (0.721)*	3.274 (0.629)*
TxTrend	-0.005 (0.005)	-0.002 (0.006)	-0.004 (0.005)
PolicyImp	0.023 (0.033)	-0.001 (0.029)	-0.011 (0.020)
PolicyImp*TxTrend	0.011 (0.008)	0.004 (0.009)	0.015 (0.009)
Grade 11 and 12		0.012 (0.005)	
SES		0.001 (0.002)	0.000 (0.001)
Asian		0.039 (0.015)	-0.029 (0.021)
Black		0.000 (0.009)	-0.002 (0.012)
Hispanic		-0.011 (0.010)	0.005 (0.011)
White		-0.009 (0.010)	0.001 (0.010)
Weights[a]			X
R-Squared	0.797	0.830	0.936

Note. *p<.05; All models include state and year fixed effects; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy-adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at 0 in policy-adopting states the year of policy implementation; [a] Models weighted by the number AP exam participants; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage of students eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of Grades 11 and 12 students divided by the total number of Grades 11 and 12 students in each respective group

80% or more of the variance in average state-level AP English Language and Composition scores across time, with the weighted model explaining the most variance ($R^2 = 0.94$, $F(79, 957) = 189.7$, $p < 0.001$).

AP Environmental Science (Difference-in-Differences)

Similar to the results for AP English Language and Composition, all difference-in-differences models for AP Environmental Science provided no evidence of policy effects or variance in pre-policy intervention slopes between the policy-adopting and comparison states (see Table 5 below). In the covariate-adjusted model, results showed as

Table 5

Difference-in-Differences Estimates for AP Environmental Science

Model	Baseline	Covariate-Adjusted	Weighted
Intercept	3.497 (0.124)*	1.153 (1.650)	3.087 (0.774)
TxTrend	-0.003 (0.011)	-0.004 (0.013)	-0.015 (0.004)
PolicyImp	0.097 (0.067)	0.083 (0.068)	0.042 (0.024)
PolicyImp*TxTrend	0.022 (0.015)	0.015 (0.019)	0.044 (0.008)
Grade 11 and 12		0.011 (0.010)	
SES		0.001 (0.003)	-0.001 (0.001)
Asian		0.132 (0.067)	-0.022 (0.028)
Black		-0.015 (0.016)	-0.027 (0.019)
Hispanic		0.021 (0.022)	0.014 (0.014)
White		0.020 (0.020)	0.007 (0.013)
Weights[a]			X
R-Squared	0.529	0.550	0.845

Note. *p<.05; All models include state and year fixed effects; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at 0 in treated states the year of policy implementation; [a] Models weighted by the number AP exam participants; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage of students eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of Grades 11 and 12 students divided by the total number of Grades 11 and 12 students in each respective group

the percentage of Asian students increases by one, AP Environmental Science exam performance increases by 0.03 points ($t = 6.48$, $SE = 0.005$, $p < 0.001$). However, the effect for Asian students becomes statistically nonsignificant in the model weighed by AP Environmental Science exam participation ($b = 0.01$, $t = 1.80$, $SE = 0.006$, $p = 0.07$).

The variance explained by the weighted AP Environmental Science model ($R^2 = 0.85$, $F(79, 891) = 67.97$, $p < 0.001$) was far greater than the baseline ($R^2 = 0.53$, $F(74, 972) = 16.88$, $p < 0.001$) and covariate-adjusted ($R^2 = 0.55$, $F(80, 890) = 15.82$, $p < 0.001$) models. The difference-in-differences models for AP Environmental Science explained less variance than the corresponding AP English Language and Composition models. The difference between the variance explained for the weighted AP English Language and Composition and AP Environmental Science models was less than the differences between the baseline and covariate-adjusted models

AP Psychology (Difference-in-Differences)

Difference-in-differences model results for AP Psychology exhibited similarities to the results for both AP English Language and Composition and AP Environmental Science (see Table 6 below). Similar to AP Environmental Science, results across all models for AP Psychology provided no evidence of policy effects or variance in pre-policy intervention slopes between the policy-adopting and comparison states. Additionally, none of the coefficients for the state characteristics reached the level of statistical significance in either the covariate-adjusted or weighted models. The amount of variance explained across all three models was less than, but closely resembled the models for AP English Language and Composition. Again, the weighted model ($R^2 = 0.88$, $F(79, 954) = 94.97$, $p < 0.001$) for AP Psychology explained the most variance.

Table 6

Difference-in-Differences Estimates for AP Psychology

	Baseline	Covariate-Adjusted	Weighted
Intercept	3.201 (0.059)*	3.185 (0.915)	1.708 (1.084)
TxTrend	-0.013 (0.007)	-0.016 (0.009)	-0.014 (0.008)
PolicyImp	-0.002 (0.048)	-0.018 (0.062)	0.002 (0.028)
PolicyImp*TxTrend	0.022 (0.015)	0.026 (0.014)	0.035 (0.014)
Grade 11 and 12		0.014 (0.006)	
SES		-0.001 (0.002)	-0.001 (0.002)
Asian		0.029 (0.027)	-0.012 (0.024)
Black		-0.007 (0.021)	0.020 (0.018)
Hispanic		-0.010 (0.012)	0.029 (0.018)
White		-0.004 (0.012)	0.024 (0.017)
Weights[a]			X
R-Squared	0.637	0.658	0.878

Note. *p<.05; All models include state and year fixed effects; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at 0 in treated states the year of policy implementation; a Models weighted by the number AP exam participants; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage of students eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of Grades 11 and 12 students divided by the total number of Grades 11 and 12 students in each respective group

AP Statistics (Difference-in-Differences)

Difference-in-differences model estimates for AP Statistics resembled the other three AP exams with no evidence of policy effects or variance in pre-policy intervention slopes (see Table 7 below). As with all previous AP exams, none of the covariates in the covariate-adjusted or weighted AP Statistics models reached the level of statistical significance. The amount of variance explained by the baseline ($R^2 = 0.56$, $F(75, 1073) = 20.53$, $p < 0.001$) and covariate-adjusted ($R^2 = 0.62$, $F(80, 938) = 21.47$, $p < 0.001$) models was slightly more than AP Environmental Science and slightly less than AP Psychology. The weighted model ($R^2 = 0.90$, $F(79, 939) = 113.9$, $p < 0.001$) for AP

Statistics explained slightly more variance than the corresponding models for both AP Environmental Science and AP Psychology.

Table 7

Difference-in-Differences Estimates for AP Statistics

	Baseline	Covariate-Adjusted	Weighted
Intercept	3.286 (0.095)*	2.403 (0.937)*	0.925 (0.963)
TxTrend	-0.001 (0.008)	-0.007 (0.010)	-0.012 (0.006)
PolicyImp	-0.054 (0.047)	-0.066 (0.046)	-0.027 (0.021)
PolicyImp*TxTrend	0.025 (0.014)	0.028 (0.015)	0.022 (0.013)
Grade 11 and 12		0.017 (0.007)	
SES		-0.001 (0.003)	-0.001 (0.002)
Asian		0.070 (0.033)	0.013 (0.026)
Black		0.000 (0.020)	0.009 (0.013)
Hispanic		0.000 (0.019)	0.031 (0.018)
White		0.002 (0.015)	0.029 (0.015)
Weights[a]			X
R-Squared	0.561	0.617	0.898

Note. *p<.05; All models include state and year fixed effects; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at 0 in treated states the year of policy implementation; a Models weighted by the number AP exam participants; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage of students eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of Grades 11 and 12 students divided by the total number of Grades 11 and 12 students in each respective group

Comparative Interrupted Time Series Models

CITS estimates departed from the pattern observed from the visual inspection of policy effects, event study estimates, and difference-in-differences findings with regard to the statistical significance of the policy-specific coefficients. Across all baseline models, only AP English Language and Composition did not produce a statistically significant policy effect. However, weighting estimates by AP exam participation resulted in statistically significant policy effects for AP English Language and Composition (both a

change in the post-policy intercept and slope). In AP Psychology, the baseline model produced a statistically significant, positive post-policy intercept and the weighted model produced a statistically significant, positive post-policy slope. AP Environmental Science and AP statistics both produced consistently statistically significant, positive post-policy slope results across all three models. Additionally, the AP Statistics covariate-adjusted and weighted models produced a negative, statistically significant post-policy intercept. The specification of the baseline CITS models, model building steps, and assumption testing are briefly discussed before a summary of main findings within each AP exam.

Intraclass coefficients (ICC) across all four exams exceeded 0.50, suggesting more than 50% of the variance in AP exam scores across time was between states, indicating hierarchical linear modeling was appropriate and necessary for modeling these data, where time is nested within individual policy-adopting and comparison states. Therefore, I used a series of deviance tests to specify the baseline CITS models, which included only an AP exam outcome, pre-policy slope, post-policy intercept, and post-policy slope. I used Singer and Willet's (2003) approach to deviance testing to determine if random effects were warranted for the time and policy-specific variables.

Table 8 below provides model comparisons and deviance tests for four different types of random effect models. With the exception of AP English Language and Composition, the best fitting model included random effects for the intercept, pre-policy slope (i.e., PreTxSlope), and post-policy intercept (i.e., PolicyImp). In these models, initial status in 1997, pre-policy trends, and immediate changes in AP exams scores were allowed to vary randomly between states. The best fitting model for AP English Language and Composition included only random effects for the model intercept. All

other parameters were fixed in the AP English Language and Composition models.

Table 8

Deviance Tests for Random Intercept, Pre-Policy Slope, and Policy Effects

	AP ELC		AP ES		AP Psychology		AP Statistics	
Model	Deviance	ChiSq	Deviance	ChiSq	Deviance	ChiSq	Deviance	ChiSq
A	870.63		894.35		140.20		504.63	
B	1243.89	373.27*	811.80	82.55*	35.11	105.09*	356.11	148.52*
C	1267.16	23.27*	791.75	20.06*	24.89	10.22*	347.99	8.13*
D	1280.41	13.25*	783.68	7.63	19.68	5.20	345.82	2.17

Note. AP ELC = AP English Language and Composition; AP ES = AP Environmental Science; Four additional models with different combinations of random effects were tested, each nested within one of the following models. Model A = Random intercept only; Model B = Model A + pre-policy slope; Model C = Model B + post-policy intercept; Model D = Model C + post-policy slope; With the exception of AP English Language & Composition, Model C produced the best fit of all models. Model A fit better than all other models for AP English Language & Composition

Tables E1 and E2 in Appendix E provide skewness and kurtosis statistics for each individual AP exam, separated by year. In general, results suggested a normal distribution without severe skewness, ranging from -0.19 to 1.32 in AP English Language and Composition, -1.14 to 0.54 in AP Environmental Science, -2.15 to -0.01 in AP Psychology, and -0.91 to 0.49 in AP Statistics. The vast majority of skewness statistics across all AP exams and years were between -1 and 1. Similarly, AP exam data did not indicate severe kurtosis, ranging from 0.56 to 3.75 in AP English Language and Composition, 0.73 to 2.89 in AP Environmental Science, -0.75 to 8.97 in AP Psychology, and -0.89 to 3.01 in AP Statistics. The exception was in 2001 (8.97), 2008 (7.73), and 2005 (7.03) in AP Psychology, suggesting a leptokurtic distribution in those years. However, as with the skewness values, the vast majority of statistics did not indicate severe kurtosis across AP exams and years.

Next, a series of assumption tests were conducted using the baseline CITS models. First, I examined the homogeneity of Level 1 variance using Levene's Test.

Table F1 in Appendix F shows that Levene's Test for each AP exam failed to reject the null hypothesis, suggesting variation in residuals across states are not statistically different. Figures F1-F4 in Appendix F present Q-Q plots of Level 1 residuals for each AP exam. The Q-Q plot for AP English Language and Composition was mostly linear, with a few visible outliers on both ends of the distribution. This same linear pattern emerged in AP Environmental Science, Psychology, and Statistics; however, the outliers on the tails of each distribution for these exams were more pronounced than in AP English Language and Composition.

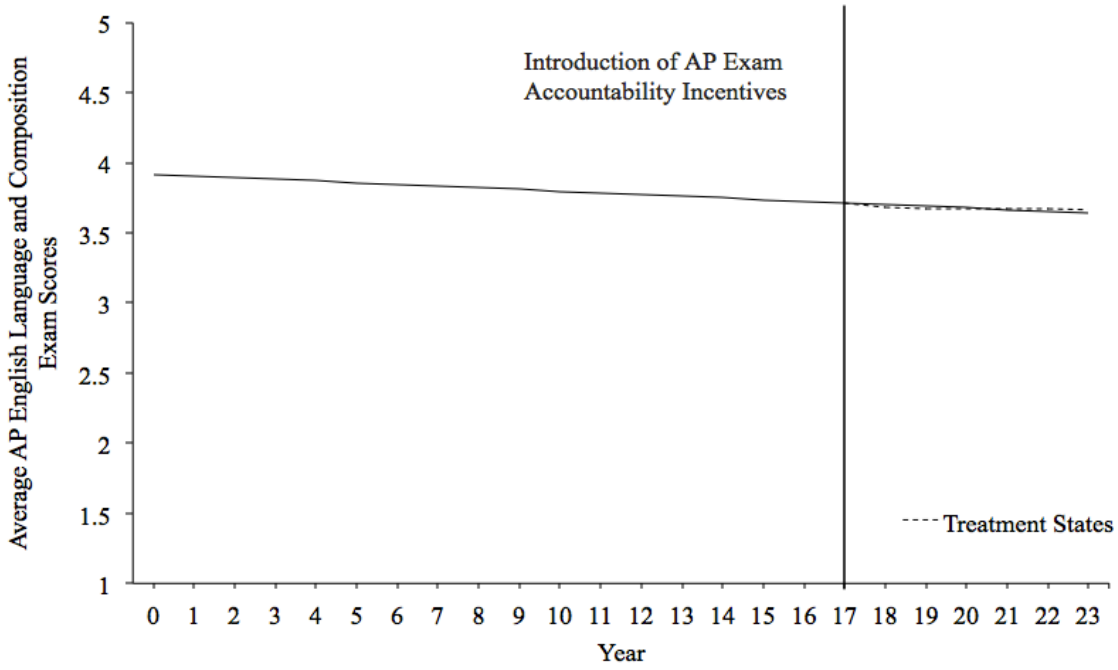
As discussed in more detail below, the weighted estimates produced half of the statistically significant policy effects in the CITS models. Therefore, Levene's Test was performed and Q-Q plots were constructed for the baseline weighted models (see Table F2 and Figures F5-F8 in Appendix F). In terms of Levene's Test, the null hypothesis for AP English Language and Composition and AP Environmental Science was rejected, but not for AP Psychology or AP Statistics. Additionally, Q-Q plots for weighted estimates across all AP exams produced more outliers than the baseline Q-Q plots. These outliers may threaten the assumption of normality for CITS models, producing a potential statistical conclusion validity threat to the overall CITS findings. Violating normality assumptions for Level 1 residuals in CITS models can bias standard errors, which in turn influences confidence intervals and subsequent hypothesis tests. However, these normality issues will not bias fixed effects estimates (Bernier et al., 2011; Raudenbush & Bryk, 2002). The results described above and presented in Appendix E suggest normality issues may be more of a concern in the weighted models.

AP English Language and Composition (CITS)

The baseline and covariate-adjusted CITS models for AP English Language and Composition provide no evidence of a change in intercept (i.e., PolicyImp) or slope (i.e., PostTxSlope) after the introduction of AP exam accountability incentives (see Table 9 below). However, the model weighted by AP English Language and Composition exam participation produced statistically significant effects for both PolicyImp and PostTxSlope. Figure 9 below graphically illustrates the post-policy intercept and slope effects for AP English Language and Composition using estimates from the weighted model. Immediately after the introduction of AP accountability incentives, AP English

Figure 9

Weighted CITS Model for AP English Language & Composition



Note. Weighted model policy effect estimates for AP English Language & Composition. The graph above illustrates the post-policy slope effect using Year 17 (i.e., 2014) as the example year of policy intervention. Note, this specific effect is only applicable to states that implemented AP exam accountability incentives in 2014. Graphically, states that implemented earlier will see more growth across time whereas those later will show less due to fewer post-policy years.

Language and Composition exam scores decreases by 0.031 points (t = -2.842, SE = 0.011, $p < 0.01$) and the post-policy slope increases by 0.009 exam points (t = 3.891, SE = 0.002, $p < 0.001$). Across all three models, the negative, statistically significant coefficient for the pre-policy slope (PreTxSlope) suggests AP English Language and Composition exhibited a linear decline in average exam scores (-0.01 across all three models) from 1997 to 2020.

Table 9

CITS Estimates for AP English Language and Composition

	Baseline	Covariate-Adjusted	Weighted
Intercept[a]	2.996 (0.038)*	3.449 (0.344)*	3.918 (0.421)*
PreTxSlope	-0.012 (0.001)*	-0.013 (0.002)*	-0.012 (0.002)*
PolicyImp	-0.011 (0.025)	-0.011 (0.026)	-0.031 (0.011)*
PostTxSlope	0.008 (0.006)	0.008 (0.006)	0.009 (0.002)*
Grade 11 and 12		0.003 (0.002)	
SES		0.000 (0.001)	0.000 (0.001)
Asian		-0.003 (0.005)	-0.020 (0.006)*
Black		-0.009 (0.004)*	-0.012 (0.004)*
Hispanic		-0.004 (0.004)	-0.004 (0.004)
White		-0.004 (0.003)	-0.009 (0.004)*
Weights[b]			X

Note. *p<.05; ICC = Intraclass coefficient; ICC = 0.779; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b]Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

In the covariate-adjusted model, the coefficient for the percentage of Black students achieved statistical significance. When the percentage of Black students increases by one, AP English Language and Composition exam performance decreases

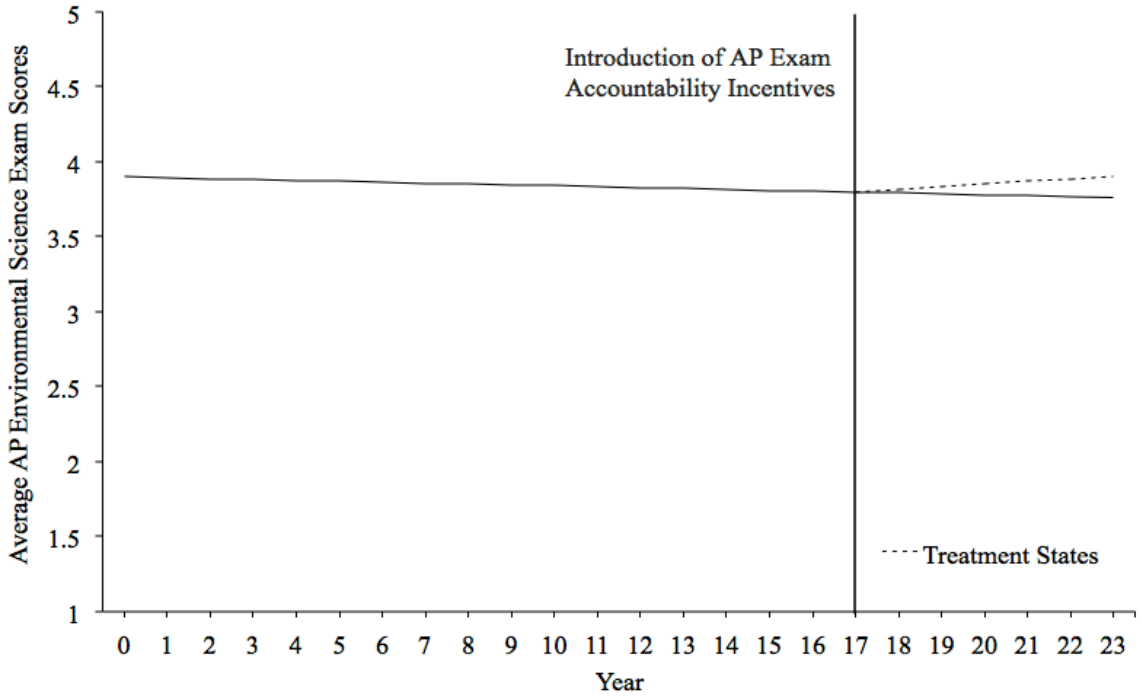
by 0.009 points (t = -2.427, SE = 0.004, $p < 0.05$). The statistical significance for Black students remained in the weighted model, while coefficients for Asian (b = -0.02, t = -3.501, SE = 0.006, $p < 0.001$) and White (b = -0.009, t = -2.171, SE = 0.004, $p < 0.05$) students also rose to the level of statistical significance.

AP Environmental Science (CITS)

Of all four AP exams, the CITS model estimates for AP Environmental Science produced results that were most consistent with the prior visual inspection of policy effects, event study estimates, and to a lesser extent, the difference-in-differences models (see Table 10 below). Figure 10 below graphically illustrates the post-policy slope policy

Figure 10

Weighted CITS Model for AP Environmental Science



Note. Weighted model policy effect estimates for AP Environmental Science. The graph above illustrates the post-policy slope effect using Year 17 (i.e., 2014) as the example year of policy intervention. Note, this specific effect is only applicable to states that implemented AP exam accountability incentives in 2014. Graphically, states that implemented earlier will see more growth across time whereas those later will show less due to fewer post-policy years.

effect for AP Environmental Science using estimates from the weighted model. For all three CITS models, the post-policy slope was positive and statistically significant, ranging from 0.024 (t = 5.992, SE = 0.004, $p < 0.001$) in the weighted model to 0.033 (t = 2.461, SE = 0.013, $p < 0.05$) in the covariate-adjusted model. Similar to AP English Language and Composition, the pre-policy slope exhibited a downward trend. However, the coefficient for PreTxSlope was only statistically significant in the baseline model (b = -0.014, t = -3.316, SE = 0.004, $p < 0.01$).

Table 10

CITS Estimates for AP Environmental Science

	Baseline	Covariate-Adjusted	Weighted
Intercept[a]	2.858 (0.078)*	2.335 (0.673)*	3.899 (0.634)*
PreTxSlope[a]	-0.014 (0.004)*	-0.001 (0.005)	-0.006 (0.004)
PolicyImp[a]	0.056 (0.058)	-0.015 (0.063)	0.012 (0.036)
PostTxSlope	0.031 (0.015)*	0.033 (0.013)*	0.024 (0.004)*
Grade 11 and 12		0.002 (0.002)	
SES		-0.008 (0.002)*	-0.004 (0.001)*
Asian		0.010 (0.008)	-0.014 (0.008)
Black		-0.003 (0.007)	-0.014 (0.006)*
Hispanic		0.004 (0.008)	-0.008 (0.007)
White		0.009 (0.007)	-0.010 (0.006)
Weights[b]			X

Note. *p<.05; ICC = Intraclass coefficient; ICC = 0.521; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b]Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

With respect to state characteristics, the coefficient for socioeconomic status was statistically significant in the covariate-adjusted model. As the percentage of students eligible for free- or reduced-priced lunch increases by 1, average AP Environmental

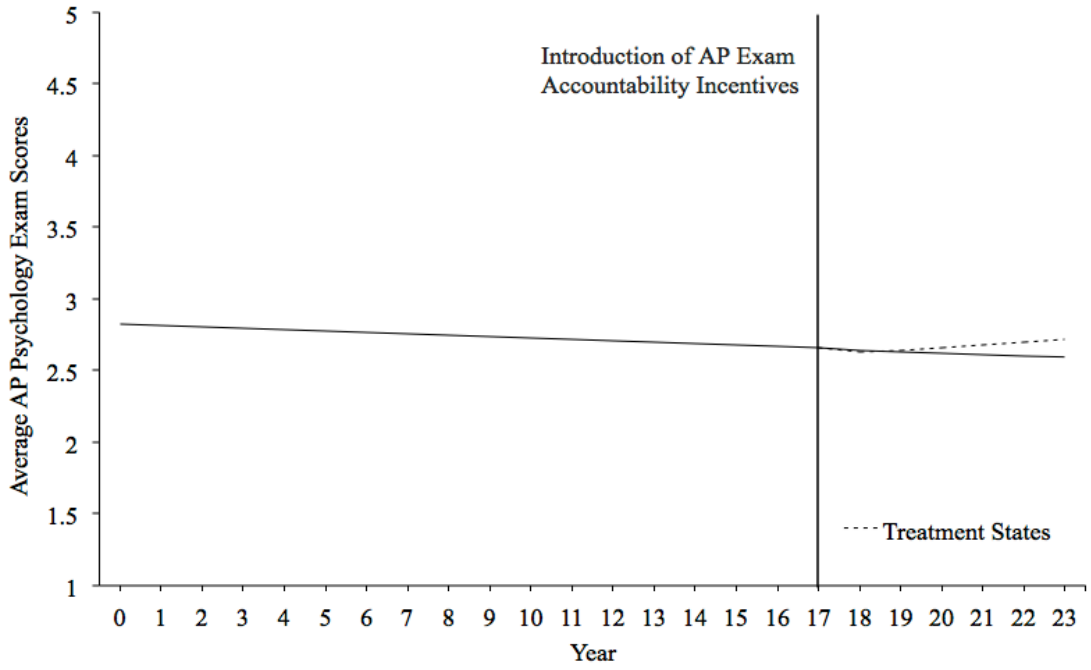
Science scores decreases by 0.008 points (t = -3.017, SE = 0.002, p < 0.01). The coefficient for socioeconomic status remained statistically significant in the weighted model (b = -0.004, t = -2.942, SE = 0.001, p < 0.01). The coefficient for the percentage of Black students also achieved statistical significance in the weighted model (b = -0.014, t = -2.19, SE = 0.006, p < 0.01). Table 11 below presents full results for the AP Environmental Science CITS models.

AP Psychology (CITS)

The results for AP Psychology showed statistically significant policy effects in the baseline and weighted models but not in the covariate-adjusted model (see Figure 11).

Figure 11

Weighted CITS Model for AP Psychology



Note. Weighted model policy effect estimates for AP Psychology. The graph above illustrates the post-policy slope effect using Year 17 (i.e., 2014) as the example year of policy intervention. Note, this specific effect is only applicable to states that implemented AP exam accountability incentives in 2014. Graphically, states that implemented earlier will see more growth across time whereas those later will show less due to fewer post-policy years.

Statistically significant post-policy intercept effects were found in the baseline model (see Table 11 below). Immediately after policy implementation, the post-policy intercept decreases by 0.095 exam points (t = -2.021, SE = 0.005, p = 0.05). The post-policy intercept was nonsignificant in the covariate-adjusted and weighted models while also decreasing in magnitude. The post-policy slope only reached the level of statistical significance in the weighted model. For every additional year after policy implementation, the post-policy slope increases by 0.028 exam points (t =8.224, SE = 0.003, $p < 0.001$). The null post-policy intercept effect in the weighted model is retained in Figure 11 due to its statistical significance in the baseline model.

Table 11

CITS Estimates for AP Psychology

	Baseline	Covariate-Adjusted	Weighted
Intercept[a]	3.158 (0.053)*	2.826 (0.558)*	2.825 (0.529)*
PreTxSlope[a]	-0.005 (0.002)*	-0.001 (0.003)	-0.010 (0.003)*
PolicyImp[a]	-0.095 (0.047)*	-0.082 (0.047)	-0.047 (0.031)
PostTxSlope	0.013 (0.010)	0.018 (0.010)	0.028 (0.003)*
Grade 11 and 12		0.006 (0.002)*	
SES		-0.002 (0.002)	0.001 (0.001)
Asian		0.005 (0.007)	0.005 (0.006)
Black		-0.006 (0.006)	-0.002 (0.005)
Hispanic		-0.004 (0.007)	0.003 (0.006)
White		0.006 (0.006)	0.005 (0.005)
Weights[b]			X

Note. *p<.05; ICC = Intraclass coefficient; ICC = 0.612; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b]Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

The pre-policy slope was only statistically significant in the baseline (b = -0.005, t = -2.273, SE = 0.002, $p < 0.05$) and weighted (b = -0.01, t = -3.471, SE = 0.003, $p < 0.001$) models, but not in the covariate-adjusted model (b = -0.001, t = -0.148, SE = 0.003, p = 0.882). The only statistically significant state characteristic, Grade 11 and 12 students, was found in the covariate-adjusted model. When the number of Grade 11 and 12 students increases by 10,000, AP Psychology exam performance increases by 0.006 points (t = 2.543, SE = 0.002, $p < 0.05$).

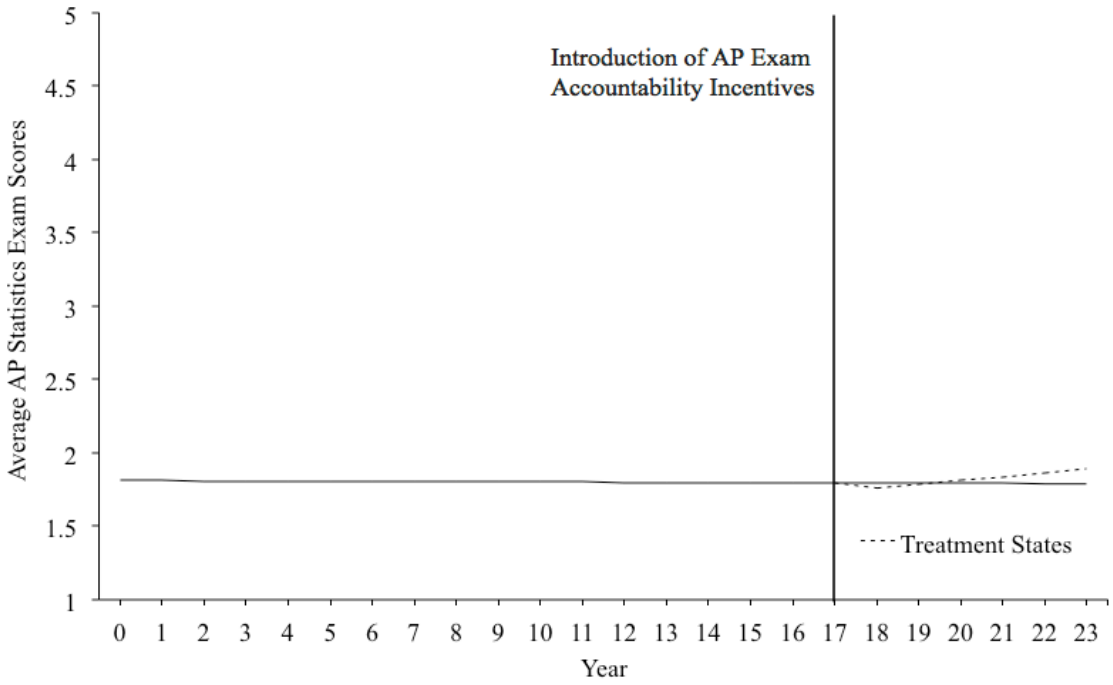
AP Statistics (CITS)

As Table 12 below shows, the baseline CITS models for AP Statistics produced a statistically significant post-policy slope (b = 0.023, t = 7.495, SE = 0.004, $p < 0.001$). Figure 12 below graphically illustrates the post-policy intercept and slope policy effect for AP Statistics using the weighted model. For every additional year after policy implementation, the post-policy slope for AP Statistics increases by 0.023 exam points. The post-policy slope retained its direction, statistical significance, and relative magnitude in the covariate-adjusted and weighted models. The post-policy intercept retained its direction, but became statistically significant in the covariate-adjusted model (b = -0.103, t = -2.214, SE = 0.046, $p < 0.05$). The statistically significant effect for the post-policy intercept remained in the weighted model, albeit with a smaller effect size (b = -0.06, t = -2.871, SE = 0.021, $p < 0.001$).

Similar to AP Environmental Science, the pre-policy slope was only statistically significant in the baseline model (b = -0.011, t = -3.673, SE = 0.003, $p < 0.001$). Two coefficients displayed statistical significance in the covariate-adjusted but not in the weighted model: Grade 11 and 12 students and socioeconomic status. As the number

Figure 12

Weighted CITS Model for AP Statistics



Note. Weighted model policy effect estimates for AP Statistics. The graph above illustrates the post-policy slope effect using Year 17 (i.e., 2014) as the example year of policy intervention. Note, this specific effect is only applicable to states that implemented AP exam accountability incentives in 2014. Graphically, states that implemented earlier will see more growth across time whereas those later will show less due to fewer post-policy years.

Grade 11 and 12 students increases by 10,000, AP Statistics exam performance increases by 0.007 points ($t = 2.765$, $SE = 0.002$, $p < 0.01$). As the percentage of students eligible for free- or reduced-priced lunch increases by 1, average AP Statistics scores decreases by 0.005 points ($t = -2.573$, $SE = 0.002$, $p < 0.05$). Finally, in the weighted model, as the percentage of White students increases by 1, AP Statistics exam performance increases by 0.013 points ($t = 2.438$, $SE = 0.006$, $p < 0.05$).

Table 12

CITS Estimates for AP Statistics

	Baseline	Covariate-Adjusted	Weighted
Intercept[a]	2.920 (0.054)*	2.857 (0.519)*	1.812 (0.547)*
PreTxSlope[a]	-0.011 (0.003)*	-0.001 (0.004)	-0.001 (0.003)
PolicyImp[a]	-0.081 (0.046)	-0.103 (0.046)*	-0.060 (0.021)*
PostTxSlope	0.023 (0.012)*	0.025 (0.011)*	0.027 (0.004)*
Grade 11 and 12		0.007 (0.002)*	
SES		-0.005 (0.002)*	-0.001 (0.001)
Asian		-0.005 (0.006)	0.005 (0.007)
Black		-0.007 (0.005)	0.002 (0.005)
Hispanic		-0.005 (0.006)	0.007 (0.006)
White		0.004 (0.005)	0.013 (0.006)*
Weights[b]			X

Note. *p<.05; ICC = Intraclass coefficient; ICC = 0.523; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b]Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Exploratory Policy Dosage Models

Exploratory policy dosage models were used to examine differential policy effects for two groups: (a) states (n = 9) that used AP outcome data as one of several indicators in high-stakes rating systems to identify and intervene in low-performing schools and (b) states (n = 10) that either simply reported AP outcome data or only used these data in the calculation of low-stakes school quality ratings. The states that used AP outcome data to generate low-stakes school quality ratings used a separate, subset of school performance indicators that did not include AP data in high-stakes ratings systems to identify and intervene in low-performing schools. The latter group of ten states, with theoretically weaker accountability incentives, served as the reference group in all policy dosage

analyses. The other nine states used AP outcome data for high stakes decisions about state mandated interventions on top of the public pressure generated from reporting and using AP data to calculate school quality rating.

Below, policy dosage relationships are reported by AP exam, focusing on differences from the main CITS analyses presented previously. The design of the policy dosage models allowed for four types of contrasts. The first and second contrasts were between the nine high-dosage policy-adopting states and all other states with regard to (a) initial status (i.e., model intercept) and the (b) pre-policy intervention slope. The third and fourth contrasts specifically examined differences between the two policy dosage groups with respect to an (a) immediate post-policy change in status and (b) a change in the post-policy intervention slope. Tables G1-G4 in Appendix G present policy dosage results for the baseline, covariate-adjusted, and weighted models.

AP English Language and Composition (Policy Dosage Models)

Only one difference emerged when the four separate policy dosage variables were added to the main CITS analyses for AP English Language & Composition. Whereas the variable for the percentage of Black students was negatively associated with growth in average exam scores in the main CITS covariate-adjusted analyses, it did not reach the level of statistical significance in the corresponding policy dosage model (b = -0.007, t = -1.90, SE = 0.004, p = 0.057). All policy effects were consistent in statistical significance, direction, and magnitude. Finally, no policy dosage policy relationships were found. The only dosage-related statistically significant finding suggests the nine high-dosage policy-adopting states had a statistically lower initial status than all other states, ranging from -0.278 (t = -3.126, SE = 0.089, $p < 0.01$) in the baseline to -0.428 (t = -3.884, SE = 0.11, p

< 0.001) in the weighted model.

AP Environmental Science (Policy Dosage Models)

Similar to AP English Language and Composition, the only notable difference between the main and policy dosage CITS models related to the percentage of Black students, which was statistically significant in the main weighted model but not the corresponding policy dosage model (b = -0.01, t = -1.705, SE = 0.006, p = 0.088). The policy-specific results for AP Environmental Science did not vary meaningfully from the main CITS to the policy dosage models. All findings were consistent in statistical significance, direction, and magnitude. Similar to AP English Language and Composition, the high-dosage policy-adopting states had a statistically lower initial status than all other states in the baseline (b = -0.504, t = -2.588, SE = 0.195, $p < 0.01$) and weighted (b = -0.39, t = -2.225, SE = 0.175, $p < 0.05$) models. No difference in the pre-policy intervention slope was observed between high-dosage states and all other states. The statistically significant coefficient for TxIdentify*Policy in the baseline and weighted models suggests that high-dosage was associated with a positive change in AP Environmental Science exam scores immediately after the introduction of AP accountability incentives. The effect size was 0.298 AP exam points (t = 2.192, SE = 0.138, $p < 0.05$) in the baseline model and 0.162 points (t = 2.187, SE = 0.074, $p < 0.05$) in the weighted model. There was no difference in the post-policy slope in any model.

AP Psychology (Policy Dosage Models)

The only differences between the main CITS and policy dosage models for AP Psychology occurred between the baseline and covariate-adjusted models. No differences were observed between the weighted models. The pre-policy slope (b = -0.005, t = -

1.891, *SE* = 0.003, *p* = 0.059) and intercept (*b* = -0.084, *t* = -1.41, *SE* = 0.060, *p* = 0.159) were statistically nonsignificant in the policy dosage but significant in the main baseline model. The only dosage-specific statistically significant finding revealed that high-dosage states had a higher initial status in average AP Psychology exam scores in comparison to all other states (*b* = 0.004, *t* = -2.524, *SE* = 0.142, *p* < .05).

AP Statistics (Policy Dosage Models)

The policy-specific results for AP Statistics varied from the main CITS to policy dosage models with respect to the post-policy intercept. While both coefficients for the post-policy intercept in the covariate-adjusted and weighted CITS models achieved statistical significance, neither did in the policy dosage model (covariate-adjusted: *b* = -0.096, *t* = -1.627, *SE* = 0.059, *p* = 0.104; weighted: *b* = 0.541, *t* = 0.935, *SE* = 0.579, *p* = 0.35). In terms of policy dosage effects, the only statistically significant finding suggested high-dosage states had a significantly lower initial status than all other states (*b* = -0.322, *t* = -2.384, *SE* = 0.135, *p* < 0.05). The statistical significance for TxIdentify was not retained in the covariate-adjusted (*b* = -0.154, *t* = -1.003, *SE* = 0.154, *p* = 0.316) or weighted (*b* = -0.174, *t* = -1.463, *SE* = 0.126, *p* = 0.487) models.

Sensitivity Analyses

Sensitivity analyses probed internal validity threats associated with the nonequivalent characteristics of the comparison group and unobservable state-level factors that may influence public school students rather than private school students. Baseline policy-adopting state and private student models were specified using the deviance test process described above (see Appendix H). For brevity, only policy-specific results are discussed below. Tables with full results for both the policy-adopting state and

private student models can be found in Appendix I and J, respectively.

Policy-Adopting State Only Models

Estimates from the main models that included all policy-adopting and comparison states were compared to estimates from models that included only observations for policy-adopting states. Any policy effects observed in the main models should remain consistent in the policy-adopting state only models, unless aspects of the comparison group composition bias estimates in the main models.

AP English Language and Composition (Policy-Adopting State). Event study and difference-in-differences estimates in the policy-adopting state model showed no policy effects, staying consistent with the main models. In the CITS models, statistically significant, positive post-policy slope effects were found in the policy-adopting state baseline ($b = 0.011$, $t = 2.23$, $SE = 0.005$, $p < 0.05$) and weighted ($b = 0.015$, $t = 4.774$, $SE = 0.003$, $p < 0.001$) CITS models. These results align with the main weighted models. However, the statistically significant post-policy slope effects in policy-adopting state baseline CITS model were not found in the main baseline model for AP English Language and Composition.

AP Environmental Science (Policy-Adopting State). Event study estimates in the weighted policy-adopting state model showed no policy effects whereas the main event study models showed delayed effects one year after policy implementation for AP Environmental Science. Similarly, no policy effects were observed across any of the main or policy-adopting state only difference-in-differences models. Statistically significant, positive post-policy slope effects were identified in the policy-adopting state covariate-adjusted ($b = 0.032$, $t = 2.039$, $SE = 0.016$, $p < 0.05$) and weighted ($b = 0.03$, $t = 3.14$, SE

= 0.01, $p < 0.01$) CITS models, aligning with results from the main covariate-adjusted and weighted CITS models for AP Environmental Science. However, unlike the main baseline model, no significant post-policy slope effects were found in the policy-adopting state baseline model ($b = 0.029$, $t = 1.833$, $SE = 0.016$, $p = 0.067$).

AP Psychology (Policy-Adopting State). Policy-specific estimates for the main and policy-adopting state models for AP Psychology were consistently nonsignificant with two exceptions. First, the negative, statistically significant post-policy intercept coefficient in the main baseline CITS model was nonsignificant in the corresponding policy-adopting state baseline model ($b = -0.017$, $t = 1.563$, $SE = 0.011$, $p = 0.118$). Similarly, the positive, statistically significant post-policy slope coefficient in the main weighted CITS model was nonsignificant in the corresponding policy-adopting state weighted model ($b = 0.017$, $t = 1.605$, $SE = 0.011$, $p = 0.108$).

AP Statistics (Policy-Adopting State). The main and policy-adopting state event study and difference-in-differences models produced no statistically significant policy effects. However, the post-policy slope was not statistically significant in the policy-adopting state baseline ($b = 0.025$, $t = 1.902$, $SE = 0.013$, $p = 0.057$) or covariate-adjusted ($b = 0.024$, $t = 1.842$, $SE = 0.013$, $p = 0.066$) models, whereas these coefficients were statistically significant in the main models. Similarly, whereas a negative, statistically significant post-policy intercept effect was observed in the main covariate-adjusted models, no such effect was found in the corresponding policy-adopting state covariate-adjusted model ($b = 0.090$, $t = 1.66$, $SE = 0.054$, $p = 0.10$).

Private School Student Models

Models that included private school instead of public school students were used to

probe unobservable state-level factors that may influence AP outcomes for public school students only. Observing consistent statistically significant effects in the main and private school student models would suggest that some event other than AP accountability incentives is responsible for the change in post-policy intercept or slope.

AP English Language and Composition (Private School Students). Similar to the policy-adopting state models for AP English Language and Composition, event study and difference-in-differences policy-specific estimates remained consistent between the main models and models that included only private school students. Additionally, no private school student CITS models produced a statistically significant policy effect, whereas positive effects were found for both the post-policy intercept and slope in the main weighted CITS model.

AP Environmental Science (Private School Students). Event study and difference-in-differences policy-specific estimates for AP Environmental Science produced no statistically significant policy effects, thus remaining consistent with the main models. Extensive positive policy effects were found in the main and policy-adopting state CITS models for AP Environmental Science, but not in the private school models. The only statistically significant effect was a negative coefficient for the post-policy intercept in the weighted private school CITS model (b = -0.088, t = 2.247, SE = 0.039, $p < 0.05$).

AP Psychology (Private School Students). The policy-specific estimates for the main and private school student AP Psychology event study and difference-in-differences models were identical with no evidence of policy effects across all models. As with the difference between the main and policy-adopting state models, the only difference

between the main and private school student model was the positive, statistically significant post-policy slope coefficient in the main weighted CITS model that did not appear in the private school student model.

AP Statistics (Private School Students). Consistent with the main model results, event study and difference-in-differences private school models showed no evidence of policy effects for AP Statistics. Additionally, no statistically significant policy effects were observed in the private school student CITS models in comparison to the negative effect for the post-policy intercept in the main covariate-adjusted CITS model and positive effect for the post-policy slope in the main baseline, covariate-adjusted, and weighted CITS models.

CHAPTER IV

DISCUSSION

This section begins with a summary of the key policy-specific findings, including different effect size interpretations. The main internal validity threats are then discussed, followed by a summary of this study's key limitations. Finally, I situate the contextualized findings within the research literatures on school accountability, the AP program, and time series data analysis. I conclude by discussing how the key findings are of practical significance to state policymakers and provide an overview of future directions for research on the ongoing effects of AP accountability incentives.

Key Policy-Specific Findings

This study examined how the introduction of AP exam accountability indicators influenced average state-level exam scores on four AP exams: English Language and Composition, Environmental Science, Psychology, and Statistics. To facilitate coherent pattern matching, this overarching research question was addressed using four separate types of analyses, two different contrasts, and two different ways of modeling policy dosage. The goal, from a causal inference perspective, is to achieve coherence across all analyses and AP exams in terms of the statistical significance, direction, and relative magnitude of the policy-specific findings. The sections below summarize the key policy-specific findings for each individual AP exam. Each section includes individual tables (13-16) for each AP exam to provide easier interpretation of the statistical significance of policy effects across analytical models.

AP English Language and Composition

The policy-specific findings for AP English Language and Composition stayed

consistent across nearly all analyses and models, with one notable exception (see Table 13 below). Whereas visual inspection of policy effects and all analytical models showed no evidence of a policy effect on average AP English Language and Composition exam scores, the weighted CITS model produced a negative, statistically significant post-policy intercept and positive post-policy slope. This same pattern emerged in the policy dosage and policy-adopting state only models, except that the post-policy intercept in the policy-adopting state only model did not rise to the level of statistical significance despite staying consistent in direction and relative magnitude. Additionally, no policy effects were found in the corresponding private school student model.

Table 13

Comparing the Policy Effect Statistical Significance: AP English Language and Composition

	Difference-in-Differences			CITS		
Policy Effect Type / Analytical Model	Baseline	Covariate -Adjusted	Weighted	Baseline	Covariate -Adjusted	Weighted
Post-Policy Intercept						
Main						
Policy-Adopting State						
Private School Students						
Post-Policy Slope						
Main						
Policy-Adopting State						
Private School Students						

Note. Null findings; Negative, statistically significant finding; Positive, statistically significant finding

Taken at face value, the statistically significant policy effects in the weighted CITS model suggest the introduction of AP accountability incentives had an immediate negative impact on average AP English Language and Composition exam scores in an average-sized policy-adopting state, but produced a long-term increase in the post-policy slope. These policy-specific findings, however, should be interpreted with caution for

three main reasons. First, the visual analyses of policy effects showed very little movement in either the post-policy intercept or slope. Second, none of the previous event study, difference-in-differences, or CITS models identified a statistically significant effect. Third, the weighted baseline models suggest normality issues may be biasing standard error estimates and subsequent hypothesis tests, particularly in AP English Language and Composition (see Appendix E). As a result, the weighted CITS estimates produce weak evidence of significant policy effects.

AP Environmental Science

Perhaps the most consistent findings within an individual AP exam were associated with Environmental Science, which demonstrated similar effects from visual inspection through the CITS analyses (see Table 14 below). The visual inspection of exam score trends suggested possible post-policy intercept and/or slope effects. Event study estimates initially aligned with the visual inspection, showing delayed, positive effects in the second year of policy implementation. Although the difference-in-differences model did not produce a statistically significant result, both the post-policy slope coefficients were positive and of a magnitude greater than those observed in the other three AP exams. All three CITS models (i.e., baseline, covariate-adjusted, weighted) produced a statistically significant, positive post-policy slope. Findings from the policy-adopting state only and private school student models bolstered these consistent policy-specific results. First, the positive, statistically significant coefficients for the post-policy slope in the policy-adopting state model (except for the baseline model, which maintained direction and relative magnitude) matched the main CITS models findings. No significant policy effects were observed in the private school student

models (outside of a negative post-policy intercept in the weighted model, which is counter to the positive effects found in the main models).

Table 14

Comparing the Policy Effect Statistical Significance: AP Environmental Science

	Difference-in-Differences			CITS		
Policy Effect Type Analytical Model	Baseline	Covariate -Adjusted	Weighted	Baseline	Covariate -Adjusted	Weighted
Post-Policy Intercept						
Main						
Policy-Adopting State						
Private School Students						
Post-Policy Slope						
Main						
Policy-Adopting State						
Private School Students						

Note. Null findings; Negative, statistically significant finding; Positive, statistically significant finding

This same post-policy result was reproduced in the policy dosage model. Additionally, the positive, statistically significant interaction between TxIdentify and PolicyImp suggests policy-adopting states that used AP exam data as one of multiple indicators in high-stakes rating systems designed to identify low-performing schools saw an immediate increase in average AP exam scores—compared to policy-adopting states that only used AP exam data as a simple, publicly reported data point or in low-stakes school quality ratings not tied to identifying low-performing schools.

AP Psychology

Results in AP Psychology mirrored those in AP English Language and Composition (see Table 15 below). Event study and difference-in-differences estimates in both exams aligned with the visual inspection of exam score trends, with all analyses suggesting no evidence of potential policy effects. As with AP English Language and Composition, CITS estimates in AP Psychology did not align with visual inspection,

event study, or difference-in-differences estimates. CITS estimates for AP Psychology identified a negative impact on the post-policy intercept (baseline model only) and a positive impact on the post-policy slope (weighted model only). The policy dosage model reproduced one of these significant results, the positive, post-policy slope in the weighted model. However, the policy-specific findings for AP Psychology should be interpreted with even more caution than those found in AP English Language and Composition. While policy-adopting state only estimates corroborated main CITS model results for AP English Language and Composition, they did not in AP Psychology. Put differently, the policy-specific findings from the main CITS models for AP Psychology were not reproduced in the policy-adopting state only models. One interpretation of these inconsistent findings is that some event other than AP accountability incentives is responsible for the post-policy slope increase in Psychology.

Table 15

Comparing the Policy Effect Statistical Significance: AP Psychology

	Difference-in-Differences			CITS		
Policy Effect Type Analytical Model	Baseline	Covariate -Adjusted	Weighted	Baseline	Covariate -Adjusted	Weighted
Post-Policy Intercept						
Main						
Policy-Adopting State						
Private School Students						
Post-Policy Slope						
Main						
Policy-Adopting State						
Private School Students						

Note. Null findings; Negative, statistically significant finding; Positive, statistically significant finding

AP Statistics

Visual inspection of policy effects in AP Statistics did not produce dramatic changes in average exam scores post-policy, although one could argue there is a slight,

delayed increase in the post-policy slope. That slight increase, however, was only statistically significant in the CITS models. All three CITS models produced a positive, statistically significant post-policy slope while the covariate-adjusted and weighted models both produced a statistically significant, negative post-policy intercept, a finding not anticipated based on the visual inspection. Similar to AP Environmental Science, the policy-adopting state and policy dosage models for AP Statistics supported the post-policy slope findings (see Table 16 below). However, the post-policy intercept finding was not reproduced in the policy-adopting state models. Finally, the private school student model for AP Statistics produced no policy effects. The consistency of results across the main, policy-adopting state, and private school models for AP Statistics is contrasted with AP Psychology, where no such alignment was found.

Table 16

Model Comparisons for Statistics

	Difference-in-Differences			CITS		
Policy Effect Type Analytical Model	Baseline	Covariate -Adjusted	Weighted	Baseline	Covariate -Adjusted	Weighted
Post-Policy Intercept						
Main						
Policy-Adopting State						
Private School Students						
Post-Policy Slope						
Main						
Policy-Adopting State						
Private School Students						

Note. Null findings; Negative, statistically significant finding; Positive, statistically significant finding

Summary of policy-specific findings

From a coherent pattern matching standpoint, policy-specific results were mixed with respect to statistical significance. Two courses—AP Environmental Science and AP Statistics—demonstrated consistent, positive post-policy slope effects across nearly all

analyses, while another course—AP English Language and Composition—produced null results except for the weighted CITS model. AP Psychology produced mixed results within the CITS models. However, results were consistent across AP exams and analyses with regards to the direction and magnitude of policy effects regardless of statistical significance. Taken together, differences in findings across outcomes suggest state policymakers may see a range of effects from introducing AP exam indicators. Policymakers may see a positive increase in the trend of average state-level AP exam performance, an immediate decrease in average AP exam performance, an immediate short-term decrease in AP exam performance that is offset by a larger long-term increase, or no effects at all. The consistency in the direction and magnitude of policy effects coupled with several instances of statistical significance deserves deeper scrutiny and further interpretation. The sections below provide additional context for the range of findings observed across AP exams.

Effect Size. Setting aside the instability of results across AP exams and potential internal validity threats (which will be discussed below), what is the practical significance of the policy-specific findings, found mostly in the CITS models? For example, the post-policy slope estimates in the CITS models for AP Environmental Science ranged from 0.02 to 0.03 exam points. This increase in slope can be considered both small, relative to the 1-5 scale for AP exams, and large, relative to the overall model slope of -0.006 exam points. Said differently, it would take roughly ten years to achieve an increase of 0.25 exam points from the introduction of accountability incentives, but without the incentives, average AP Environmental Science exam scores would likely continue a downward trend. The same can be said of AP Psychology and AP Statistics,

with one caveat. Although both of these exams saw a similar post-policy slope increase (ranging from 0.01 to 0.03 in AP Psychology and from 0.02 to 0.03 in AP Statistics) that moved each exam from a negative to positive trajectory across time, each model produced an immediate decrease in average AP exam scores following the introduction of AP exam accountability indicators, ranging from -0.05 to -0.10 exam points. The same cannot be said for AP English Language and Composition. If taken at face value, the CITS weighted estimates suggest states that introduce AP accountability incentives will see an immediate decrease in average AP English Language and Composition exam scores (-0.03) and the corresponding increase in the post-policy slope will not be large enough (0.009) to overcome the negative trend across time (-0.012). Thus, the practical effect is a one-time decrease in average AP English Language and Composition exam scores after the introduction of AP exam accountability incentives.

Another way to examine the magnitude of effects is by converting point estimates into standard deviation effect sizes using the average standard deviation across all years within individual AP exams. Table 17 below presents standard deviation effect sizes for the policy-specific variables. As Table 17 shows, the statistically significant post-policy slope estimates of 0.02 to 0.03 exam points in AP Environmental Science equate to five- (0.049) and seven-hundredths (0.068) of a standard deviation. The post-policy slope estimates ranging from 0.01 to 0.03 in AP Psychology equate to three- (0.027) and eight-hundredths (0.082) of a standard deviation. The post-policy slope estimates ranging from 0.02 to 0.03 in AP Statistics equate to four- (0.035) and eight-hundredths (0.076) of a standard deviation. The single post-policy slope estimate in English Language and Composition equates to three-hundredths (0.03) of a standard deviation. The largest

effect sizes were the negative post-policy intercept estimates. The post-policy intercept in AP English Language and Composition fell by a tenth (-0.104) of a standard deviation, by just over a tenth (-0.128) to approximately one quarter (-0.259) in AP Psychology, and by just over a tenth (-0.147) to one quarter (-0.253) in AP Statistics.

Table 17

Comparison of Standard Deviation Effect Sizes for Policy-Specific Variables

AP Exam	Method	Baseline	Covariate-adjusted	Weighted
		Post-Policy Intercept		
English Language and Composition	Diff-in-Diff	0.077	-0.003	-0.037
	CITS	-0.037	-0.037	-0.104
Environmental Science	Diff-in-Diff	0.199	0.170	0.086
	CITS	0.115	-0.031	0.025
Psychology	Diff-in-Diff	-0.005	-0.049	0.005
	CITS	-0.259	-0.224	-0.128
Statistics	Diff-in-Diff	-0.132	-0.162	-0.066
	CITS	-0.199	-0.253	-0.147
		Post-Policy Slope		
English Language and Composition	Diff-in-Diff	0.037	0.013	0.050
	CITS	0.027	0.027	0.030
Environmental Science	Diff-in-Diff	0.045	0.031	0.090
	CITS	0.064	0.068	0.049
Psychology	Diff-in-Diff	0.060	0.071	0.096
	CITS	0.035	0.049	0.076
Statistics	Diff-in-Diff	0.061	0.069	0.054
	CITS	0.056	0.061	0.066

Note. Negative, statistically significant finding; Positive, statistically significant finding

Table 17 also provides a comparison of the magnitude of effects between the difference-in-differences and the CITS models. The standard deviation effect sizes for the post-policy intercept exhibited more variance than the post-policy slope estimates. For

example, all of the statistically significant post-policy intercept estimates are negative and between a tenth to just over a quarter of a standard deviation. Effect sizes within AP exams and models (e.g., baseline) are roughly within a tenth of a standard deviation between difference-in-differences and CITS models, except in AP Psychology where the differences are greater. Post-policy slope estimates are uniformly positive and of similar magnitude across AP exams, methods, and models with more than two-thirds of estimates between a three and eight-hundredths of a standard deviation. In general, Table 17 shows that the difference-in-differences and CITS models produced similar results in terms of direction and standard deviation effect sizes, but not statistical significance. Finally, Table 17 also illustrates the consistency of direction—negative post-policy intercept and positive post-policy slope—of all the statistically significant findings.

The difference in annual exam participation across AP exams is evident in the change in standard deviations across time. The standard deviation in AP English Language and Composition hovers near 0.3 across all years, whereas the other three exams begin higher in 1997 before gradually moving toward 0.3, with AP Environmental Science starting the highest (SD = 0.77 in 1998), following by AP Statistics (SD = 0.63 in 1997), then AP Psychology (SD = 0.44 in 1997)—which directly corresponds to the average number of AP exam participants across all years. As a result of the gradual regression toward 0.3, when computing standard deviations using the five most recent years of data, the effect sizes presented above are larger by approximately 0.01 standard deviations, except for AP English Language and Composition where standard deviations across time remained relatively consistent.

If taking the CITS estimates at face value, it appears state policymakers will be

taking a risk, albeit a relatively small one, when introducing AP exam accountability incentives. Setting aside the opportunity costs associated with investing resources in creating and implementing AP exam accountability indicators, the findings from this study suggest the best case scenario would be a marginal increase in average AP exam scores of between three and eight-hundredths of a standard deviation, which is small but enough to place the state on a positive trajectory going into the future. The worst case scenario would be an immediate, marginal decrease in average AP exam scores. The other post-policy scenario entails a tradeoff of an immediate decrease of between a tenth and one-third standard deviation in average AP exam scores and a marginal increase in the trajectory of change across time of between three and eight-hundredths of a standard deviation, a large enough increase to place the state on a positive post-policy trajectory.

Weighting Effects. The weighted models accounted for more than half of the statistically significant policy effect findings, nearly all of which occurred in the CITS models. There are at least three potential explanations for why the weighted models produced a disproportionate amount of statistically significant policy effects. First, it may suggest averaged-sized states are the most likely to see effects from introducing AP accountability incentives. It also might suggest that in large states, with many more AP exam participants and more stable growth trajectories, AP accountability incentives may not be strong enough to produce observable effects.

Second, methodological issues, such as the normality of Level 1 residuals in the weighted models, could also help explain why the weighted models accounted for the disproportionate amount of statistically significant policy effect findings relative to other models. In fact, the standard errors for the policy-specific variables are noticeably smaller

in weighted models relative to the baseline and covariate-adjusted models (standard errors in the baseline and covariate-adjusted models within AP exam models are nearly identical). For example, the standard errors for the post-policy intercept in the weighted models are approximately one half the size of standard errors in the baseline and covariate-adjusted models, on average across AP exams. The standard errors for the post-policy slope are approximately one-third the size, on average.

Finally, two other weighting issues relate to substantial variation in AP exam participation among states and the number of implementation years in policy-adopting states. Appendix K presents individual trajectories in AP exam scores for each policy-adopting state (see Figures K1-K16) as well as average variation in AP exam scores, the number of implementation years for policy-adopting states, and individual state rank according to average AP exam participation (see Table K1). The combination of these descriptive data allows for the visual inspection of heterogeneous policy effects. First, states with more implementation years contribute more to post-policy estimates than states with fewer implementation years. The second type of weighting effect comes from actual AP exam participation weights in the weighted models, which generally corresponds to lower variance in AP exam scores across time within states. Therefore, these larger states with less variance are weighted more heavily in policy effect estimates.

Results in AP Environment Science and AP Statistics, both of which exhibited significant, positive post-policy slope effects, helps illustrate how the number of years of policy implementation in large states may be contributing to statistically significant policy effects in the weighted models. In AP Environmental Science, states such as Florida (Figure K5) and Georgia (Figure K6) exhibit clear upward trajectories in exam

performance post-policy. These states also have at least nine post-policy data points and are among the top ten in average AP exam participation. Therefore, the extra weight states like Florida and Georgia carry from both a large number of post-policy data points and low variance in AP exam scores across time may be driving some of the statistically significant policy effects observed, especially in the weighted models. Future research should further examine the effects of both AP exam participation weights as well as the effect of early policy adopters on the estimation of average policy effects.

State Characteristics. State characteristic data provided additional context to the key findings. For example, state characteristic data showed that policy-adopting and comparison states are different compositionally, with policy-adopting states having higher percentages of students eligible for free or reduced priced lunch, more Black and Hispanic students, and fewer Asian and White students. Policy-adopting states also spent fewer dollars per student than comparison states, on average. Additionally, policy-adopting states were larger, with an average of 145,679 and 49,891 AP exam participants in policy-adopting and comparison states in 2019, respectively. In fact, the six largest policy-adopting states (California, Texas, Florida, Illinois, Georgia, and Pennsylvania) accounted for nearly half (49.16%) of all public school AP exam participants in 2019. As addressed later in the discussion, the differences between policy-adopting and comparison states produce two avenues for future research, one exploring how state size relates to the decision to adopt AP exam accountability indicators, and the other examining different comparison group configurations (Hallberg et al., 2018).

Although not the focus of this study, the models that included covariates produced mixed and inconsistent results with respect to statistical significance and a general lack of

significance overall (less than 6% of all coefficients in all main difference-in-differences and CITS models). Aligning with past research showing students who are economically disadvantaged and students of color score lower on AP exams than their peers on average (Kolluri, 2018), the findings from this study show, in some but not all models, that as the percentage of students eligible for free and reduced lunch and Black students increases within a state, average AP exam scores decrease. Also aligning with past research, as the percentage of White students increases within a state, average AP exam scores increase. Yet, contrary to past research, the findings presented in this study suggest as the percentage of Asian students increases within a state, average AP exam scores decrease. Though it should be noted this finding was observed in only one model (the weighted CITS model in English Language and Composition), and the variable for Asian students was harmonized to include students who identify as Native Hawaiian or Other Pacific Islander, thus creating a more heterogeneous overall subgroup of students. Finally, the results from this study suggest increases in size may be associated with an increase in average AP exam scores. These results should be interpreted cautiously given the limited amount of research on the association between changes in state size and average AP exam scores. Future scholarship should examine in more detail how changes in state and school demographics relate to trends in AP exam scores.

Internal Validity

A host of potential internal validity threats make it challenging to take the policy-specific findings at face value. Probing internal validity threats centers on identifying factors that may influence inferences about cause and effect relationships. Hallberg et al. (2018) highlight three potential internal validity threats that often plague short interrupted

time series designs similar to the one employed in this study: history, selection, and instrumentation. History threatens internal validity when a separate event (e.g., policy, program) occurs simultaneously with the intervention being analyzed. In the context of AP exam accountability incentives, history threats include any event that influences the state-level AP outcome variables being analyzed. An example of a historical threat would be a new College Board policy that provides students with additional time to complete an AP exam. If this College Board policy is implemented in the same year a state's AP exam indicator is introduced, and if this policy leads to a significant increase in AP exam scores, a positive effect associated with school accountability incentives could be confounded with the change in College Board policy. This threat would theoretically impact all AP exam participants, regardless if they reside in an policy-adopting or comparison state, or if they are a public or private school student.

One way to probe the internal validity threat of history is to create a separate model with a nonequivalent comparison group that theoretically should not be influenced by the policy intervention; private school students in policy-adopting states in the case of this study. The vast majority of policy-specific coefficients across all private school student models produced null findings. Since the private school models had the same number of Level 2 units, small effects that are undetectable rather than statistical power issues are likely producing the null results. Although only one piece of evidence, these results do not suggest history events that influenced all AP exam participants generated the policy-specific findings observed in the main policy-adopting and comparison state models. This is not entirely surprising considering most threats to the validity of the findings presented in this study come from federal and state policies, many of which are

specific to public schools and students.

Local history threats, such as state-level AP policies, occur when "one group experiences a set of unique events that the other does not" (Shadish et al., 2002, p. 182). These state-level policies must be introduced in the same year as AP exam accountability incentives are adopted to threaten internal validity. The use of nonequivalent comparison groups and switching replications helped address issues associated with history threats that influence all policy-adopting and comparison states, to some extent. Unfortunately, local history is more difficult to combat with additional research design elements. What data that does exist suggests policy-adopting states are more active in the AP policy arena than comparison states, on average. A 2016 report from the Education Commission of the States provides state profiles centered around ten different AP policies. When excluding accountability policies, policy-adopting states had a higher average number of AP policies (3.68) than comparison states (2.72). The average among the six policy-adopting states that accounted for nearly half of all AP exam participants in 2019 was even higher (4.33). These figures suggest local history threats could be at play. However, it is beyond the scope of this study to explore local history threats in sufficient detail. The discussion below on how the findings are situated in AP literature further explores what bearing these local history threats have on the findings presented in this study and how these threats can be leveraged to improve future research.

Selection is another internal validity threat that can impact short interrupted time series designs that use aggregated student-level data. Selection becomes an issue if the group of students pre-policy is meaningfully different than the group post-policy (Hallberg et al., 2018). Although the demographic composition of students within each

state does differ from year to year, it does not appear these differences were so extreme that they influenced the AP outcomes of interest. The relative lack of significance of time-varying covariates suggests changes in state characteristics do not exert much influence on changes in state-level AP exam scores across time. Furthermore, the amount of variance explained from the baseline to the covariate-adjusted difference-in-differences models was negligible in most cases.

Finally, instrumentation threatens internal validity when changes to the way the outcome variable is calculated occur at or near the introduction of the policy intervention. All publicly available information suggests none of the four AP exams selected for empirical analysis have undergone course or exam redesigns. It is possible, though, that the College Board did not publicize small revisions to these AP courses and exams. Both nonequivalent comparison groups (i.e., comparison states, private school students in policy-adopting states) and switching replications helped address potentially undetectable instrumentation issues. At the very least, similar to exploring history effects, the null findings in the private school student model did not provide evidence that instrumentation threatens the validity of the main model policy-specific results. Regardless, this internal validity threat is impossible to fully rule out without access to College Board psychometric information on the four AP exams examined in this study.

Limitations

Internal validity threats, which often can only be partially probed, can be considered a type of limitation. Three other limitations are worth noting. First, the staggered implementation of AP exam accountability indicators across states produced both opportunities and challenges from a methodological standpoint. On the one hand,

staggered implementation creates an opportunity to strengthen causal inferences if a coherent pattern of policy effects is found. For example, the staggered implementation helps account for single event validity threats that impact all states. On the other hand, varying policy implementation time points across states poses significant methodological challenges. Although this study attempted to be as precise as possible with the exact implementation date within individual states using online search procedures, it is difficult to determine when schools first became aware of new accountability metrics without speaking directly to those involved in experiencing the policy intervention (e.g., school administrators, teachers). Additionally, even within the year of policy implementation, states differed in the time of year they announced the introduction of new AP exam indicators (e.g., spring, fall). However, the decision rule described in the Methods section above was designed to ensure schools had at least one entire semester to react to the introduction of new indicators, and often schools had much more time to respond.

Second, and related, the list of other policies that could interact with the effect of AP accountability incentives is long. Some of these policies are discussed further below, but at a minimum, it is reasonable to assume that other state-level AP policies (Education Commission of the States, 2006; 2016) exert some influence on AP accountability incentives. For example, could the existence of complimentary AP policies strengthen the overall effect of accountability incentives on state-level average AP exam scores? Do certain combinations of AP policies produce the ideal mix for improving AP exam participation and performance? Similarly, several other types of policies, such as those focused on promoting postsecondary readiness (e.g., college and career readiness standards) or others focused on improving teacher capacity (e.g., teacher evaluation

systems), may also influence the effects of accountability incentives. As discussed below, future research focused on AP accountability incentives should include document analyses and interviews with state officials in an effort to further isolate the effect of AP accountability incentives and to estimate the influence of other policies on AP outcomes.

Finally, descriptive data and contextual information demonstrated that policy-adopting and comparison states are different demographically and geographically. Policy-adopting and comparison states also differed in meaningful ways with regard to AP exam performance and participation, with policy-adopting states scoring lower on exams but having higher participation, on average. The use of time-varying state characteristic data in the covariate-adjusted and weighted models helped account for some of these differences, but not all of them, and certainly not for unobserved state differences. Ultimately, the observed and unobserved differences may have biased estimates if they relate to the reason policy-adopting states introduced AP exam data or if time-varying confounds impact policy-adopting and comparison states differently. As described below, future research should consider alternative comparison group configurations and further explore differences between policy-adopting and comparison states.

Implications

The following sections situate the contextualized findings from this study in literature on school accountability, AP, and methodological research focused on different approaches to modeling time series data in policy analyses.

School Accountability

This study represents perhaps the first attempt to link school accountability to AP exam outcomes. Nearly all of the school accountability literature examining academic

outcomes has focused on standardized test scores within states or NAEP scores across states. Typical effect sizes found from NAEP studies are estimated in the one-third (up to 0.34) standard deviation range (Dee & Jacob, 2011; Wong et al., 2015). As Figlio and Loeb (2011) found, most "other estimated effects range from no effect to up to 0.20 standard deviations" (p. 412). The mixed findings from this study follow the general pattern of modest, positive effect sizes found across school accountability research. Although relatively small, they are not far off from estimates of teacher effectiveness in the literature; Hanushek and Rivkin (2010) estimate teachers add about 0.13 and 0.17 standard deviations of value to student achievement, on average. However, only AP Environmental Science, and AP Statistics to a lesser extent, produced consistent, positive effects across analyses. These positive effects, although promising, should be viewed cautiously given the potential internal validity threats and limitations described above. Furthermore, the negative post-policy intercept estimates also indicate positive post-policy slope effects may not be large enough to offset the immediate, decrease in average AP exam scores post-policy (i.e., negative post-policy intercept change).

Although the mixed findings in this study did not produce the type of coherent pattern matching needed to make strong causal inferences, there are many unanswered questions that when answered, will produce a more complete picture of the effect of AP exam accountability incentives. For example, one question that remains is why policy-adopting states seemed to experience a decline in AP exam scores immediately after the introduction of accountability incentives, which was followed by an increase in the post-policy slope. This phenomenon was found across all four AP exams.

In speculating about possible causes of this phenomenon, it is important to note

that policy-adopting states generally required schools to only report on average AP exam scores *across all students*. Only four states report average AP exam accountability data for subgroups of students. Past research shows that average AP exam scores decrease as participation increases (Kolluri, 2018). Although trend lines do not suggest increases in AP exam participation at the point of intervention for policy-adopting states, future research should carefully examine how this policy reform influenced exam participation rates for all students and subgroups of students. On the other hand, the increase in the post-policy slope could point to the opposite effect. That is, school leaders may have initially responded to AP accountability incentives by expanding the number of test-takers, but after seeing a corresponding decrease in average exam scores which hurt the school's accountability rating, these same school leaders may have employed gatekeeping tactics to ensure only high achieving students sat for AP exams. NCLB showed this is a distinct possibility. Research on NCLB found some schools reacted to accountability incentives by systematically reclassifying low-performing students as students with special needs (e.g., Booher-Jennings, 2005; Cullen & Reback, 2006; Figlio & Getzler, 2007; Jacob, 2005), encouraging absences (Cullen & Reback, 2006), or disproportionate disciplinary practices (Figlio, 2006). These exclusionary tactics all had one common effect. Each reduced the number of students who took the standardized tests used for school accountability purposes. Therefore, future research should carefully examine if AP accountability incentives differentially impacts subsequent exam participation and performance for all available subgroups of students.

Future research should also examine two unique, closely related aspects of AP accountability incentives. As described above, AP exams contain higher stakes than

traditional state standardized tests and NAEP, the two outcome measures studied most often in school accountability research. How does the existence of high stakes for students interact with accountability polices, and how is that relationship different than when low-stakes assessments are used for accountability purposes? The stakes for students closely relates to the precise score they receive on an AP exam, with a score of 3 or higher being the minimum required to be eligible for college credit. With more and more postsecondary institutions requiring an AP exam score of 4 or 5 for college credit, it is reasonable to ask what practical effect accountability incentives have even if average scores increase. Would states see greater economic benefits (in the form of more tuition savings for students) if schools were incentivized to have more students score a 4 or higher on AP exams? This type of question points to the need for a cost-benefit analysis that compares AP accountability incentives to alternative policy approaches.

This study also explored how policy dosage may influence the effects of AP exam accountability incentives. Specifically, exploratory CITS analyses grouped policy-adopting states into two categories, the nine states that tied AP exam data to high-stakes ratings used to identify low-performing schools subjected to state mandated sanctions and the ten states that did not. The ten states that did not use AP exam data for high stakes ratings either used AP exam data as a simple, publicly reported data point or in the calculation of low-stakes school quality ratings. The hypothesis being tested is that using AP exam data as one of multiple measures in high-stakes rating systems is a stronger incentive than using AP exam data only as a publicly reported data point or in the calculation of low-stakes school quality rating. Only the policy dosage analyses in AP Environmental Science seem to provide evidence that confirms this exploratory

hypothesis. The policy dosage findings specific to AP Environmental Science align with past research showing that stronger accountability sanctions are associated with larger effects (Dee & Jacob, 2011; Hanushek & Raymond, 2005; Wong et al., 2015).

In AP Environmental Science, policy-adopting states that used AP exam data as one of multiple measures in high-stakes ratings saw an increase of approximately five- to seven-hundredths of a standard deviation in average exam scores immediately after policy implementation. Interestingly, the post-policy intercept in the main CITS models for AP Environmental Science was negative and was not statistically significant. Moreover, although no other statistically significant dosage effects were found in the three other AP exams, the positive coefficient for TxIdentify*PolicyImp across all models and courses suggests future research should further explore the influence of policy dosage on the effect of AP exam accountability incentives. This positive effect suggests policy-adopting states that attach the strongest accountability consequences to AP exam data may see stronger immediate effects than policy-adopting states with lesser consequences. It is also possible the relatively small sample size of policy-adopting states in each policy dosage group may be responsible for the largely null findings if there is not enough statistical power to detect small policy effects.

Of course, policy dosage can be conceptualized many different ways. Researchers vary widely in the how they model accountability consequences with each taking a slightly different approach, both conceptually and analytically (Dee & Jacob, 2011; Hanushek & Raymond, 2005; Lee & Wong, 2004; Wong et al., 2015). One commonality is that these researchers all model accountability consequences at the system-level. New era accountability systems that include a more expansive set of school performance

indicators necessitates a different approach. The reason is two-fold. First, some states use two sets of school quality calculations, one that results in low-stakes school ratings and another that uses a smaller subset of indicators in high-stakes rating systems to identify low-performing schools that receive state mandated interventions (Polikoff et al., 2014). AP exam indicators fall into both categories depending on the policy-adopting state. Therefore, the conceptualization of policy dosage with respect to AP exam indicators is more nuanced than for state standardized tests, which historically compassed nearly the entirely of states' school accountability systems.

The list of policy dosage considerations for modeling the strength of AP exam accountability incentives includes but is not limited to the following: (a) explicit vs. implicit incentives, (b) the weight placed on the indicator in school quality ratings, (c) the construction of the indicator itself, (d) the type of interventions or consequences attached to AP exam performance, and (e) the effect of other AP or college readiness policies. First, Figlio and Ladd (2015) assert there are two types of accountability incentives, explicit and implicit. Explicit incentives can be positive (e.g., increased school resources) or negative (e.g., state takeover, mandated restructuring, or closing), with varying degrees of strength. Research also shows implicit accountability incentives (e.g., public pressure) can have effects on schools (Figlio & Kenny, 2009). The policy intervention in all 19 policy-adopting states carry some level of implicit incentives via the public pressure generated from posting school-level AP exam data publicly, but as discussed above in the context of policy dosage, these states vary widely with respect to explicit incentives.

Policy-adopting states also varied considerably in how much weight was placed on AP exam indicators in school quality ratings and the design of the indicator itself. For

example, states that use AP exam data for low- or high-stakes school ratings either use standalone AP indicators or use AP data as one measure within a composite indicator. For example, AP exam data accounts for one of eight standalone indicators used in Illinois to generate a summative designation of an Exemplary, Commendable, Targeted, or Comprehensive school. In New Mexico, AP exam scores are one of 11 ways students can earn points on the College and Career Readiness indicator for their school, which accounts for 15% of a school's A-F grade. These different indicator designs can be separated into compensatory or complementary categories (Brookhart, 2009). Standalone AP indicators within a multiple measure system can be thought of as compensatory since performance on any one of the indicators can offset poor performance on another—but AP performance matters one way or another. For complementary indicators, on the other hand, performance on one indicator among many suffices. In New Mexico, students could earn points for their school with AP exam scores, but if they earned points through any one of the other ten options, their AP exam score performance would not matter. There is also significant variation within these broad categorizations across states. The point is, no research has examined the effects, tradeoffs, and unintended consequences of each of the different approaches states use to create AP exam indicators. There is also no available research on the potential unintended consequences of using AP exam indicators that measure status rather than growth. Do AP status indicators result in schools and teachers preforming educational triage in the same way status indicators for state standardized tests have in the past (e.g., Lauen & Gaddis, 2016)? Future research should consider these design nuances and the subsequent impact on policy dosage.

Additionally, seven of the 19 policy-adopting states (see Table 1) included AP

exam participation data as part of the state's overall AP accountability indicator. Questions remain about how this type of combination indicator influences subsequent AP exam participation and exam performance. Does the inclusion of AP participation guard against schools potentially encouraging only high-performing students from taking AP exams in an effort to increase the percentage of students who score a 3 or higher? Do combination indicators create tensions in schools if efforts to increase access result in a decrease in the percentage of students who score a 3 or higher, as at least one study found (Rowland & Shircliffe, 2016)? Future research should explore these questions and others related to the types of incentives created by combination indicators and the subsequent effect on both AP exam participation and performance.

Adding to the complexity are differences in what interventions policy-adopting states employ in low-performing schools that are identified with AP exam data, among other indicators. Reviews of ESEA waivers and ESSA state plans show that states are generally unclear on the specific ways they will support low-performing schools (Aldeman, 2017a, 2017b; Polikoff et al., 2014). The brief document analysis performed for this study backs up these claims. Preliminary evidence from ESEA waivers and ESSA state plans in policy-adopting states do not shed light on the specific accountability mechanisms by which states will improve AP outcomes in low-performing schools. Each policy-adopting state's ESEA waiver and ESSA state plan was reviewed for the number of times "Advanced Placement" or "AP" was mentioned. The average number of mentions for ESEA flexibility waivers was 16.11 with a range of 0-43 mentions. The average for ESSA was 11.21 with a range of 2-28 mentions. These mentions were coded using an inductive approach. The major themes that emerged in both ESEA waivers and

ESSA state plans can be grouped into six categories:

- stating state goals specific to AP,
- describing the policy environment supporting AP,
- touting positive AP participation and performance outcomes,
- providing the rationale for AP accountability indicators,
- detailing the design of AP accountability indicators, and
- specifying what accountability mechanisms will improve AP outcomes.

The two most important categories for providing context to the key findings presented in this study include what accountability mechanisms are tied to those indicators and how the broader policy environment may or may not influence the effect of AP accountability incentives. Of the two categories, the one that included the least attention and specificity was the description states provided, or more commonly, did not provide, on the accountability mechanisms that would be used to improve AP outcomes. In fact, no policy-adopting state even alluded to specific accountability mechanisms in ESEA waivers, but these states did provide a more robust description of the supporting AP policy environment than ESSA state plans did.

Eight policy-adopting states described how Title IV, Part A funds, allocated through ESSA, would be used to expand AP access and improve AP outcomes. Title IV, Part A funds are distributed to school districts based on the Title 1 allocation formula. Thus, these funds are not specifically for low-performing schools or directly tied to AP exam accountability indicators. The only mention of specific accountability mechanisms discovered came from Arizona's ESSA plan, which stated Local Education Agencies identified as low-performing could implement evidence-based interventions focused on

providing accelerated learning opportunities, such as AP. However, no detail was provided to what types of evidence-based interventions schools can implement or what makes an intervention "evidence-based." Of course, the lack of detail regarding specific accountability mechanisms in high-level policy documents does not necessarily mean states are not providing resources, support, and other forms of technical assistance specific to improving AP exam outcomes for all schools and low-performing schools in particular. Further research, including comprehensive document analyses of state documentation and interviews with state officials, are needed to fully describe the capacity building efforts states use to improve AP outcomes. A deeper exploration of states AP policy initiatives will also help to further isolate the policy effect associated with AP exam accountability incentives and to refine the policy dosage categorizations.

Advanced Placement

The findings from this study fit within the small group of studies examining AP policy (Beach et al., 2019; Jackson, 2010; Jeong, 2009; Klopfenstein, 2004; Klugman, 2013; Kramer, 2016; Rowland & Shircliffe, 2016). However, the majority of these studies focus on resource-based policies and incentives whereas the accountability policies studied here do not provide schools with resources to induce participation or performance outcomes. Instead, the only resources provided via the accountability system, if they are any at all, generally come when schools are identified as low-performing, although some schools may receive resources in the form of professional development from general state support provided universally to all schools (e.g., AP-specific resources generated from mandatory school-based needs assessments; Title IV, Part A funds). To my knowledge, the only quantitative study to examine the influence of

school accountability incentives found that Pennsylvania's introduction of an AP access indicator led to an immediate increase in AP course offerings across all schools, suggesting schools respond to AP accountability incentives (Beach et al., 2019). If taken at face value, the results from this study suggest schools' short-term reaction to AP accountability incentives led to an immediate, negative post-policy intercept change, whereas the long-term reaction led to an increase in the post-policy slope.

As stated above, no immediate evidence is available to explain precisely why accountability incentives are associated with short-term decline and long-term growth in state-level AP exam performance. The findings from the brief document review described above point to the possibility that other state-level AP policies may interact with accountability incentives. Although state officials were generally light on the accountability details associated with AP indicators, they often signaled the belief that current and past policies had a positive effect on AP outcomes. Not surprising, many of these policies focused on providing test fee waivers for students, merit-based incentives, and other resource-based policies that are the primary focus of other AP policy research. However, states mentioned many other initiatives, such as providing access to AP courses through state run virtual programs, training programs for teachers put on in partnership with the College Board, and state-level task forces or committees dedicated to AP and other advanced coursework. This comes as no surprise given the breadth of state-level AP policies cataloged by the Education Commission of the States (2006; 2016).

Reports from the Education Commission of the States (2006; 2016) demonstrate varying levels of policy support for AP participation and performance across states. Questions remain about what AP policies are complimentary or detrimental to the

effectiveness of AP accountability incentives. An examination of the Education Commission of the State's 2016 report shows that policy-adopting states, on average, had a higher level of policy activity among the nine types of policies cataloged; policy-adopting states averaged 3.68 AP policies as compared to 2.72 for comparison states (excluding accountability policies). The most frequent policy among the states was providing support for access (n = 15), followed by teacher training (n = 11), financial support to improve access and performance (n = 10), subsidizing AP exam fees (n = 10), and mandating postsecondary institutions accept AP exam scores for college credit (n = 10). The existence of these policies indicates local history validity threats may be present, but timing is everything. For these other policies to be a true threat to validity, they need to be implemented at the same time as AP accountability incentives and influence AP exam outcomes. These other policies are part of a broader AP policy environment, an environment that could strengthen, weaken, or have no effect on AP exam accountability incentives. Beyond AP, the larger state-level policy environment that centers on college and career readiness may also influence AP accountability incentives. Data from the College and Career Readiness and Success Center at the American Institutes of Research (2019) shows that states also vary considerably in the level of policy support for college and career readiness initiatives (e.g., P-20 longitudinal data systems, support for advanced coursework and career and technical education pathways). Beyond AP and college and career readiness, teacher evaluation (Kraft et al., 2018) and other tangential polices may also interact in ways that strengthen or weaken AP accountability incentives.

Ultimately, a deep exploration of state AP and other college and career readiness policies across time is needed to fully describe and isolate the effect of AP exam

accountability incentives. There are several questions a deep exploration of this nature could help answer. For example, are accountability policies more effective in states with a strong commitment to the AP program, through state policy and support? Are accountability policies less effective if certain, other AP policies are in conflict? What effects do other CCR-focused policies (e.g., resources to support college readiness programs, incentives to provide access to accelerated coursework) have on the effect of AP accountability incentives? Do other policies, such as teacher evaluation systems, exert influence on AP accountability incentives.

Another important question this type of deep exploration could help answer is what caused these 19 states to introduce AP accountability indicators? A thorough understanding of past and current state-level AP policies would help expose existing ties to the College Board, the AP program, and its other flagship program, the SAT. Targeted Internet searches did not produce detailed information on the College Board's lobbying activities on a state-by-state basis. However, interviews with state officials and systematic document analyses could uncover connections between the College Board and certain states, which may make it more likely these states would adopt AP accountability incentives. For its part, the College Board does mention in at least one document that AP exam data can be used for "college and career readiness accountability measures" (College Board, 2015, p. 5) and in another document touts positive effects of doing so in Pennsylvania using descriptive statistics (College Board, 2020c). However, the College Board's website did not include many other easily accessible documents that advocated for the adoption of AP exam accountability incentives.

Of course, College Board lobbying and advocacy may not be the main factor

explaining why AP accountability incentives were introduced in these 19 states. Perhaps these states are simply building on an already deep commitment to the AP program that is evident in the higher level of AP policy activity. These states, on average, also had lower exam scores than comparison states. Were these policy-adopting states driven by the desire to improve lower than desired average state-level AP exam scores? Finally, the clustering of states in the South and Midwest, coupled with a slight advantage to conservative-leaning states, suggests a more detailed analysis of the political environment in these states as well as geographical considerations should be examined in future research. At the very least, the lack of statistical coherence found in this study specific to the effect of AP accountability incentives, and the contextual information presented above, necessitates the need for a mixed methods approach in addition to the rigorous quantitative examination of time series data.

Policy Analyses using Aggregated Time Series Data

The literature examining the methodological implications associated with different approaches to modeling time series data in quasi-experiments is limited, which is somewhat surprising given the availability of large scale longitudinal datasets has grown rapidly in recent years. A few researchers have compared and contrasted difference-in-differences and CITS designs. These researchers find that both design types can yield internally valid estimates that produce the same policy effect interpretation with similar standard errors (Gordon & Heinrich, 2004; Somers et al., 2013). However, a typical conclusion is that the CITS design is more rigorous because it "implicitly controls for differences in the baseline mean and trends between the treatment and comparison group" (Somers et al., 2013, p. 3). One potential reason for the lack of studies comparing

difference-in-differences and CITS analyses could be that both use slightly different names for the same design. For example, "the different ways in which the econometrics and CITS literatures use the terms fixed effects and random effects can be confusing" (Gordon & Heinrich, 2004, p. 174).

Yet, both designs are intended to answer the same types of questions and can produce similar results (Gordon & Heinrich, 2004; Somers et al., 2014). However, this study demonstrates that the difference-in-differences and CITS models can also produce findings that differ in meaningful ways. The CITS models produced all but one of the statistically significant, policy-specific findings across all analyses—the difference-in-differences models produced none. As a result, these findings suggest that researchers may come to very different conclusions about the effectiveness of policy interventions depending on what design they use for modeling policy effects where treated units implement the intervention at different times.

The differences in statistically significant policy effects between the two designs may have resulted from the violation of the parallel trends assumption in the difference-in-differences estimates. The policy-adopting states exhibited a downward trend in AP exam scores prior to policy implementation. Since the difference-in-differences design assumed pre-policy trends were parallel between policy-adopting and comparison states, the post-policy increases in AP exam scores appear to be parallel with pre-policy estimates, which would explain the null policy effects across all AP exams. Conversely, the CITS models explicitly controlled for this downward pre-policy trend. The small, positive post-policy slope effects were compared relative to the downward trend, thus capturing the small, positive post-policy slope effects.

Unfortunately, no studies could be located that use both a difference-in-differences and CITS design for modeling the staggered implementation of a policy intervention. This gap in the literature is notable for two reasons. First, the way econometricians typically model staggered implementation is different than what is recommended in the CITS literature. In the difference-in-differences setup, time is centered on the year of policy implementation in treatment units while the comparison group is coded as a constant -1 and does not contribute to pre-policy slopes estimates (Athey & Imbens, 2018; Goodman-Bacon, 2018). In the CITS approach, time remains on the same series for treatment and comparison groups and both groups contribute to pre-policy slope estimates (Raudenbush & Bryk, 2002; Singer & Willet, 2003). It should be noted that the approach to modeling staggered implementation within a difference-in-differences design is a relatively new area of econometrics and only one application in education was found (Liebowitz et al., 2019). Similarly, only one study in education was found that employed the recommended coding scheme for modeling staggered implementation using a CITS design (Boal & Nakamoto, 2020).

Second, across education and in school accountability in particular, researchers will increasingly need to model the staggered implementation of policy interventions. For example, states have introduced numerous new school quality indicators in addition to AP exam indicators (Aldeman, 2017a, 2017b; Polikoff et al., 2014). There are countless other examples in education more broadly, where schools or classrooms implement certain programs or polices at different times. Perhaps the lack of research in these areas is due to the limitations and challenges inherent to using time series data to analyze policy shocks, which some researchers attribute to an outsized focus on other topics, such

as charter school effects, which are more conducive to randomized trials and more rigorous quasi-experiments (Saw & Schneider, 2015).

The meaningful discrepancies between the two approaches, and the increasing need to model staggered implementation, signals a need for a comprehensive methodological study that analyzes a field-based quasi-experiment such as one presented in this study. A within-study design is another approach that could further examine the design characteristics and tradeoffs associated with modeling staggered implementation using a difference-in-differences or a CITS design. As mentioned above, two of the clearest distinctions between the two designs are the modeling of time and the selection and coding of comparison groups. Additionally, there are several different approaches to constructing comparison groups, such as using all available comparison units, statistical matching, local or geographical matching, and hybrid options (Hallberg et al., 2018; Jacob et al., 2016). This study used the option of all available comparison units, which "can provide unbiased estimates of causal effects in the CITS context as long as time-varying confounds do not operate differentially across the treatment and comparison groups" (Hallberg, 2018, p. 298). The lack of statistical significance among the time-varying covariates, although not conclusive, does provide some evidence that changes in state characteristics did not bias the policy-specific results. Local and statistical matching was limited by the small sample size of 50 states and DC, which resulted in a lack of comparable large states or comparable states in the South and Midwest, the two regions most policy-adopting states were clustered in.

One new issue pertaining to the construction of comparison groups deserves greater scrutiny in future research, especially within the difference-in-differences

literature. While existing state-level studies in education that leveraged staggered implementation included 44 treated and seven comparison units (Liebowitz et al., 2019), the ratio in the current study was reversed, with 19 treated and 32 comparison units. Future research should explore the implications that arise from an imbalance between policy-adopting and comparison states. Weighting estimates by group size or participation levels, as this study did, is another issue closely tied to constructing comparison groups. As stated earlier, weighted estimates accounted for more than half of the statistically significant policy effect findings found in the CITS models. Future research should explore whether this is policy-driven or a methodological pattern. Finally, analysts studying problems that require methods similar to those employed here will be confronted with several other decisions, such as how to use fixed and random effects, what role covariates should play, and the implications of using clustered standard errors or hierarchical linear modeling to account for clustering effects. A comprehensive guide addressing the above issues would be a valuable resource for researchers.

Conclusion

In conclusion, the findings from this study represent some of the first pieces of evidence on how accountability incentives influence state-level AP outcomes. This study suggests state policymakers could see a range of effects from using AP exam data as a school accountability indicator, though most of these effects are relatively small in magnitude compared to overall trends in AP exam scores across time. Policymakers may see a positive increase in the trend of average state-level AP exam performance on a magnitude of between three- and eight-hundredths of a standard deviation. Policymakers could also experience short-term disappointment in the form of an immediate decrease in

average AP exam performance. That immediate decrease, however, will be alleviated by a long term increase in average performance that is larger in magnitude than the immediate decrease. Or, policymakers could experience an immediate decrease so large, and long-term increase so small, that the net effect is a one-time decrease in average AP exam performance. On the other hand, the inconsistent evidence across AP exams with respect to statistical significance produces weak evidence of policy effects, which may mean the only risks policymakers are taking in introducing AP exam accountability incentives are the opportunity costs that result from the time and resources dedicated to choosing, designing, and implementing new accountability indicators.

This type of effect size interpretation produces a key question for state policymakers. Is it worth it? What are benefits and costs associated with introducing AP exam accountability indicators in school accountability systems? How many resources are used to choose, design, and implement AP accountability indicators? If there are benefits in the form of improved student AP outcomes, is the ratio of benefits to costs warranted? Would state resources be better spent on other policy approaches for improving students' AP outcomes? How does policy dosage and other policy factors impact this cost-benefit calculation? Future research should attempt to answer these questions with respect to AP and other accountability incentives.

The findings from this study and lessons from past research also show state policymakers should pay close attention to the incentives generated by the design of AP accountability indicators. Although only speculations at the point, the potential dip in AP exam scores immediately after the introduction of AP exam accountability indicators, coupled with the long-term increase across time, suggests schools may be employing

exclusionary participation tactics similar to what was observed under NCLB (Cullen & Reback, 2006; Figlio & Getzler, 2007). That is, it seems plausible that an initial, positive reaction from schools to the introduction of AP exam indicators would be to increase the number of students who sit for AP exams. Across time, research has suggested increases in AP exam participation lead to lower overall scores (Kolluri, 2018). A reaction to the lower average scores may incentivize schools to discourage low-performing students to sit for AP exams, either through gatekeeping tactics or by enacting other structural or social barriers. The fact that only one in five (21%) policy-adopting states report AP exam accountability data by subgroups of students ensures most schools would not be held accountable for this type of exclusionary behavior. Further examination of AP exam performance, participation, and access data as well as interviews with state officials, school administrators, and teachers are needed to develop a more nuanced understanding of the incentives created by AP exam accountability indicators.

The search for factors that may influence AP accountability incentives extends beyond the design of indicators. Reports from the Education Commission of States (2006; 2016) show at least nine other types of AP policies exist across all states. The brief analysis of those policies shows that policy-adopting states appear to have more AP policies than comparison states on average. Research shows that these policies can interact with the intent of school accountability incentives. For example, efforts to improve average school-level AP exam performance may conflict with efforts to increase AP participation (Rowland & Shircliffe, 2016). In addition, research from Klugman (2013) showed that one combination of AP policies in California led to greater inequities in access to AP courses. It seems equally plausible that some combinations of AP policies

may improve the effectiveness of AP accountability incentives. For example, it seems plausible that providing funding for AP-specific teacher training would provide a way for schools to access an additional avenue of support—in addition to the public pressure or threat of sanctions from accountability incentives—for improving students' AP exam scores. Ultimately, more research is needed to fully evaluate the nuanced relationships among different AP policies, but it is incumbent on state policymakers to factor these policies into decisions about adopting and designing AP exam accountability incentives.

Outside of the AP program itself, state policymakers should also be aware of and evaluate how AP exam incentives fit into the larger accountability system and broader state policy ecosystem. One important difference among states are the number of indicators employed in school accountability systems. For example, in some states AP exam data accounts for one of five overall indicators (e.g., Pennsylvania, Oklahoma). More commonly, AP accountability is one of many sub-indicators within a larger college and career indicator. In some of these states, students can earn credit for their school by qualifying AP exam scores or sufficient performance on one of several other sub-indicators. School improvement specialists suggest that schools can only be expected to produce meaningful change on a small number of goals (Kaufman, 2012). Is it reasonable to expect that schools can improve on several different accountability indicators at one time? How do schools prioritize addressing performance on many different accountability indicators at one time? How large of an effect should state policymakers expect from the introduction of AP exam accountability indicators?

Of course, a state's accountability system resides within the broader education ecosystem that contains a multitude of policies interacting within and across different

levels of system. This system includes several types of state policy initiatives, such as those focused on postsecondary readiness, teacher quality, and school funding. Setting aside the number and scope of various policy initiatives, school leaders and teachers respond to various goals from stakeholders (e.g., city governments, district administrators, parents). Although no one study can fully expect to account for the myriad of factors that influence AP accountability incentives, the results and contextual information presented in this study suggest future research should employ a mixed methods approach. A rigorous mixed methods approach has the potential to further isolate the effects of AP accountability incentives and provide needed evidence on how school leaders and teachers respond to these incentives.

Finally, the lessons learned throughout this study, as well as those that would be gleaned from a more expansive research agenda, will also inform future studies on other school quality indicators. Very little research has examined the effect of school accountability incentives on the ever expanding list of the indicators states employ. What, for example, is the effect of school accountability on college admission exams (i.e., SAT, ACT), an indicator used in several states (Aldeman, 2017a, 2017b). Does introducing accountability indicators for chronic absenteeism actually result in improved attendance in schools? How does school accountability effect graduation rates? The overall findings from this study, as well as the accompanying contextual information, suggest the answer to these questions is unlikely to be simple. Continuing to produce rigorous evidence on the effects of school accountability on these and other outcomes is needed further explore how well public resources are being spent in the pursuit of school improvement.

APPENDIX A

CODING OF TIME AND POLICY INTERVENTION VARIABLES

Table A1.

Coding of Time and Policy Intervention Variables

Variable	Coding Scheme using 2014 as Example Year of Policy Intervention
TxTrend	1997 = -17, 1998 = -16 … 2013 = -1, 2014 = 0, 2015 =1, 2016 = 2…
PreTxSlope	1997 = 0, 1998 = 1, 1999 = 2… 2019 = 23
Policy	1997-2013 = 0, 2014-2019 = 1
PostTxSlope	1997-2013 = 0, 2014 = 0, 2015 =1, 2016 = 2…

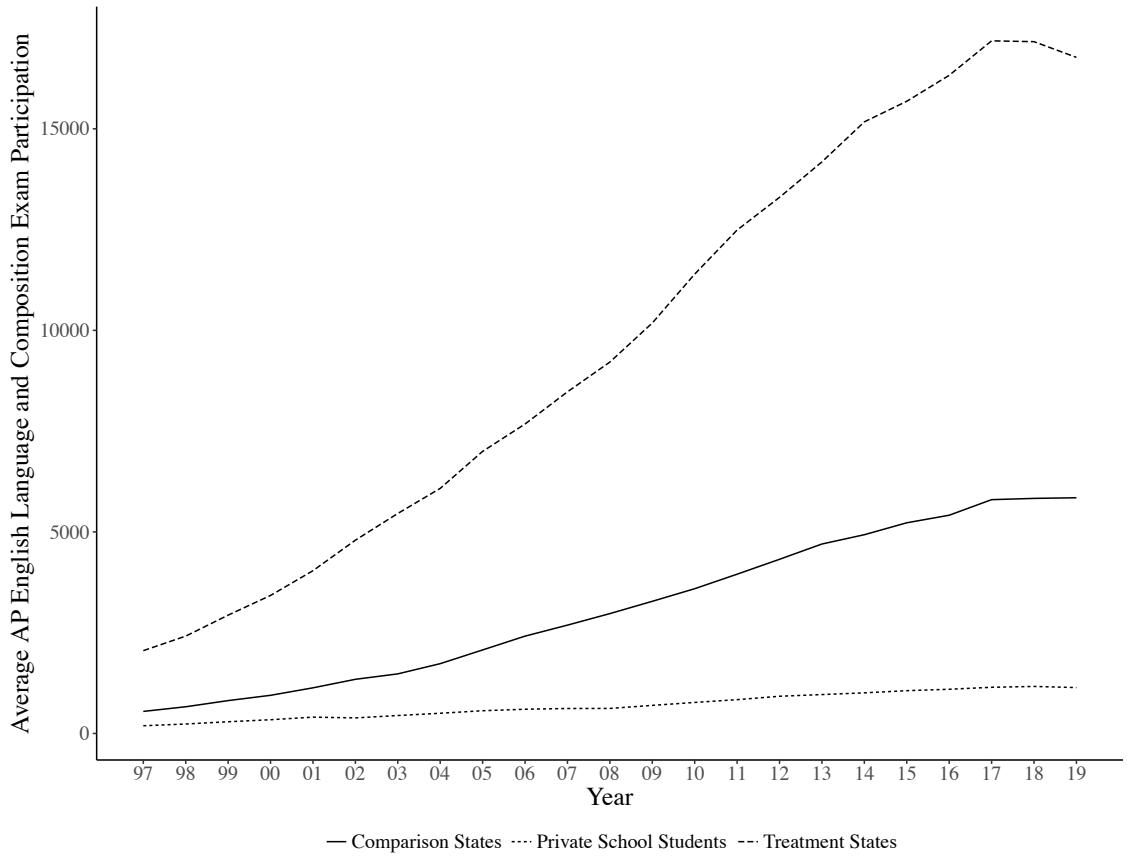
Note. Dummy coded variables for TxTrend used in event studies; TxTrend and Policy used in difference-in-differences models; PreTxSlope, Policy, and PostTxSlope used in CITS models

APPENDIX B

TRENDS IN AP EXAM PERFORMANCE

Figure B1

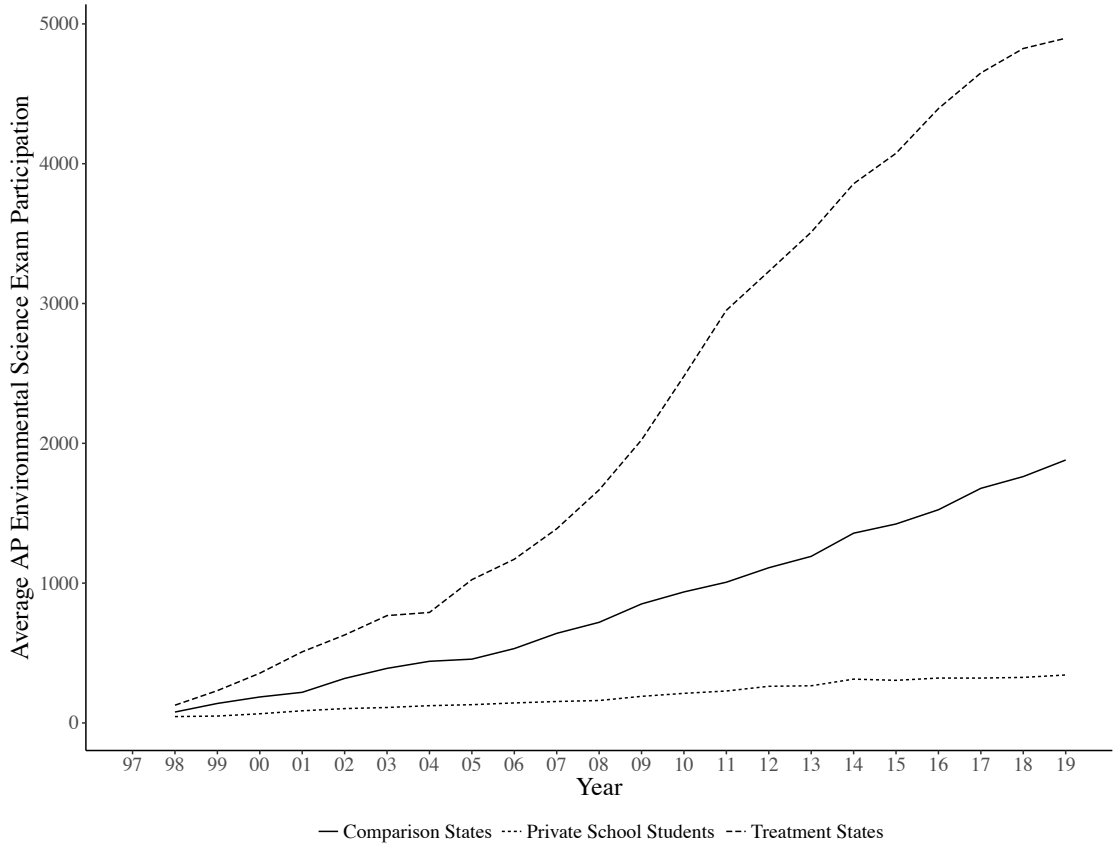
Average AP English Language and Composition Exam Participation



Note. Average AP English Language and Composition exam participation for public school students in policy-adopting (n = 19) and comparison states (n = 32) and private school students in all states (n = 51) from academic years 1996-97 to 2018-19

Figure B2

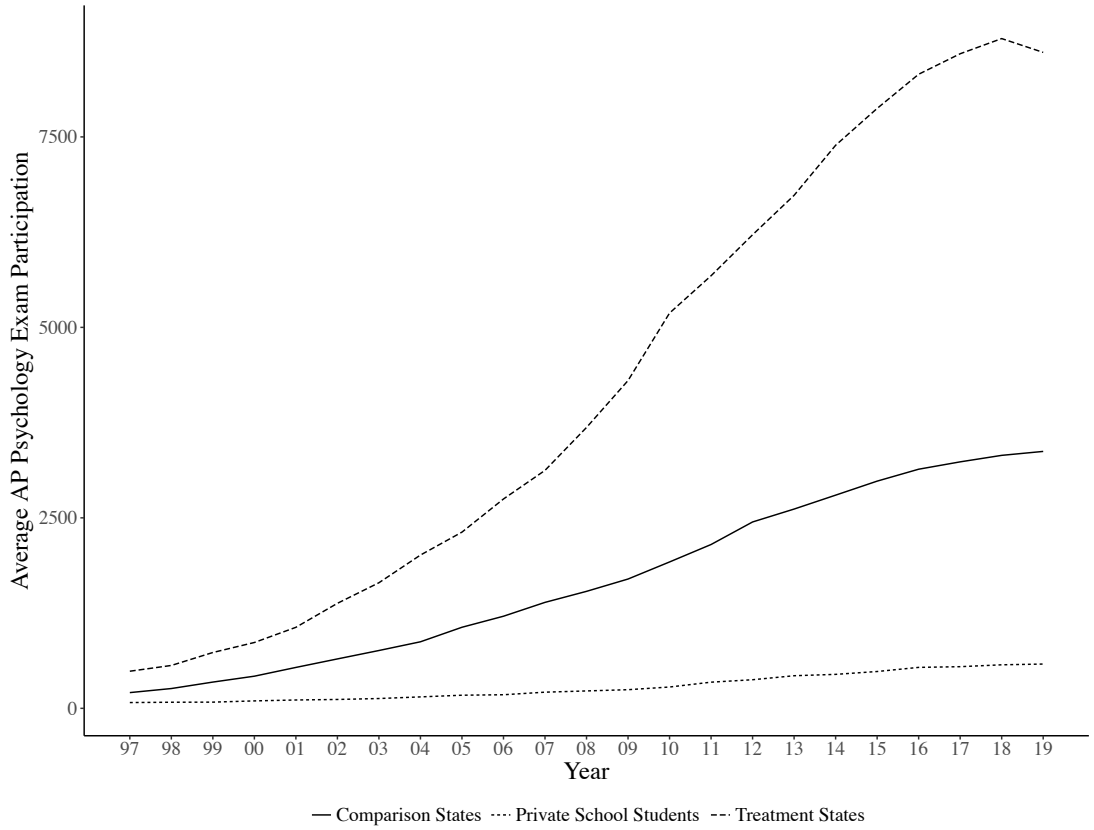
Average AP Environmental Science Exam Participation



Note. Average AP Environmental Science exam participation for public school students in policy-adopting (n = 19) and comparison states (n = 32) and private school students in all states (n = 51) from academic years 1996-97 to 2018-19

Figure B3

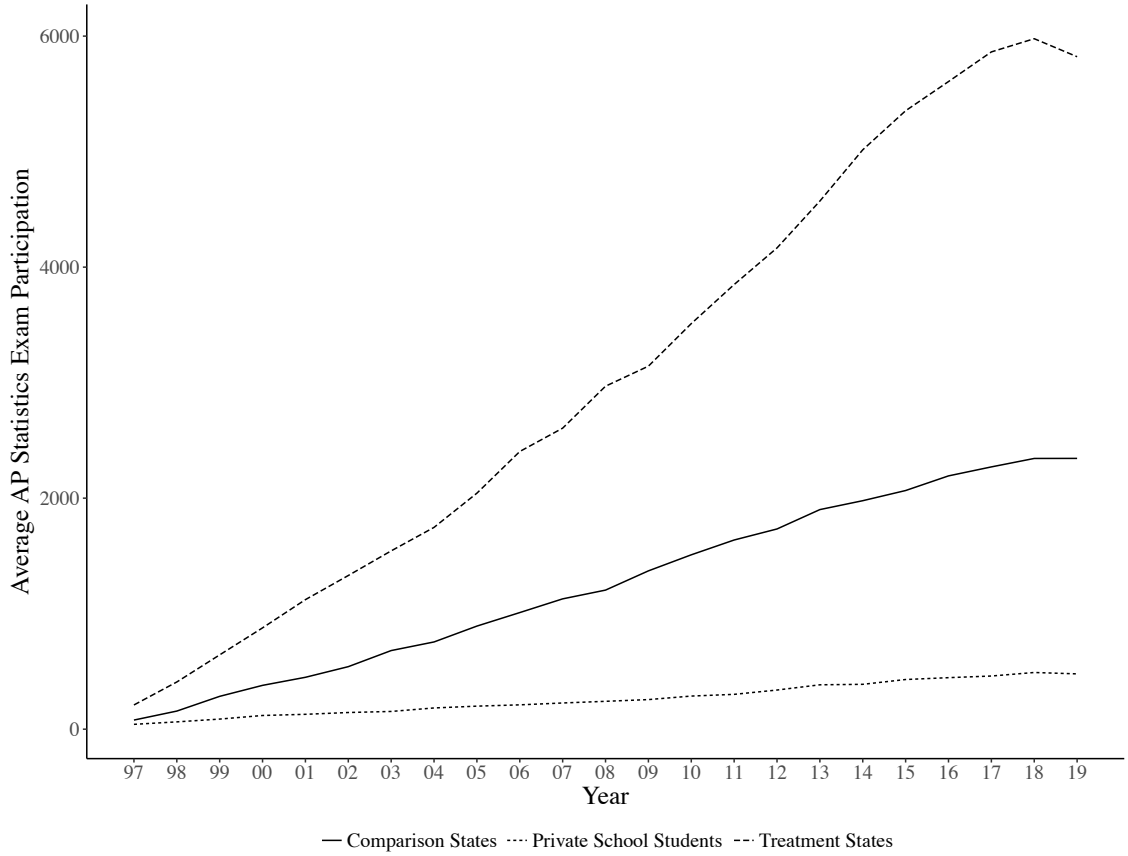
Average AP Psychology Exam Participation



Note. Average AP Psychology exam participation for public school students in policy-adopting (n = 19) and comparison states (n = 32) and private school students in all states (n = 51) from academic years 1996-97 to 2018-19

Figure B4

Average AP Statistics Exam Participation



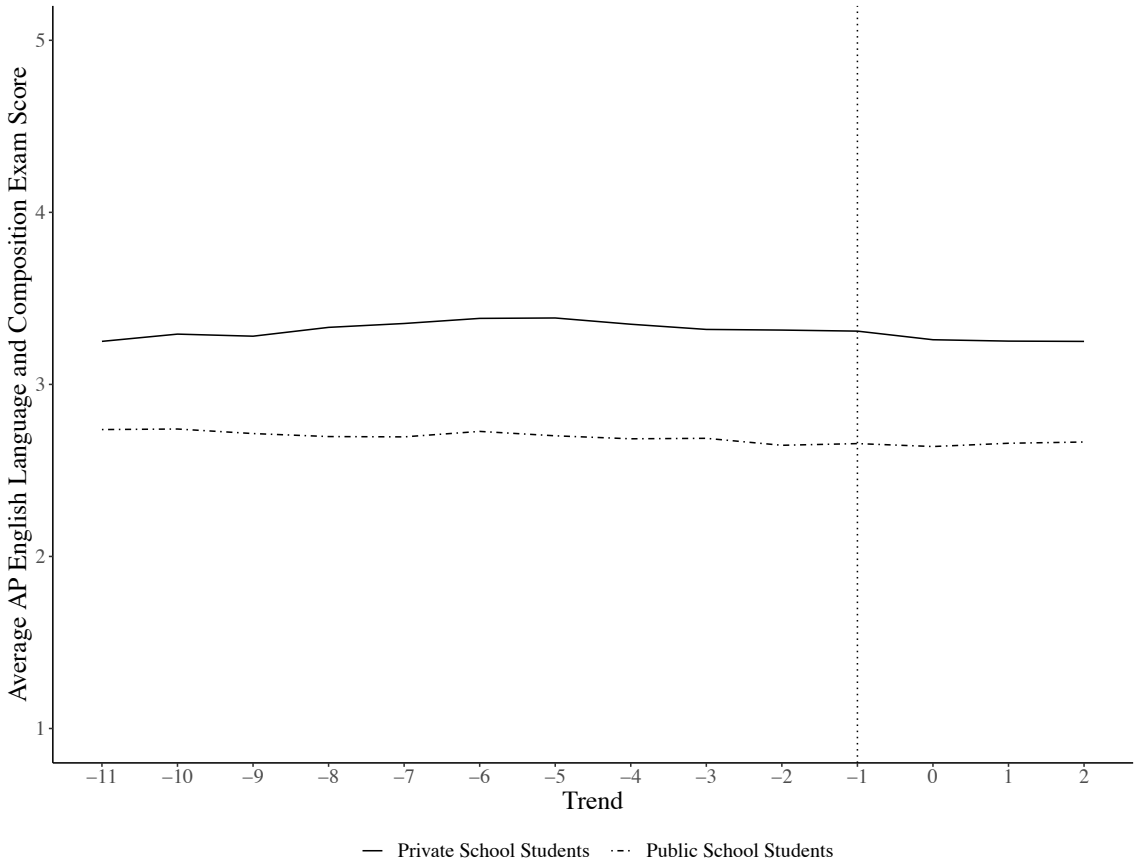
Note. Average AP Statistics exam participation for public school students in policy-adopting (n = 19) and comparison states (n = 32) and private school students in all states (n = 51) from academic years 1996-97 to 2018-19

APPENDIX C

VISUAL INSPECTION OF POLICY EFFECTS

Figure C1

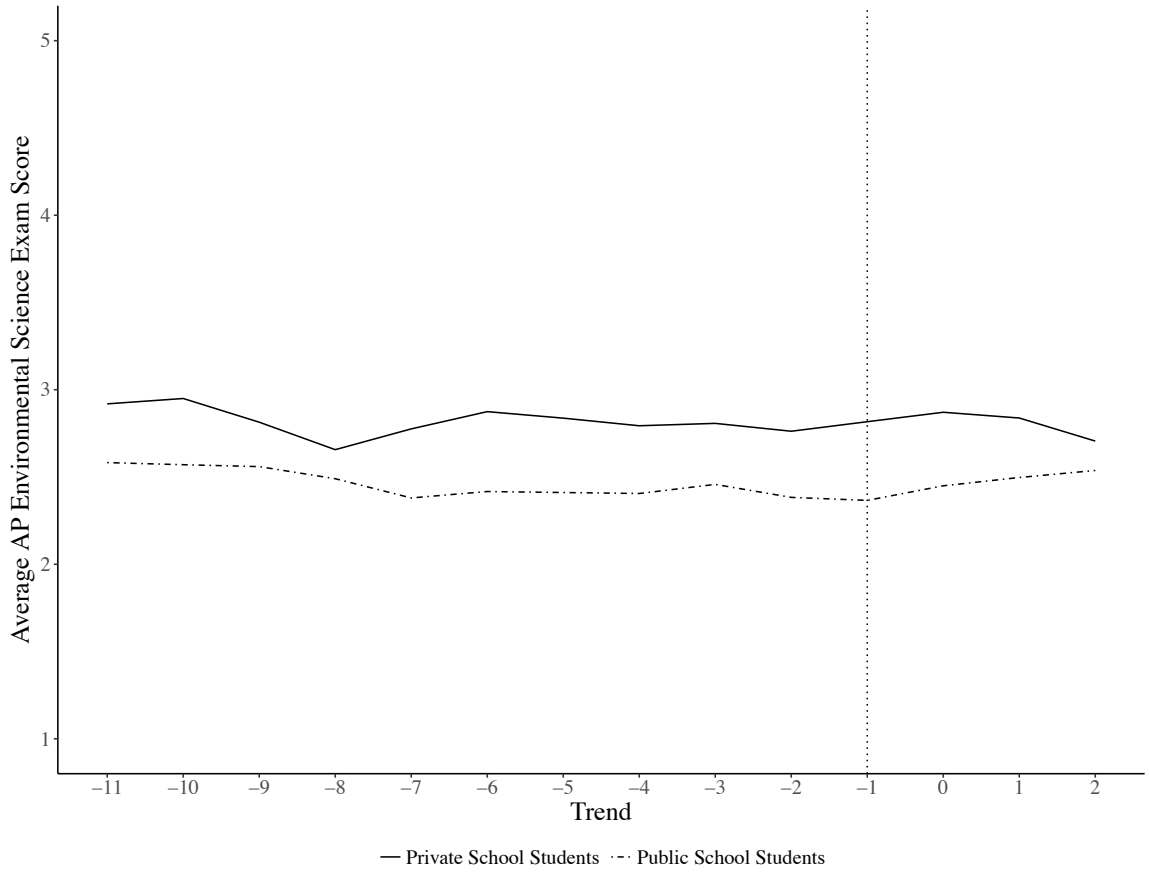
Policy Effects for AP English Language and Composition Exam Performance



Note. Average AP English Language and Composition exam scores for public and private school students in policy-adopting state (n = 19) states centered on the year of policy implementation. At Trend = 0, all 19 policy-adopting states introduced an AP exam accountability indicator into its school accountability system. All policy-adopting states have 11 years of pre- and 2 years of post-policy data

Figure C2

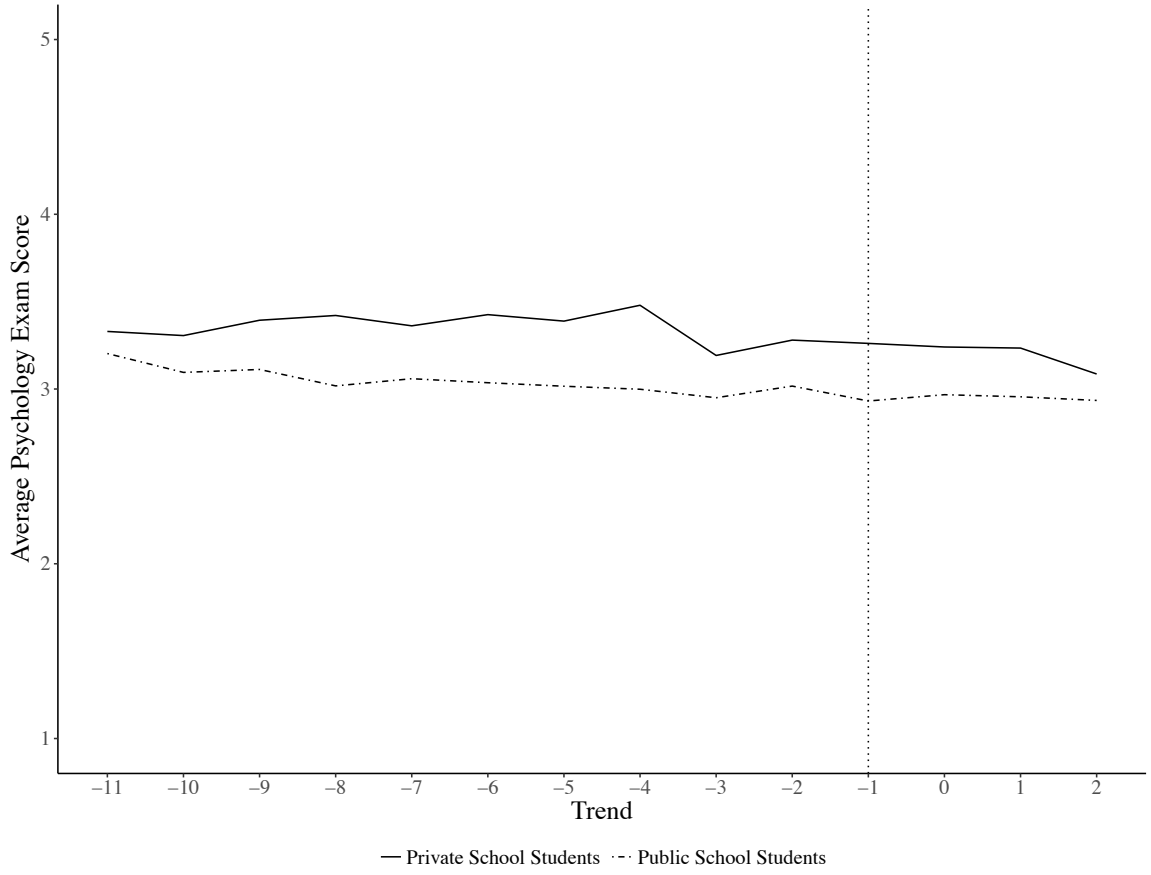
Policy Effects for AP Environmental Science Exam Performance



Note. Average AP Environmental Science exam scores for public and private school students in policy-adopting state (n = 19) states centered on the year of policy implementation. At Trend = 0, all 19 policy-adopting states introduced an AP exam accountability indicator into its school accountability system. All policy-adopting states have 11 years of pre- and 2 years of post-policy data

Figure C3

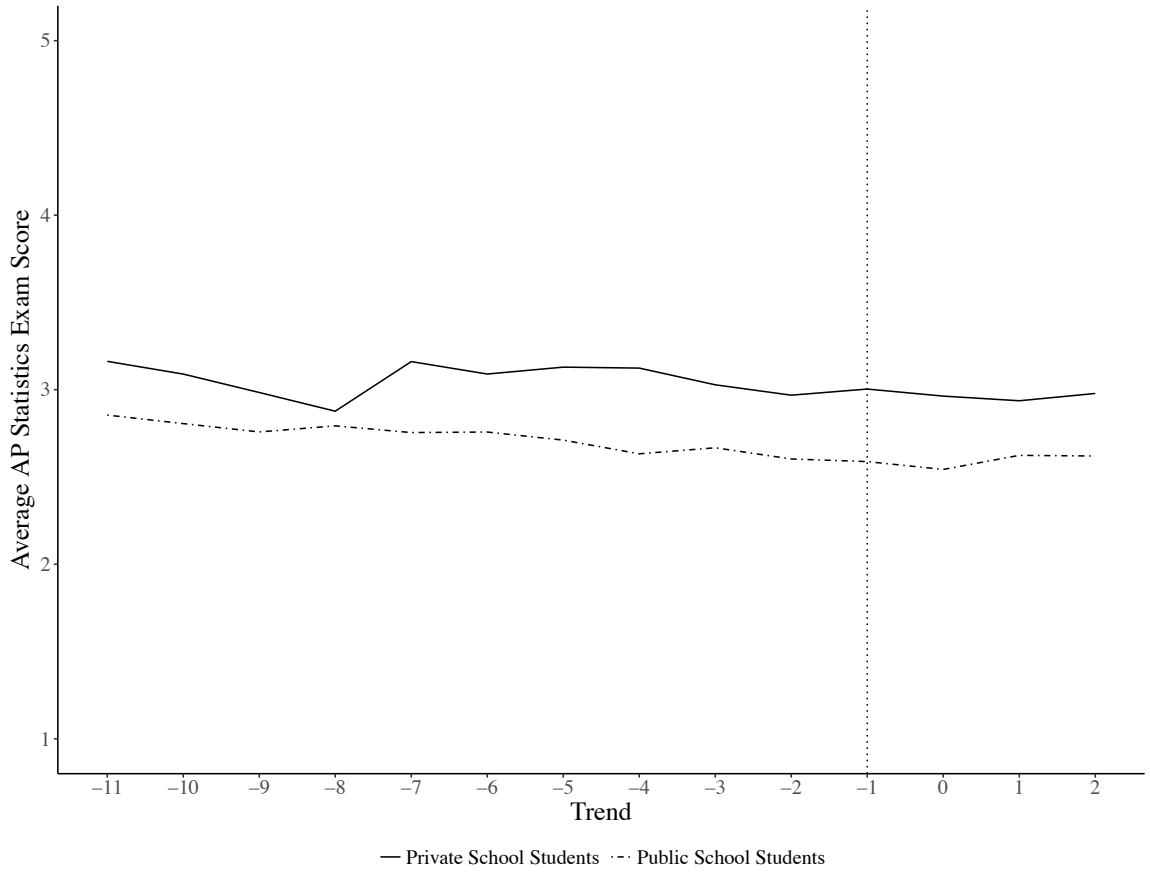
Policy Effects for AP Psychology Exam Performance



Note. Average AP Psychology exam scores for public and private school students in policy-adopting state (n = 19) states centered on the year of policy implementation. At Trend = 0, all 19 policy-adopting states introduced an AP exam accountability indicator into its school accountability system. All policy-adopting states have 11 years of pre- and 2 years of post-policy data

Figure C4

Policy Effects for AP Statistics Exam Performance



Note. Average AP Statistics exam scores for public and private school students in policy-adopting state (n = 19) states centered on the year of policy implementation. At Trend = 0, all 19 policy-adopting states introduced an AP exam accountability indicator into its school accountability system. All policy-adopting states have 11 years of pre- and 2 years of post-policy data

Figure C5

Policy Effects for AP English Language and Composition Exam Participation

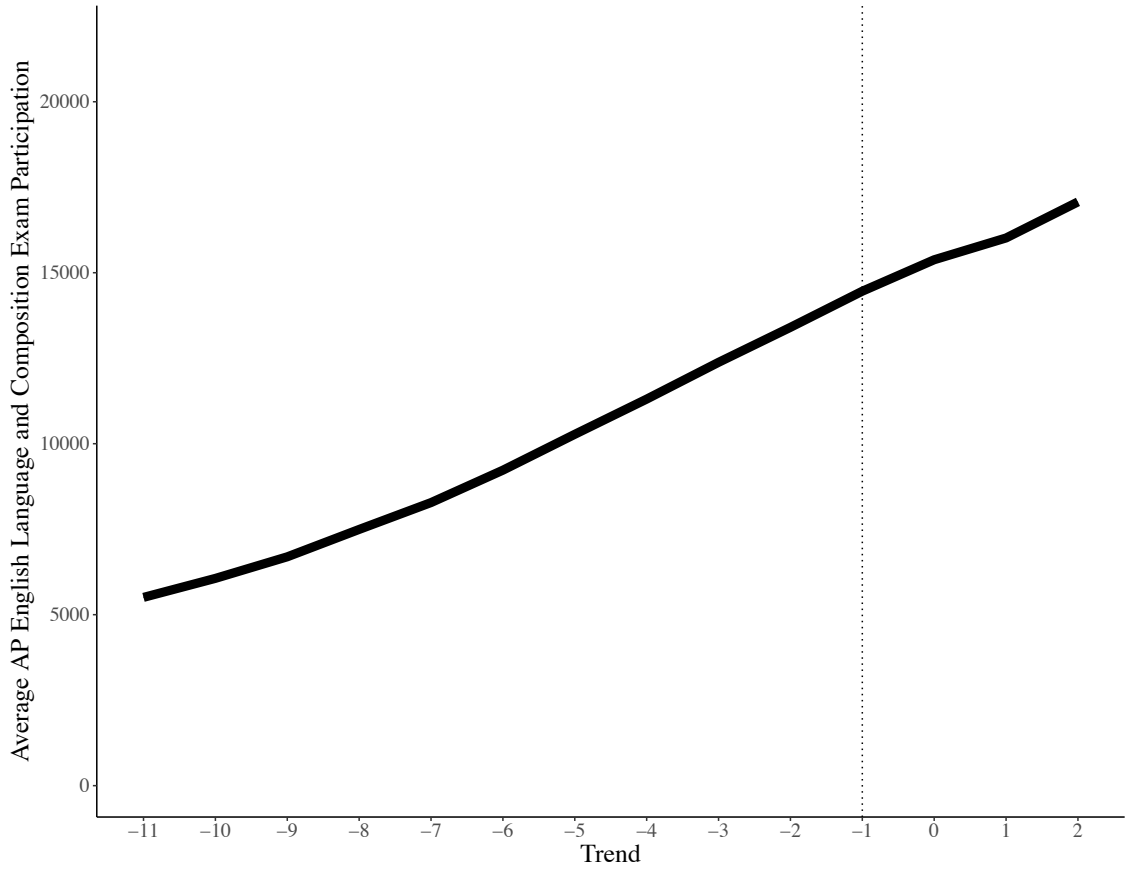
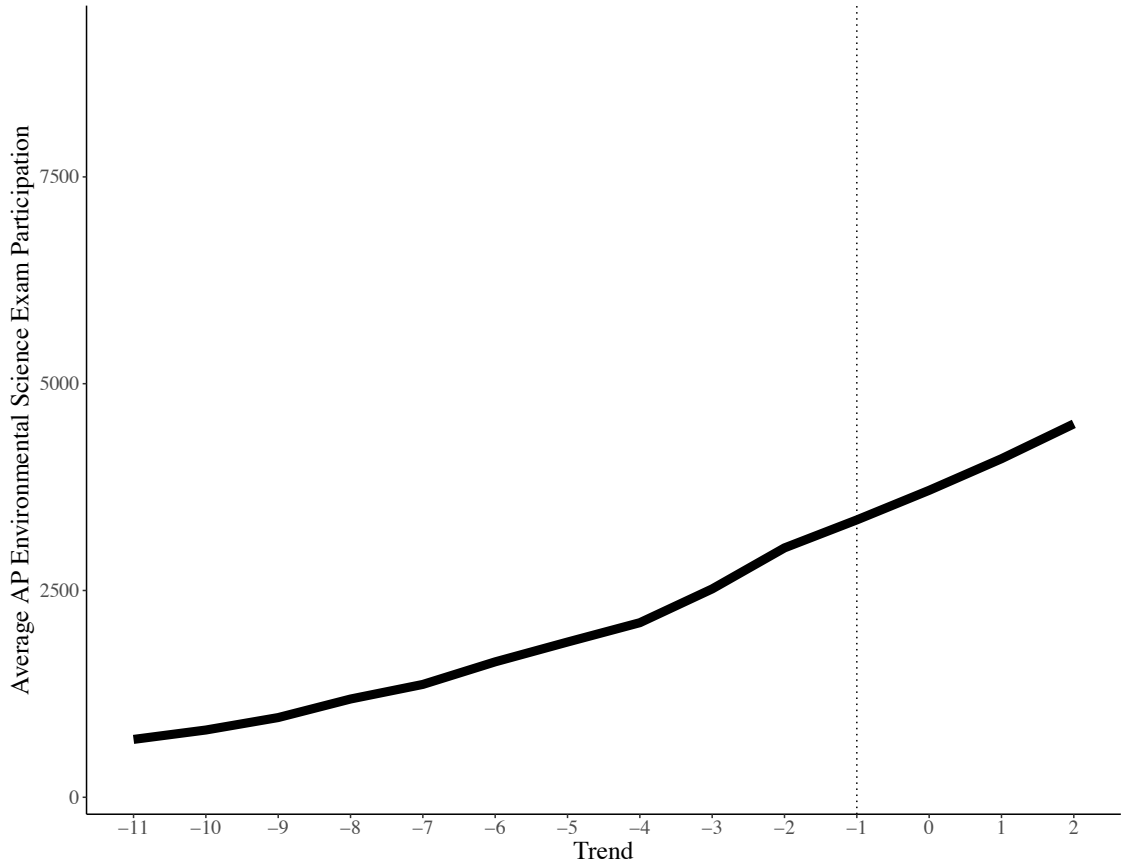


Figure 13. Average AP English Language and Composition exam participation for public school students in policy-adopting (n = 19) states centered on the year of policy implementation. At Trend = 0, all 19 policy-adopting states introduced an AP exam accountability indicator into its school accountability system. All policy-adopting states have 11 years of pre- and 2 years of post-policy data

Figure C6

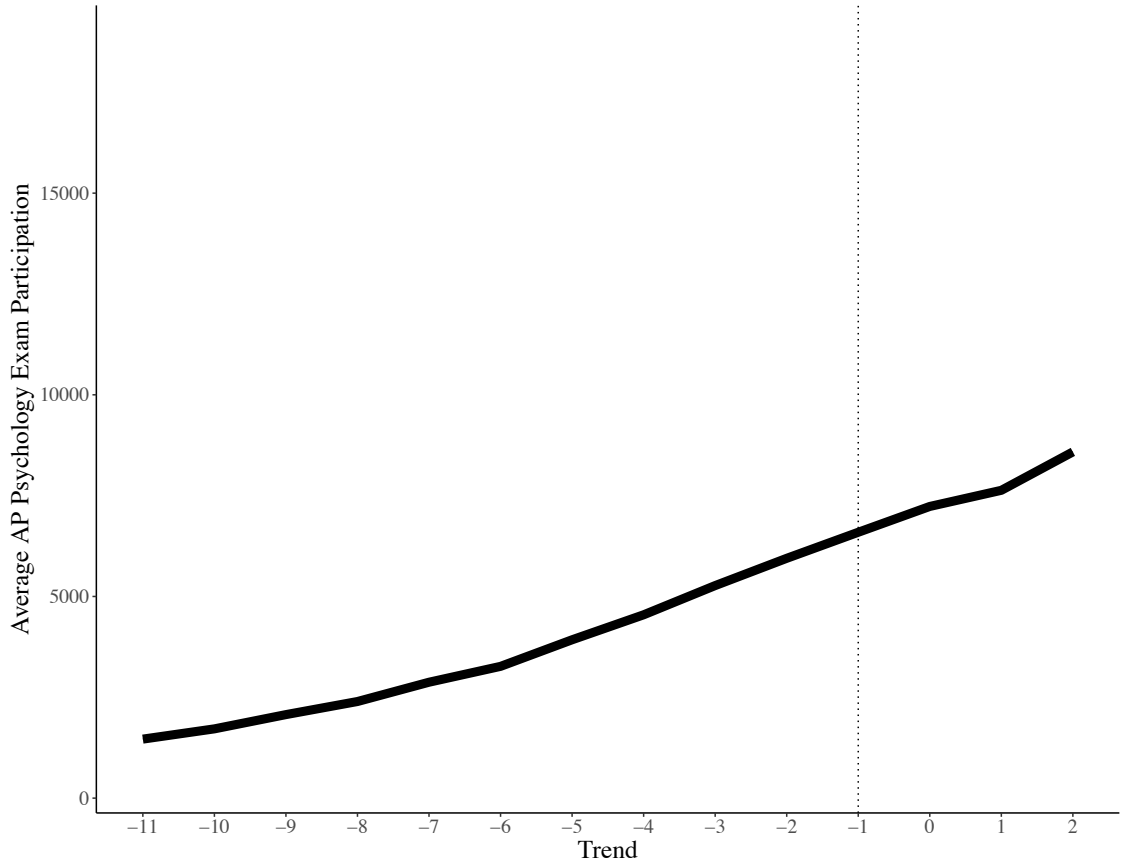
Policy Effects for AP Environmental Science Exam Participation



Note. Average AP Environmental Science exam participation for public school students in policy-adopting (n = 19) states centered on the year of policy implementation. At Trend = 0, all 19 policy-adopting states introduced an AP exam accountability indicator into its school accountability system. All policy-adopting states have 11 years of pre- and 2 years of post-policy data

Figure C7

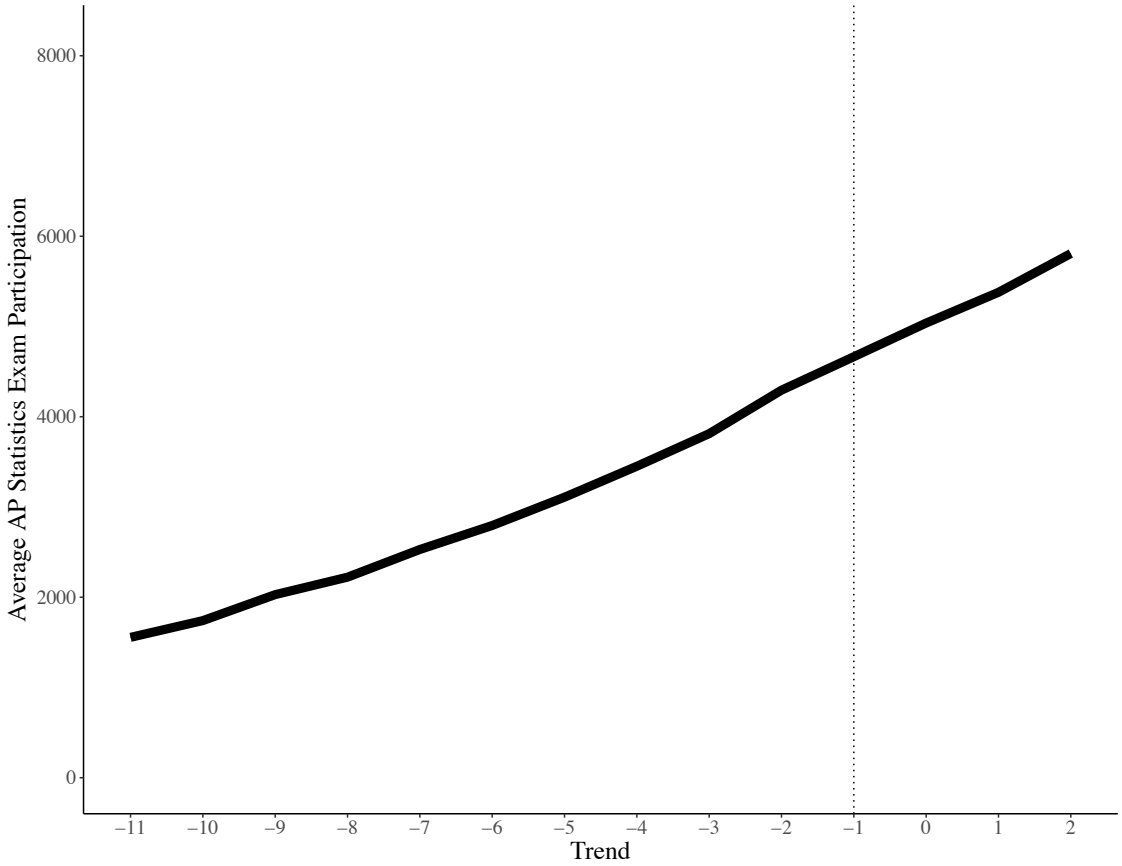
Policy Effects for AP Psychology Exam Participation



Note. Average AP Psychology exam participation for public school students in policy-adopting (n = 19) states centered on the year of policy implementation. At Trend = 0, all 19 policy-adopting states introduced an AP exam accountability indicator into its school accountability system. All policy-adopting states have 11 years of pre- and 2 years of post-policy data

Figure C8

Policy Effects for AP Statistics Exam Participation



Note. Average AP Statistics exam participation for public school students in policy-adopting (n = 19) states centered on the year of policy implementation. At Trend = 0, all 19 policy-adopting states introduced an AP exam accountability indicator into its school accountability system. All policy-adopting states have 11 years of pre- and 2 years of post-policy data

APPENDIX D

EVENT STUDY ESTIMATES

Table D1

Event Study Estimates of the Effect of AP Accountability Indicator: Baseline

	AP ELC	AP ES	AP Psychology	AP Statistics
11 Yr. Pre	-0.086 (0.100)	0.117 (0.117)	0.207 (0.078)*	0.147 (0.094)
10 Yr. Pre	-0.082 (0.104)	0.047 (0.116)	0.223 (0.079)*	0.101 (0.093)
9 Yr. Pre	-0.074 (0.112)	0.079 (0.115)	0.177 (0.077)*	0.070 (0.093)
8 Yr. Pre	-0.093 (0.092)	-0.011 (0.116)	0.077 (0.076)	0.123 (0.093)
7 Yr. Pre	-0.097 (0.088)	-0.092 (0.114)	0.135 (0.077)	0.097 (0.093)
6 Yr. Pre	-0.090 (0.094)	-0.062 (0.114)	0.090 (0.075)	0.126 (0.093)
5 Yr. Pre	-0.086 (0.097)	-0.034 (0.113)	0.076 (0.076)	0.072 (0.092)
4 Yr. Pre	-0.101 (0.082)	-0.014 (0.112)	0.051 (0.075)	0.027 (0.092)
3 Yr. Pre	-0.097 (0.085)	0.035 (0.112)	0.031 (0.074)	0.061 (0.091)
2 Yr. Pre	-0.108 (0.074)	-0.005 (0.112)	0.079 (0.075)	0.014 (0.092)
1 Yr. Pre	0	0	0	0
Zero Yr.	-0.091 (0.092)	0.069 (0.112)	0.039 (0.077)	-0.008 (0.092)
1 Yr. Post	-0.077 (0.105)	0.128 (0.112)	0.003 (0.077)	0.046 (0.092)

Note. *p < .05; All models include state and year fixed effects; Standard errors in parentheses; Coefficients for 12+ Yrs. Pre include all observations 12 or more years prior to policy implementation and coefficients for 2+ Yrs. Post include all observations 2 or more years post policy implementation—all other coefficients are balanced by including observations from all policy-adopting states; Zero Yr. = year of policy implementation; Reference group for all models includes all observations in all years for comparison states and all observations for policy-adopting states the year prior to policy implementation

Table D2

Event Study Estimates of the Effect of AP Accountability Indicators: Covariate-Adjusted

	AP ELC	AP ES	AP Psychology	AP Statistics
11 Yr. Pre	-0.007 (0.046)	0.102 (0.119)	0.198 (0.078)*	0.164 (0.092)
10 Yr. Pre	0.045 (0.046)	0.008 (0.121)	0.216 (0.079)*	0.172 (0.092)
9 Yr. Pre	0.014 (0.046)	0.075 (0.118)	0.170 (0.078)*	0.085 (0.091)
8 Yr. Pre	0.000 (0.045)	-0.032 (0.117)	0.069 (0.076)	0.165 (0.089)
7 Yr. Pre	-0.001 (0.045)	-0.130 (0.116)	0.128 (0.077)	0.139 (0.09)
6 Yr. Pre	0.002 (0.045)	-0.097 (0.115)	0.084 (0.076)	0.154 (0.089)
5 Yr. Pre	0.004 (0.045)	-0.070 (0.115)	0.070 (0.076)	0.100 (0.089)
4 Yr. Pre	-0.010 (0.044)	-0.047 (0.113)	0.048 (0.075)	0.051 (0.088)
3 Yr. Pre	-0.008 (0.044)	0.000 (0.113)	0.028 (0.075)	0.081 (0.088)
2 Yr. Pre	-0.020 (0.044)	-0.031 (0.114)	0.078 (0.075)	0.022 (0.089)
1 Yr. Pre	0	0	0	0
Zero Yr.	-0.003 (0.046)	0.018 (0.117)	0.039 (0.077)	-0.028 (0.091)
1 Yr. Post	-0.007 (0.046)	0.063 (0.118)	0.002 (0.078)	0.014 (0.091)

Note. *p < .05; All models include state and year fixed effects; Standard errors in parentheses; [a] State demographic controls include the number of Grade 11 and 12 students, percentage eligible for free- and reduced-price lunch, and the percentage of Asian, Black, Hispanic, and White students; Coefficients for 12+ Yrs. Pre include all observations 12 or more years prior to policy implementation and coefficients for 2+ Yrs. Post include all observations 2 or more years post policy implementation—all other coefficients are balanced by including observations from all policy-adopting states; Zero Yr. = year of policy implementation; Reference group for all models includes all observations in all years for comparison states and all observations for policy-adopting states the year prior to policy implementation

Table D3

Event Study Estimates of the Effect of AP Accountability Indicators: Weighted

	AP ELC	AP ES	AP Psychology	AP Statistics
11 Yr. Pre	0.062 (0.023)*	0.182 (0.049)*	0.121 (0.042)*	0.148 (0.034)*
10 Yr. Pre	0.072 (0.022)*	0.172 (0.044)*	0.173 (0.038)*	0.142 (0.032)*
9 Yr. Pre	0.059 (0.021)*	0.216 (0.041)*	0.159 (0.036)*	0.121 (0.030)*
8 Yr. Pre	0.055 (0.020)*	0.167 (0.036)*	0.142 (0.032)*	0.164 (0.028)*
7 Yr. Pre	0.044 (0.019)*	0.115 (0.035)*	0.135 (0.031)*	0.130 (0.028)*
6 Yr. Pre	0.039 (0.018)*	0.128 (0.032)*	0.092 (0.029)*	0.116 (0.026)*
5 Yr. Pre	0.030 (0.019)	0.064 (0.037)	0.081 (0.030)*	0.088 (0.028)*
4 Yr. Pre	0.018 (0.017)	0.059 (0.029)*	0.023 (0.026)	0.077 (0.024)*
3 Yr. Pre	0.005 (0.016)	0.052 (0.027)	0.008 (0.024)	0.049 (0.023)*
2 Yr. Pre	-0.013 (0.016)	0.031 (0.026)	0.003 (0.024)	0.002 (0.023)
1 Yr. Pre	0	0	0	0
Zero Yr.	-0.009 (0.015)	0.045 (0.025)	0.016 (0.023)	-0.030 (0.022)
1 Yr. Post	0.000 (0.015)	0.058 (0.024)*	0.026 (0.023)	0.004 (0.021)

Note. *p < .05; All models include state and year fixed effects and are weighted by the number of students who completed an AP exam; Standard errors in parentheses; All models include state demographic controls: percentage eligible for free- and reduced-price lunch and the percentage of Asian, Black, Hispanic, and White students; The coefficients for 12+ Yrs. Pre include all observations 12 or more years prior to policy implementation and coefficients for 2+ Yrs. Post include all observations 2 or more years post policy implementation. These coefficients are excluded due to being unbalanced—all other coefficients are balanced by including observations from all policy-adopting states; Zero Yr. = year of policy implementation; Reference group for all models includes all observations in all years for comparison states and all observations for policy-adopting states the year prior to policy implementation

APPENDIX E

SKEWNESS AND KURTOSIS STATISTICS

Table E1

Skewness Values

Year	AP ELC	AP ES	AP Psychology	AP Statistics
1997	-0.59 (0.33)		-0.66 (0.34)	0.19 (0.35)
1998	-0.19 (0.33)	0.09 (0.38)	-0.65 (0.34)	-0.38 (0.34)
1999	-0.25 (0.33)	-0.13 (0.37)	-0.15 (0.34)	0.12 (0.34)
2000	-0.41 (0.33)	0.43 (0.37)	-0.39 (0.34)	0.49 (0.34)
2001	-0.41 (0.33)	-0.07 (0.35)	-2.05 (0.33)	-0.11 (0.34)
2002	-0.68 (0.33)	-0.23 (0.35)	-0.50 (0.34)	-0.17 (0.33)
2003	-0.50 (0.33)	0.48 (0.35)	-0.53 (0.33)	0.21 (0.34)
2004	-0.46 (0.33)	-0.70 (0.34)	-0.03 (0.34)	-0.01 (0.34)
2005	-0.47 (0.33)	-0.71 (0.34)	-2.00 (0.33)	-0.86 (0.34)
2006	-0.94 (0.33)	-0.32 (0.34)	-1.54 (0.33)	-0.67 (0.34)
2007	-0.86 (0.33)	-0.59 (0.35)	-1.46 (0.33)	-0.83 (0.34)
2008	-1.32 (0.33)	-1.14 (0.34)	-2.15 (0.33)	-0.88 (0.34)
2009	-0.67 (0.33)	-0.07 (0.35)	-1.25 (0.33)	-0.72 (0.34)
2010	-0.85 (0.33)	-0.56 (0.34)	-1.56 (0.33)	-0.91 (0.34)
2011	-0.76 (0.33)	-0.47 (0.34)	-1.12 (0.33)	-0.72 (0.34)
2012	-0.74 (0.33)	-0.38 (0.34)	-0.60 (0.33)	-0.50 (0.34)
2013	-0.67 (0.33)	-0.36 (0.34)	-0.72 (0.33)	-0.40 (0.34)
2014	-0.73 (0.33)	-0.25 (0.34)	-1.09 (0.33)	-0.47 (0.34)
2015	-0.58 (0.33)	-0.03 (0.34)	-0.73 (0.33)	-0.62 (0.34)
2016	-0.76 (0.33)	-0.38 (0.34)	0.01 (0.33)	-0.60 (0.34)
2017	-0.66 (0.33)	-0.73 (0.34)	-0.91 (0.33)	-0.31 (0.34)
2018	-0.81 (0.33)	0.54 (0.33)	-0.45 (0.33)	-0.73 (0.33)
2019	-0.54 (0.33)	-0.17 (0.33)	-0.42 (0.33)	-0.62 (0.33)

Table E2

Kurtosis Values

Year	AP ELC	AP ES	AP Psychology	AP Statistics
1997	1.08 (0.66)		0.17 (0.66)	-0.45 (0.69)
1998	0.79 (0.66)	0.97 (0.74)	0.12 (0.66)	0.05 (0.66)
1999	0.10 (0.66)	0.87 (0.73)	0.53 (0.66)	0.28 (0.66)
2000	-0.08 (0.66)	0.66 (0.72)	0.13 (0.66)	-0.19 (0.66)
2001	0.47 (0.66)	0.26 (0.69)	8.97 (0.66)	-0.89 (0.66)
2002	0.81 (0.66)	0.15 (0.69)	-0.03 (0.66)	0.78 (0.66)
2003	-0.56 (0.66)	2.84 (0.68)	-0.01 (0.66)	3.01 (0.66)
2004	-0.42 (0.66)	0.34 (0.67)	-0.75 (0.66)	2.00 (0.66)
2005	0.29 (0.66)	0.97 (0.67)	7.03 (0.66)	1.03 (0.66)
2006	0.80 (0.66)	-0.66 (0.67)	4.93 (0.66)	0.48 (0.66)
2007	0.98 (0.66)	-0.31 (0.69)	3.32 (0.66)	1.79 (0.66)
2008	3.75 (0.66)	1.97 (0.66)	7.73 (0.66)	0.94 (0.66)
2009	0.99 (0.66)	0.74 (0.68)	2.20 (0.66)	0.67 (0.66)
2010	1.40 (0.66)	0.58 (0.66)	4.02 (0.66)	0.81 (0.66)
2011	0.92 (0.66)	1.39 (0.66)	1.79 (0.66)	0.08 (0.66)
2012	0.99 (0.66)	-0.73 (0.66)	0.06 (0.66)	-0.37 (0.66)
2013	0.57 (0.66)	-0.42 (0.66)	0.55 (0.66)	-0.59 (0.66)
2014	0.60 (0.66)	-0.43 (0.66)	1.72 (0.66)	-0.35 (0.66)
2015	0.37 (0.66)	-0.14 (0.66)	0.78 (0.66)	-0.49 (0.66)
2016	0.59 (0.66)	-0.43 (0.67)	0.86 (0.66)	-0.27 (0.66)
2017	0.10 (0.66)	0.73 (0.66)	1.17 (0.66)	-0.79 (0.66)
2018	0.17 (0.66)	1.43 (0.66)	0.59 (0.66)	0.51 (0.66)
2019	-0.06 (0.66)	0.05 (0.66)	-0.49 (0.66)	0.05 (0.66)

APPENDIX F

CITS ASSUMPTION TESTING

BASELINE MODELS

Table F1

Levene's Test for Homogeneity of Level 1 Variance

AP ELC		AP ES		AP Psychology		AP Statistics	
DF	F value	DF	F value	DF	F value	DF	F value
1, 1171	3.655	1, 1045	1.429	1, 1165	0.265	1, 1147	0.284

Note. AP ELC = AP English Language and Composition; AP ES = AP Environmental Science

Figure F1

Q-Q Plot for AP English Language and Composition: Baseline Model

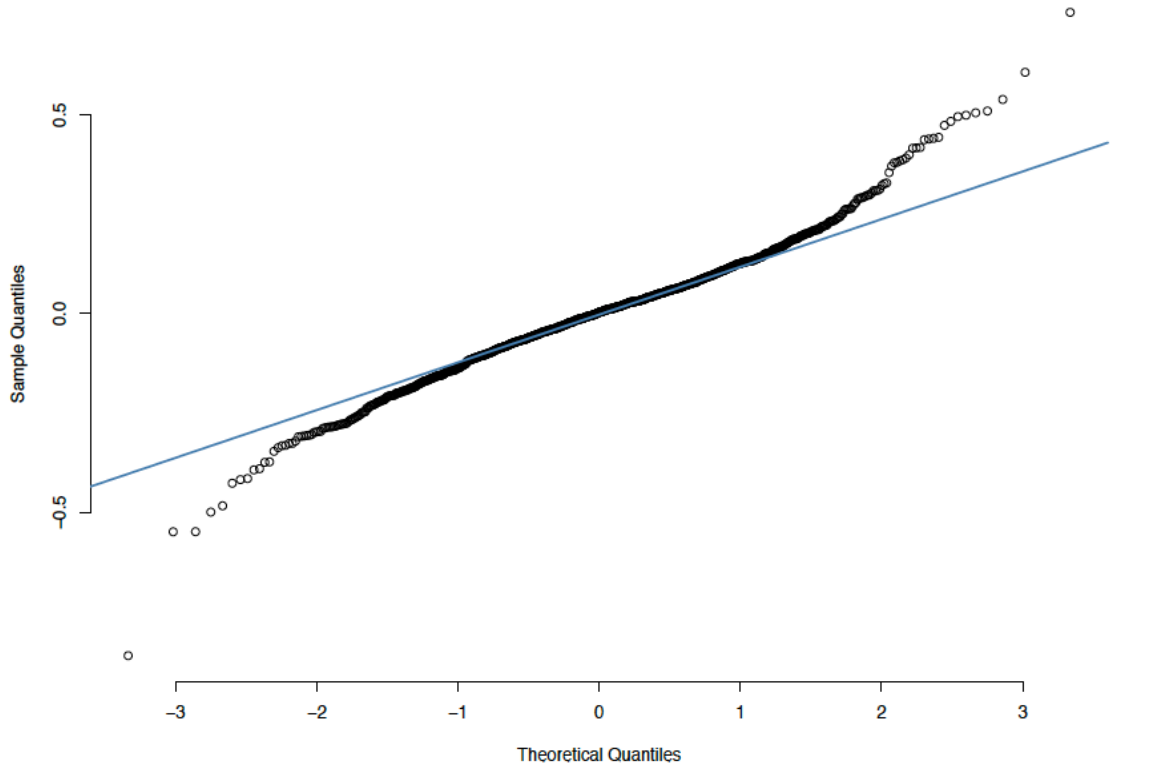


Figure F2

Q-Q Plot for AP Environmental Science: Baseline Model

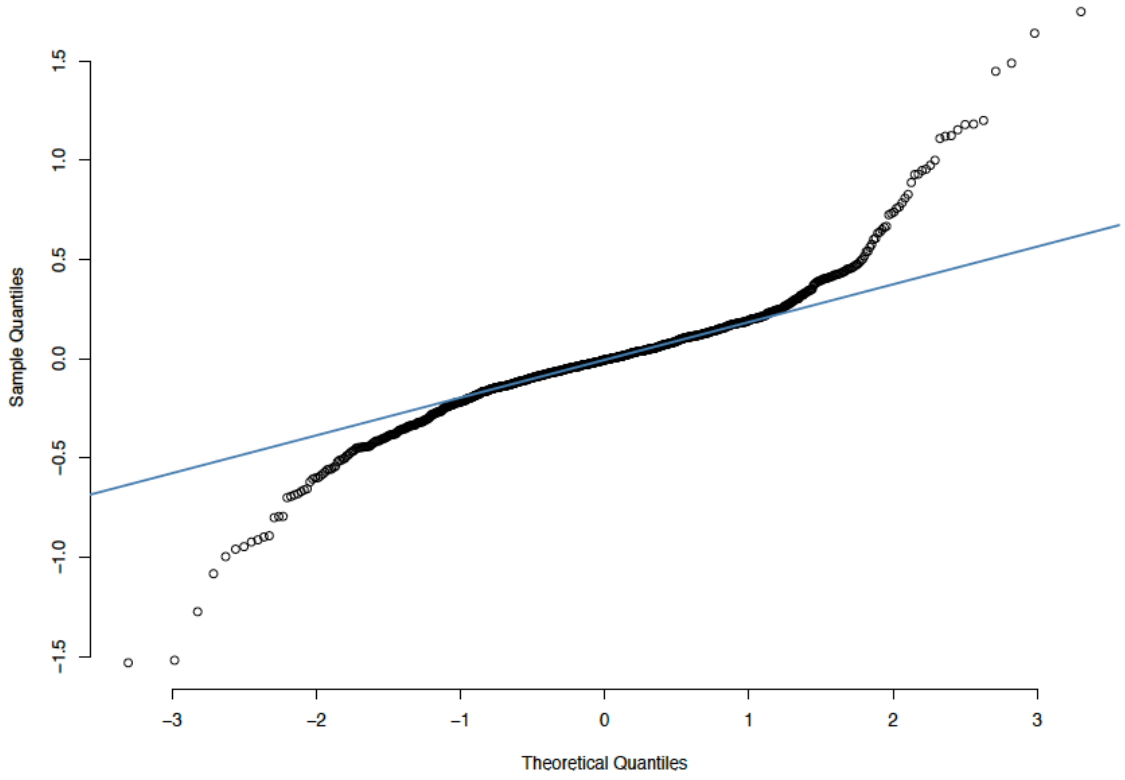


Figure F3

Q-Q Plot for AP Psychology: Baseline Model

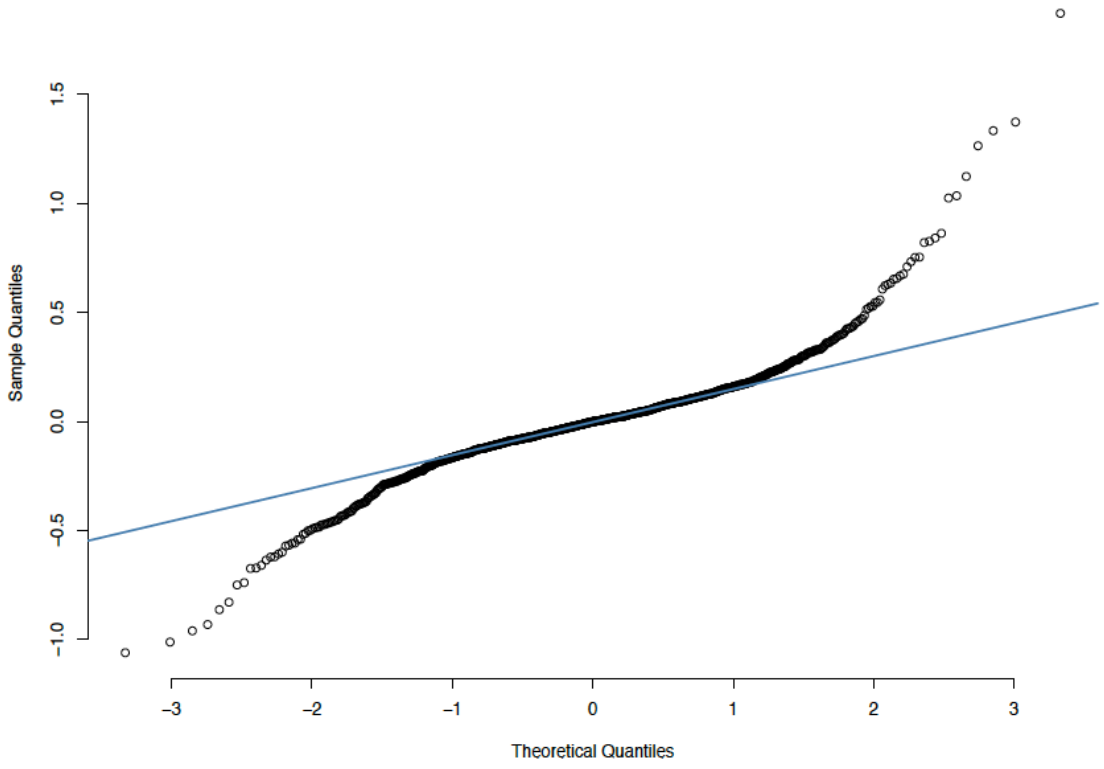


Figure F4

Q-Q Plot for Statistics: Baseline Model

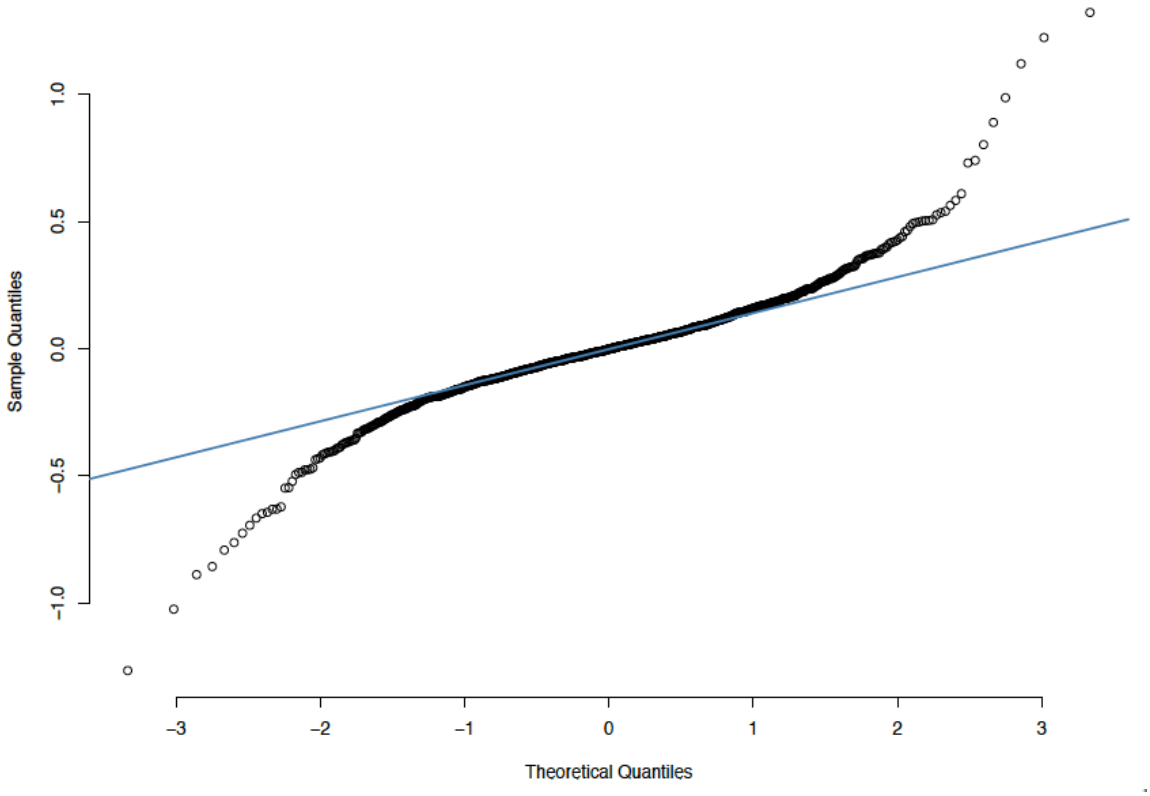


Table F2

Levene's Test for Homogeneity of Level 1 Variance - Weighted

AP ELC		AP ES		AP Psychology		AP Statistics	
DF	F value	DF	F value	DF	F value	DF	F value
1, 1171	10.343*	1, 1045	4.171*	1, 1165	0.678	1, 1147	0.81

Note. AP ELC = AP English Language and Composition; AP ES = AP Environmental Science

Figure F5

Q-Q Plot for AP English Language and Composition: Weighted Baseline

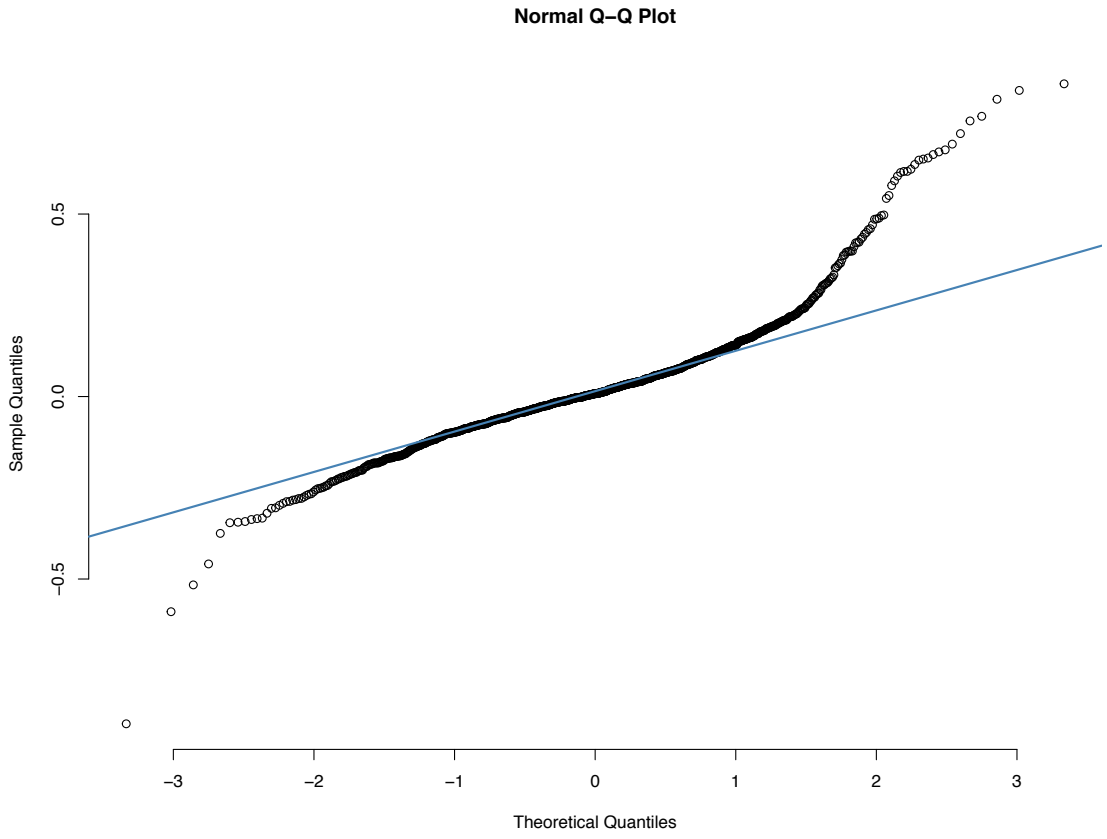


Figure F6

Q-Q Plot for AP Environmental Science: Weighted Baseline

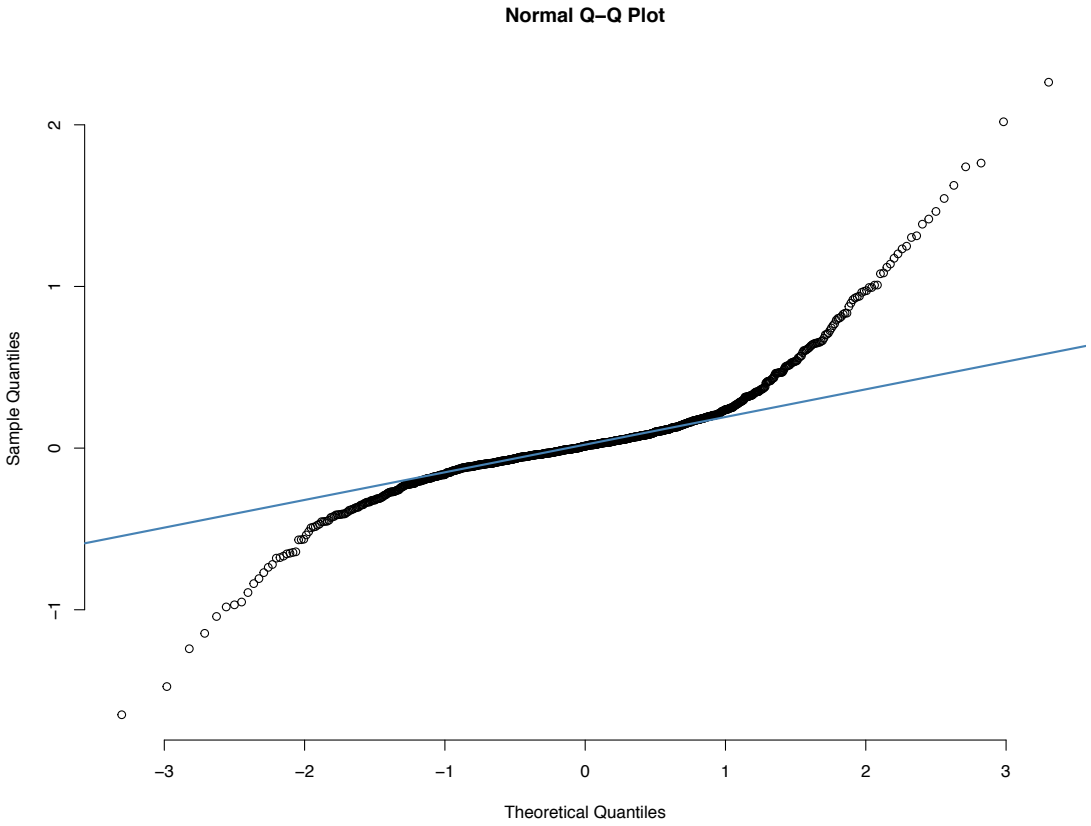


Figure F7

Q-Q Plot for AP Psychology: Weighted Baseline

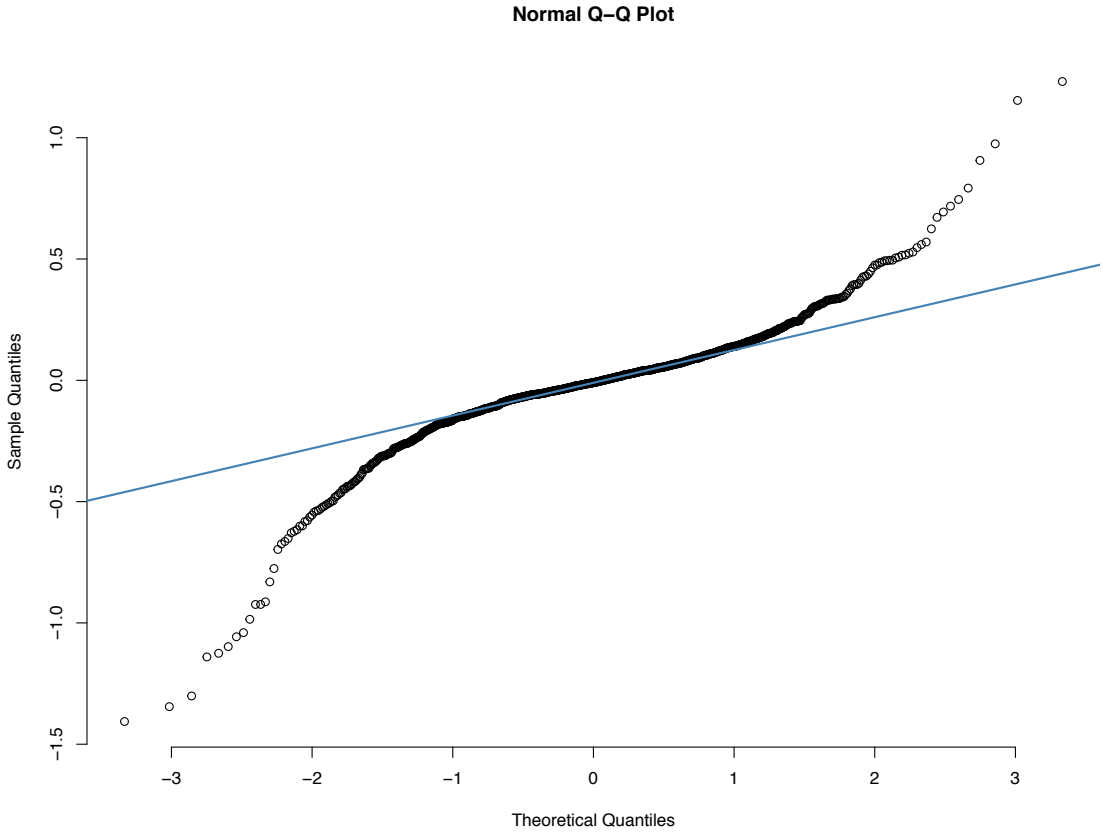
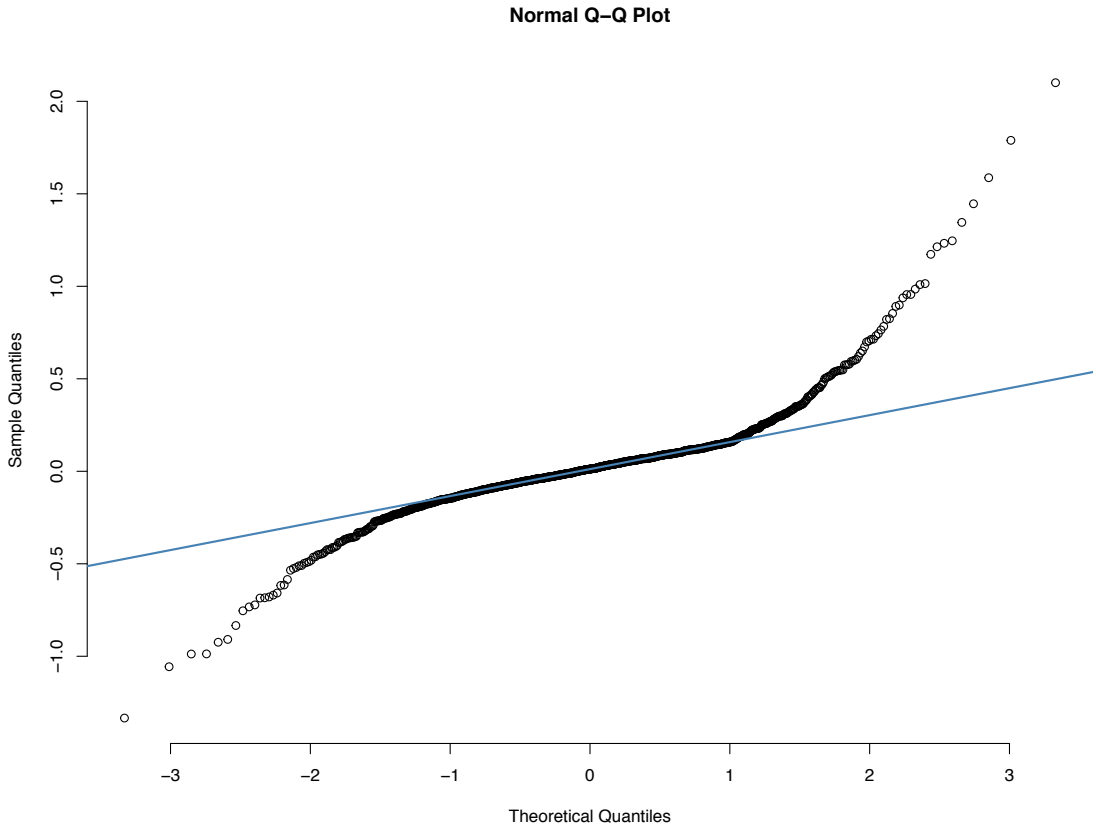


Figure F8

Q-Q Plot for AP Statistics: Weighted Baseline



APPENDIX G

POLICY DOSAGE MODELS

Table G1

CITS Estimates for AP English Language and Composition: Policy Dosage Models

	Model 1	Model 2	Model 3
Intercept[a]	3.048 (0.037)*	3.379 (0.341)*	3.997 (0.417)*
PreTxSlope	-0.011 (0.001)*	-0.013 (0.002)*	-0.013 (0.002)*
PolicyImp	-0.026 (0.033)	-0.028 (0.033)	-0.043 (0.015)*
PostTxSlope	0.008 (0.008)	0.010 (0.007)	0.010 (0.003)*
Grade 11 and 12		0.004 (0.002)*	
SES		0.000 (0.001)	0.000 (0.001)
Asian		-0.004 (0.005)	-0.018 (0.005)*
Black		-0.007 (0.004)	-0.011 (0.004)*
Hispanic		-0.002 (0.004)	-0.004 (0.004)
White		-0.003 (0.003)	-0.010 (0.004)*
TxIdentify	-0.278 (0.089)*	-0.316 (0.096)*	-0.428 (0.110)*
TxIdentify*PreTxSlope	-0.007 (0.002)*	-0.005 (0.003)	0.001 (0.002)
TxIdentify*PolicyImp	0.086 (0.053)	0.073 (0.056)	0.021 (0.022)
TxIdentify*PostTxSlope	0.003 (0.012)	-0.001 (0.012)	-0.001 (0.005)
Participation weights			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.703; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table G2

CITS Estimates for AP Environmental Science: Policy Dosage Models

	Model 1	Model 2	Model 3
Intercept[a]	2.947 (0.081)*	2.589 (0.661)*	3.844 (0.621)*
PreTxSlope[a]	-0.014 (0.005)*	-0.003 (0.005)	-0.006 (0.004)
PolicyImp[a]	-0.090 (0.076)	-0.103 (0.078)	-0.049 (0.043)
PostTxSlope	0.037 (0.017)*	0.045 (0.017)*	0.018 (0.005)*
Grade 11 and 12		0.003 (0.002)	
SES		-0.007 (0.002)*	-0.004 (0.001)*
Asian		0.006 (0.008)	-0.014 (0.008)
Black		-0.004 (0.006)	-0.010 (0.006)
Hispanic		0.003 (0.008)	-0.005 (0.007)
White		0.007 (0.007)	-0.009 (0.006)
TxIdentify	-0.504 (0.195)*	-0.252 (0.192)	-0.390 (0.175)*
TxIdentify*PreTxSlope	-0.001 (0.011)	-0.003 (0.012)	-0.005 (0.008)
TxIdentify*PolicyImp	0.298 (0.138)*	0.277 (0.142)	0.162 (0.074)*
TxIdentify*PostTxSlope	-0.022 (0.027)	-0.030 (0.027)	0.014 (0.008)
Participation weights			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.521; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table G3

CITS Estimates for AP Psychology: Policy Dosage Models

	Model 1	Model 2	Model 3
Intercept[a]	3.195 (0.057)*	2.707 (0.557)*	2.859 (0.530)*
PreTxSlope[a]	-0.005 (0.003)	0.000 (0.004)	-0.010 (0.003)*
PolicyImp	-0.084 (0.060)	-0.079 (0.060)	-0.042 (0.040)
PostTxSlope	0.020 (0.013)	0.022 (0.012)	0.024 (0.005)*
Grade 11 and 12		0.006 (0.002)*	
SES		-0.002 (0.002)	0.001 (0.001)
Asian		0.005 (0.007)	0.003 (0.006)
Black		-0.005 (0.006)	-0.002 (0.005)
Hispanic		-0.001 (0.007)	0.005 (0.006)
White		0.007 (0.006)	0.005 (0.005)
TxIdentify	-0.205 (0.137)	0.033 (0.142)	0.004 (0.142)*
TxIdentify*PreTxSlope	-0.003 (0.007)	-0.010 (0.008)	-0.016 (0.006)
TxIdentify*PolicyImp	0.043 (0.098)	0.030 (0.101)	0.034 (0.065)
TxIdentify*PostTxSlope	-0.020 (0.021)	-0.011 (0.020)	0.012 (0.007)
Participation weights			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.612; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table G4

CITS Estimates for AP Statistics — Policy Dosage Models

	Model 1	Model 2	Model 3
Intercept[a]	2.978 (0.057)*	2.872 (0.521)*	1.908 (0.550)*
PreTxSlope[a]	-0.012 (0.004)*	-0.002 (0.004)	-0.002 (0.004)
PolicyImp[a]	-0.101 (0.060)	-0.096 (0.059)	0.541 (0.579)
PostTxSlope	0.032 (0.015)*	0.029 (0.014)*	0.026 (0.005)*
Grade 11 and 12		0.007 (0.002)*	
SES		-0.005 (0.002)*	-0.002 (0.001)
Asian		-0.006 (0.006)	0.003 (0.007)
Black		-0.006 (0.005)	0.003 (0.005)
Hispanic		-0.004 (0.006)	0.008 (0.006)
White		0.003 (0.005)	0.013 (0.006)*
TxIdentify	-0.322 (0.135)*	-0.154 (0.154)	-0.184 (0.126)
TxIdentify*PreTxSlope	0.003 (0.009)	0.002 (0.010)	0.004 (0.007)
TxIdentify*PolicyImp	0.047 (0.099)	-0.024 (0.101)	-0.578 (0.928)
TxIdentify*PostTxSlope	-0.023 (0.024)	-0.013 (0.023)	0.003 (0.007)
Participation weights			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.523; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

APPENDIX H

DEVIANCE TESTS FOR SENSITIVITY ANALYSES

Table H1

Deviance Tests for Random Intercept, Pre-Policy Slope, and Policy Effects: Policy-Adopting State Only Models

Model	AP ELC		AP ES		AP Psychology		AP Statistics	
	Deviance	ChiSq	Deviance	ChiSq	Deviance	ChiSq	Deviance	ChiSq
A	462.05		189.15		106.03		232.71	
B	557.39	95.34*	133.52	55.63*	54.52	51.51*	203.00	29.71*
C	584.56	27.18*	105.99	27.53*	45.94	8.59*	197.03	5.97
D	606.12	21.55*	89.30	16.69*	39.52	6.42	195.06	1.96

Note. AP ELC = AP English Language and Composition; AP ES = AP Environmental Science; Four additional models with different combination of random effects were tested, each nested within one of the following models. With the exception of AP English Language & Composition, Models A-D produced the best fit of all models. Model A fit better than all other models for AP English Language & Composition; Model A = Random intercept only; Model B = Model A + pre-policy slope; Model C = Model B + post-policy intercept; Model D = Model C + post-policy slope

Table H2

Deviance Tests for Random Intercept, Pre-Policy Slope, and Policy Effects: Private School Student Models

Model	AP ELC		AP ES		AP Psychology		AP Statistics	
	Deviance	ChiSq	Deviance	ChiSq	Deviance	ChiSq	Deviance	ChiSq
A	235.32		1539.6		1605		1486.9	
B	170.22	65.10*	1520.1	19.57*	1517	88.01*	1457.5	29.38*
C	161.06	9.16*	1509.1	10.98*	1515.7	1.33	1451.9	5.64
D	156.41	4.66	1507	2.13	1505	10.74*	1449.8	2.15

Note. AP ELC = AP English Language and Composition; AP ES = AP Environmental Science; Four additional models with different combination of random effects were tested, each nested within one of the following models. Model A = Random intercept only; Model B = Model A + pre-policy slope; Model C = Model B + post-policy intercept; Model D = Model C + post-policy slope; With the exception of AP English Language & Composition, Models A-D produced the best fit of all models. Model A fit better than all other models for AP English Language & Composition

APPENDIX I

AP ACCOUNTABILITY STATE ONLY MODELS

EVENT STUDY ESTIMATES

Table I1

Event Study Estimates: Policy-Adopting State Only Models

	AP ELC	AP ES	AP Psychology	AP Statistics
11 Yr. Pre	0.111 (0.037)*	0.368 (0.063)*	0.328 (0.064)*	0.173 (0.055)*
10 Yr. Pre	0.124 (0.034)*	0.393 (0.057)*	0.376 (0.058)*	0.173 (0.051)*
9 Yr. Pre	0.104 (0.032)*	0.393 (0.052)*	0.341 (0.053)*	0.164 (0.047)*
8 Yr. Pre	0.106 (0.029)*	0.304 (0.045)*	0.311 (0.047)*	0.189 (0.042)*
7 Yr. Pre	0.084 (0.026)*	0.258 (0.042)*	0.293 (0.043)*	0.162 (0.039)*
6 Yr. Pre	0.097 (0.024)*	0.249 (0.037)*	0.214 (0.038)*	0.151 (0.034)*
5 Yr. Pre	0.070 (0.024)*	0.168 (0.040)*	0.184 (0.038)*	0.112 (0.035)*
4 Yr. Pre	0.053 (0.020)*	0.120 (0.031)*	0.100 (0.032)*	0.098 (0.029)*
3 Yr. Pre	0.030 (0.018)	0.103 (0.028)*	0.065 (0.028)	0.067 (0.026)*
2 Yr. Pre	-0.001 (0.018)	0.045 (0.027)	0.029 (0.028)	0.007 (0.026)
1 Yr. Pre	0	0	0	0
Zero Yr.	-0.019 (0.018)	0.019 (0.026)	-0.008 (0.027)	-0.040 (0.025)
1 Yr. Post	-0.017 (0.017)	0.012 (0.025)	-0.029 (0.026)	-0.013 (0.024)

Note. *p < .05; All models include state and year fixed effects; Standard errors in parentheses; [a] Models weighted by the number of students who completed an AP exam; State demographic controls include the percentage eligible for free- and reduced-price lunch, and the percentage of Asian, Black, Hispanic, and White students; Coefficients for 12+ Yrs. Pre include all observations 12 or more years prior to policy implementation and coefficients for 2+ Yrs. Post include all observations 2 or more years post policy implementation—all other coefficients are balanced by including observations from all policy-adopting states; Zero Yr. = year of policy implementation; Reference group for all models includes all observations in all years for comparison states and all observations for policy-adopting states the year prior to policy implementation

Table I2

Difference-in-Differences Estimates for AP English Language and Composition: Policy-Adopting State Only Models

	Model 1	Model 2	Model 3
Intercept	4.177 (0.057)*	-12.596 (7.500)	-2.757 (9.471)
TxTrend	0.073 (0.003)*	-0.993 (0.468)	-0.359 (0.516)
PolicyImp	0.029 (0.037)	-0.024 (0.026)	-0.013 (0.014)
PolicyImp*TxTrend	0.011 (0.010)	-0.009 (0.014)	0.013 (0.006)
Grade 11 and 12		0.005 (0.007)	
SES		0.001 (0.002)	0.001 (0.002)
Asian		0.001 (0.071)	-0.076 (0.029)
Black		-0.004 (0.007)	-0.004 (0.018)
Hispanic		-0.025 (0.015)	-0.005 (0.012)
White		-0.037 (0.016)	-0.013 (0.013)
Participation weights[a]			X
R-Squared	0.862	0.887	0.939

Note. *p < .05; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at zero in treated states the year of policy implementation; [a] Models weighted by the number of students who completed an AP exam; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table I3

Difference-in-Differences Estimates for AP Environmental Science: Policy-Adopting State Only Models

	Model 1	Model 2	Model 3
Intercept	9.142 (1.593)*	3.949 (19.09)	17.700 (16.45)
TxTrend	0.350 (0.086)*	-0.029 (1.161)	0.799 (0.925)
PolicyImp	0.094 (0.062)	0.053 (0.061)	0.046 (0.027)
PolicyImp*TxTrend	0.025 (0.014)	0.019 (0.023)	0.044 (0.008)
Grade 11 and 12		-0.001 (0.015)	
SES		-0.003 (0.005)	0.000 (0.002)
Asian		-0.037 (0.103)	-0.024 (0.042)
Black		-0.007 (0.013)	-0.026 (0.025)
Hispanic		0.014 (0.032)	0.020 (0.018)
White		-0.016 (0.015)	0.003 (0.021)
Participation weights[a]			X
R-Squared	0.613	0.613	0.868

Note. *p < .05; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at zero in treated states the year of policy implementation; [a] Models weighted by the number of students who completed an AP exam; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table I4

Difference-in-Differences Estimates for AP Psychology: Policy-Adopting State Only Models

	Model 1	Model 2	Model 3
Intercept	10.667 (0.088)*	23.300 (37.01)	25.828 (17.66)
TxTrend	0.395 (0.005)*	0.956 (1.958)	1.447 (1.036)
PolicyImp	-0.047 (0.057)	-0.064 (0.047)	-0.005 (0.035)
PolicyImp*TxTrend	0.013 (0.021)	0.019 (0.025)	0.037 (0.017)
Grade 11 and 12		0.015 (0.009)	
SES		0.001 (0.004)	0.004 (0.002)
Asian		-0.023 (0.104)	-0.048 (0.060)
Black		-0.045 (0.071)	0.004 (0.022)
Hispanic		-0.038 (0.037)	0.036 (0.015)
White		-0.023 (0.045)	0.045 (0.027)
Participation weights[a]			X
R-Squared	0.719	0.719	0.890

Note. *p < .05; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at zero in treated states the year of policy implementation; [a] Models weighted by the number of students who completed an AP exam; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table I5

Difference-in-Differences Estimates for AP Statistics: Policy-Adopting State Only Models

	Model 1	Model 2	Model 3
Intercept	3.823 (0.152)*	-7.984 (18.14)	7.285 (16.36)
TxTrend	0.050 (0.005)*	-0.637 (1.069)	0.384 (0.959)
PolicyImp	-0.001 (0.058)	-0.024 (0.051)	-0.026 (0.025)
PolicyImp*TxTrend	0.031 (0.021)	0.036 (0.021)	0.022 (0.015)
Grade 11 and 12		-0.002 (0.007)	
SES		0.002 (0.005)	0.003 (0.003)
Asian		-0.040 (0.096)	-0.024 (0.049)
Black		0.007 (0.016)	0.001 (0.019)
Hispanic		0.010 (0.018)	0.029 (0.017)
White		-0.014 (0.026)	0.027 (0.024)
Weights[a]			X
R-Squared	0.552	0.597	0.916

Note. *p < .05; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at zero in treated states the year of policy implementation; [a] Models weighted by the number of students who completed an AP exam; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table I6

CITS Estimates for AP English Language and Composition: Policy-Adopting State Only Models

	Model 1	Model 2	Model 3
Intercept[a]	2.882 (0.068)*	4.986 (0.526)*	5.240 (0.766)*
PreTxSlope	-0.016 (0.001)*	-0.024 (0.002)*	-0.019 (0.003)*
PolicyImp	0.032 (0.024)	0.009 (0.024)	-0.010 (0.015)
PostTxSlope	0.011 (0.005)*	0.005 (0.005)	0.015 (0.003)*
Grade 11 and 12		0.002 (0.003)	
SES		0.000 (0.001)	0.000 (0.001)
Asian		-0.001 (0.022)	-0.081 (0.016)*
Black		-0.014 (0.005)*	-0.021 (0.009)*
Hispanic		-0.017 (0.007)*	-0.014 (0.008)
White		-0.026 (0.005)*	-0.024 (0.008)*
Weights[b]			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.779; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b]Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table I7

CITS Estimates for AP Environmental Science: Policy-Adopting State Only Models

	Model 1	Model 2	Model 3
Intercept[a]	2.670 (0.133)*	1.740 (0.831)*	3.894 (0.924)*
PreTxSlope[a]	-0.020 (0.008)*	-0.014 (0.009)	-0.013 (0.007)
PolicyImp[a]	0.126 (0.069)	0.093 (0.066)	0.041 (0.039)
PostTxSlope[a]	0.029 (0.016)	0.032 (0.016)*	0.030 (0.001)*
Grade 11 and 12		0.002 (0.003)	
SES		-0.004 (0.003)	-0.005 (0.002)*
Asian		0.019 (0.028)	-0.022 (0.022)
Black		0.002 (0.008)	-0.015 (0.009)
Hispanic		0.012 (0.010)	-0.003 (0.010)
White		0.013 (0.008)	-0.011 (0.009)
Weights[b]			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.521; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b]Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table I8

CITS Estimates for AP Psychology: Policy-Adopting State Only Models

	Baseline	Covariate-Adjusted	Weighted
Intercept[a]	3.152 (0.124)*	4.020 (1.303)*	3.968 (1.244)*
PreTxSlope[a]	-0.011 (0.005)*	-0.010 (0.007)	-0.010 (0.007)
PolicyImp[a]	-0.050 (0.058)	-0.053 (0.059)	-0.053 (0.059)
PostTxSlope	0.016 (0.010)	0.017 (0.011)	0.017 (0.011)
Grade 11 and 12		0.000 (0.004)	
SES		-0.001 (0.003)	-0.001 (0.003)
Asian		0.036 (0.034)	0.036 (0.026)
Black		-0.021 (0.013)	-0.020 (0.012)
Hispanic		-0.013 (0.015)	-0.013 (0.014)
White		-0.005 (0.013)	-0.004 (0.013)
Weights[b]			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.693; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b] Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table I9

CITS Estimates for AP Statistics: Policy-Adopting State Only Models

	Model 1	Model 2	Model 3
Intercept[a]	2.849 (0.083)*	1.693 (0.928)	0.595 (0.934)
PreTxSlope[a]	-0.012 (0.005)*	-0.002 (0.006)	-0.008 (0.004)*
PolicyImp	-0.066 (0.052)	-0.090 (0.054)	-0.062 (0.017)*
PostTxSlope	0.025 (0.013)	0.024 (0.013)	0.029 (0.004)*
Grade 11 and 12		0.003 (0.004)	
SES		-0.004 (0.003)	0.002 (0.002)
Asian		0.041 (0.029)	0.018 (0.021)
Black		0.005 (0.009)	0.012 (0.010)
Hispanic		0.005 (0.011)	0.023 (0.010)*
White		0.015 (0.009)	0.024 (0.009)*
Weights[b]			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.522; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting state states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b] Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

APPENDIX J

PRIVATE STUDENT MODELS

EVENT STUDY ESTIMATES

Table J1

Event Study Estimates: Private School Students Only

	AP ELC	AP ES	AP Psychology	AP Statistics
11 Yr. Pre	-0.001 (0.034)	0.242 (0.087)*	-0.100 (0.079)	0.025 (0.061)
10 Yr. Pre	0.164 (0.065)*	0.300 (0.154)	-0.083 (0.151)	0.130 (0.123)
9 Yr. Pre	-0.031 (0.032)	0.179 (0.081)*	0.038 (0.071)	-0.007 (0.058)
8 Yr. Pre	0.108 (0.061)	0.229 (0.140)	0.131 (0.145)	-0.162 (0.114)
7 Yr. Pre	-0.035 (0.030)	0.047 (0.071)	0.081 (0.065)	0.021 (0.054)
6 Yr. Pre	0.065 (0.055)	0.133 (0.123)	0.212 (0.122)	-0.124 (0.099)
5 Yr. Pre	-0.004 (0.028)	0.062 (0.064)	-0.017 (0.058)	0.019 (0.050)
4 Yr. Pre	0.081 (0.051)	-0.004 (0.118)	0.051 (0.106)	-0.037 (0.097)
3 Yr. Pre	-0.027 (0.026)	-0.022 (0.061)	-0.035 (0.053)	-0.043 (0.046)
2 Yr. Pre	0.086 (0.047)	0.116 (0.110)	0.071 (0.098)	-0.149 (0.090)
1 Yr. Pre	0	0	0	0
Zero Yr.	0.030 (0.043)	0.072 (0.102)	0.039 (0.086)	-0.115 (0.08)
1 Yr. Post	0.030 (0.025)	-0.075 (0.057)	0.050 (0.049)	-0.020 (0.044)

Note. *p < .05; All models include state and year fixed effects; Standard errors in parentheses; [a] Models weighted by the number of students who completed an AP exam; State demographic controls include the percentage eligible for free- and reduced-price lunch, and the percentage of Asian, Black, Hispanic, and White students; Coefficients for 12+ Yrs. Pre include all observations 12 or more years prior to policy implementation and coefficients for 2+ Yrs. Post include all observations 2 or more years post policy implementation—all other coefficients are balanced by including observations from all policy-adopting states; Zero Yr. = year of policy implementation; Reference group for all models includes all observations in all years for comparison states and all observations for policy-adopting states the year prior to policy implementation

Table J2

Difference-in-Differences Estimates for AP English Language and Composition: Private School Students Only

	Model 1	Model 2	Model 3
Intercept	3.118 (0.079)*	2.950 (0.215)*	3.511 (0.359)*
TxTrend	0.003 (0.004)	-0.001 (0.003)	0.001 (0.003)
PolicyImp	-0.014 (0.030)	0.009 (0.042)	0.027 (0.046)
PolicyImp*TxTrend	-0.007 (0.013)	-0.013 (0.011)	-0.004 (0.009)
Grade 11 and 12			
SES			
Asian		-0.023 (0.005)	0.002 (0.001)
Black		0.045 (0.019)*	-0.023 (0.007)*
Hispanic		0.002 (0.003)*	0.017 (0.012)
White		0.000 (0.006)	-0.007 (0.006)
Weights[a]			X
R-Squared	0.558	0.558	0.719

Note. *p < .05; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at zero in treated states the year of policy implementation; [a] Models weighted by the number of students who completed an AP exam; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table J3

Difference-in-Differences Estimates for AP Environmental Science: Private School Students Only

	Model 1	Model 2	Model 3
Intercept	3.602 (0.136)*	4.392 (0.782)*	3.771 (0.413)
TxTrend	-0.007 (0.010)	-0.008 (0.012)	-0.026 (0.017)
PolicyImp	0.079 (0.089)	0.143 (0.130)	-0.008 (0.083)
PolicyImp*TxTrend	-0.014 (0.021)	-0.009 (0.023)	0.032 (0.021)
Asian		-0.003 (0.003)	-0.001 (0.003)
Black		-0.006 (0.013)	-0.003 (0.012)
Hispanic		0.027 (0.034)	0.012 (0.011)
White		-0.005 (0.015)	0.002 (0.007)
Weights[a]			X
R-Squared	0.250	0.284	0.597

Note. *p < .05; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at zero in treated states the year of policy implementation; [a] Models weighted by the number of students who completed an AP exam; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table J4

Difference-in-Differences Estimates for AP Psychology: Private School Students Only

	Model 1	Model 2	Model 3
Intercept	3.993 (0.132)	4.305 (0.890)*	4.372 (0.547)
TxTrend	0.003 (0.012)	0.000 (0.016)	0.005 (0.007)
PolicyImp	-0.098 (0.073)	-0.118 (0.102)	0.023 (0.057)
PolicyImp*TxTrend	-0.004 (0.019)	0.008 (0.033)	-0.007 (0.012)
Asian		0.001 (0.015)	0.000 (0.003)
Black		-0.001 (0.017)	-0.014 (0.015)
Hispanic		-0.040 (0.026)	-0.002 (0.020)
White		0.002 (0.014)	0.001 (0.009)
Weights[a]			X
R-Squared	0.226	0.173	0.514

Note. *p < .05; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at zero in treated states the year of policy implementation; [a] Models weighted by the number of students who completed an AP exam; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table J5

Difference-in-Differences Estimates for AP Statistics: Private School Students Only

	Model 1	Model 2	Model 3
Intercept	3.548 (0.157)*	2.984 (0.743)*	3.106 (0.373)
TxTrend	0.010 (0.010)	0.013 (0.009)	0.000 (0.006)
PolicyImp	-0.149 (0.067)*	-0.163 (0.088)	-0.020 (0.035)
PolicyImp*TxTrend	-0.012 (0.014)	-0.006 (0.018)	-0.010 (0.013)
Asian		0.004 (0.018)	0.008 (0.007)
Black		0.001 (0.014)	-0.004 (0.007)
Hispanic		0.014 (0.021)	0.009 (0.012)
White		0.001 (0.013)	0.002 (0.006)
Weights[a]			X
R-Squared	0.275	0.226	0.500

Note. *p < .05; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; TxTrend is coded -1 in all years for all comparison states and centered at zero in treated states the year of policy implementation; [a] Models weighted by the number of students who completed an AP exam; Standard errors clustered at the state-level in parentheses; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table J6

CITS Estimates for AP English Language and Composition: Private School Students Only

	Model 1	Model 2	Model 3
Intercept[a]	3.325 (0.042)*	2.738 (0.207)*	2.694 (0.182)*
PreTxSlope	-0.001 (0.003)	0.003 (0.002)	0.002 (0.002)
PolicyImp	-0.021 (0.045)	-0.036 (0.057)	-0.002 (0.030)
PostTxSlope	-0.014 (0.010)	-0.014 (0.014)	-0.011 (0.006)
Asian		0.006 (0.004)	0.010 (0.003)*
Black		0.002 (0.005)	0.003 (0.004)
Hispanic		0.013 (0.006)*	0.008 (0.005)
White		0.007 (0.003)*	0.008 (0.002)*
Weights[a]			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.380; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b]Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table J7

CITS Estimates for AP Environmental Science: Private School Students Only

	Model 1	Model 2	Model 3
Intercept[a]	2.989 (0.061)*	3.935 (0.481)*	3.533 (0.278)*
PreTxSlope[a]	-0.003 (0.005)	-0.007 (0.007)	-0.010 (0.004)*
PolicyImp[a]	-0.071 (0.094)	-0.027 (0.131)	-0.088 (0.039)*
PostTxSlope	-0.015 (0.021)	-0.021 (0.030)	0.014 (0.010)
Asian		-0.002 (0.009)	0.000 (0.005)
Black		-0.012 (0.010)	-0.001 (0.007)
Hispanic		-0.021 (0.011)	-0.014 (0.007)
White		-0.011 (0.006)	-0.007 (0.004)*
Weights[b]			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.305; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b]Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table J8

CITS Estimates for AP Psychology: Private School Students Only

	Model 1	Model 2	Model 3
Intercept[a]	3.289 (0.083)*	3.768 (0.421)*	3.676 (0.253)*
PreTxSlope[a]	0.000 (0.005)	0.000 (0.007)	-0.011 (0.003)*
PolicyImp	-0.112 (0.080)	-0.110 (0.127)	-0.012 (0.038)
PostTxSlope[a]	-0.012 (0.020)	-0.009 (0.032)	-0.010 (0.009)
Asian		0.002 (0.009)	0.002 (0.004)
Black		0.005 (0.009)	0.001 (0.004)
Hispanic		-0.013 (0.010)	-0.015 (0.005)*
White		-0.007 (0.005)	-0.002 (0.003)
Weights[b]			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.235; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b]Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

Table J9

CITS Estimates for AP Statistics: Private School Students Only

	Model 1	Model 2	Model 3
Intercept[a]	2.873 (0.067)*	2.828 (0.358)*	2.868 (0.221)*
PreTxSlope[a]	0.008 (0.004)	0.012 (0.004)*	0.011 (0.003)*
PolicyImp	-0.134 (0.078)	-0.102 (0.117)	-0.041 (0.032)
PostTxSlope	-0.002 (0.020)	-0.001 (0.030)	-0.015 (0.008)
Asian		0.007 (0.007)	0.004 (0.004)
Black		0.007 (0.008)	0.006 (0.005)
Hispanic		-0.002 (0.009)	-0.001 (0.006)
White		0.000 (0.004)	-0.001 (0.003)
Weights[b]			X

Note. *p < .05; ICC = Intraclass coefficient; ICC = 0.396; PreTxSlope is coded 0 in 1996-97; 1 in 1997-98 … and 22 in 2018-19; PolicyImp is coded 0 in all years for all comparison states and yet-to-implement policy adopting states in the years prior to policy implementation and 1 for policy-adopting states in every year of policy implementation; PostTxSlope is an interaction between PreTxSlope and PolicyImp; [a] Level 2 random effects; [b]Models are weighted by the number of students who completed an AP exam; Grade 11 and 12 students divided by 10,000 for scaling purposes; SES = percentage eligible for free- and reduced-price lunch; Asian, Black, Hispanic, and White variables are the number of students in each respective group divided by the number of Grades 11 and 12 students

APPENDIX K

TRENDS IN AP EXAM PERFORMANCE FOR POLICY-ADOPTING STATES

Figure K1

State by State Trajectories for AP English Language and Composition: Group 1

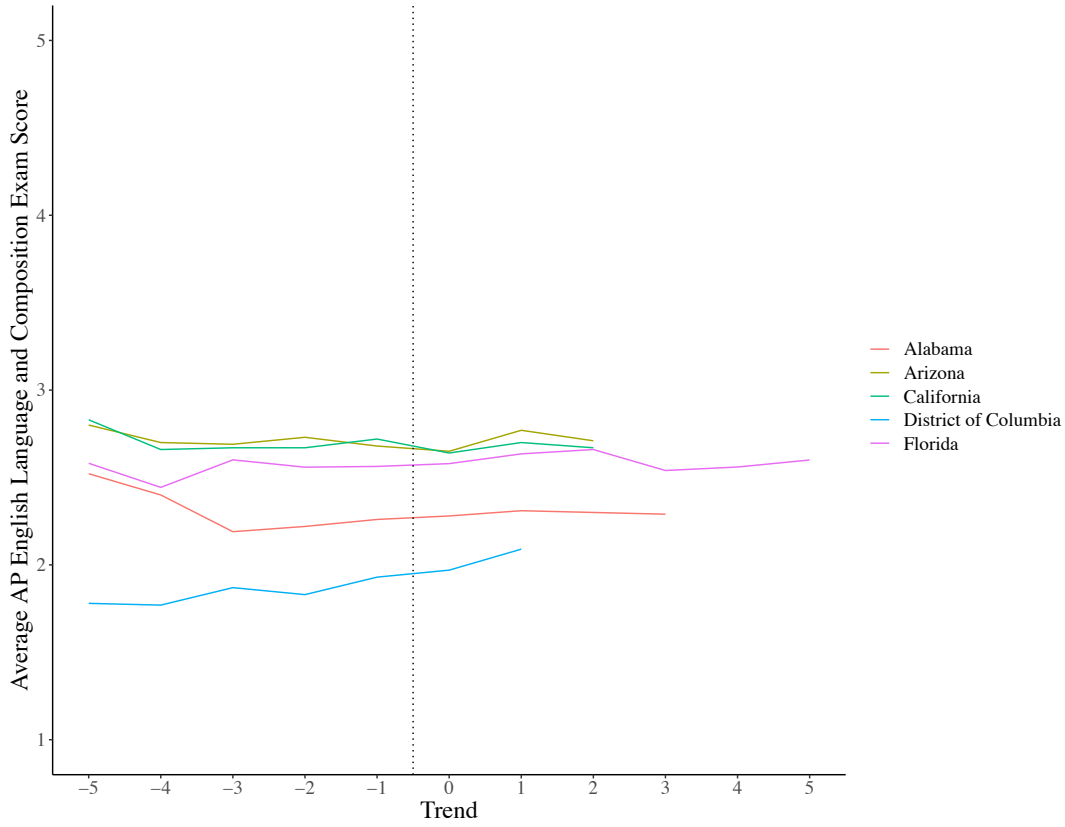


Figure K2

State by State Trajectories for AP English Language and Composition: Group 2

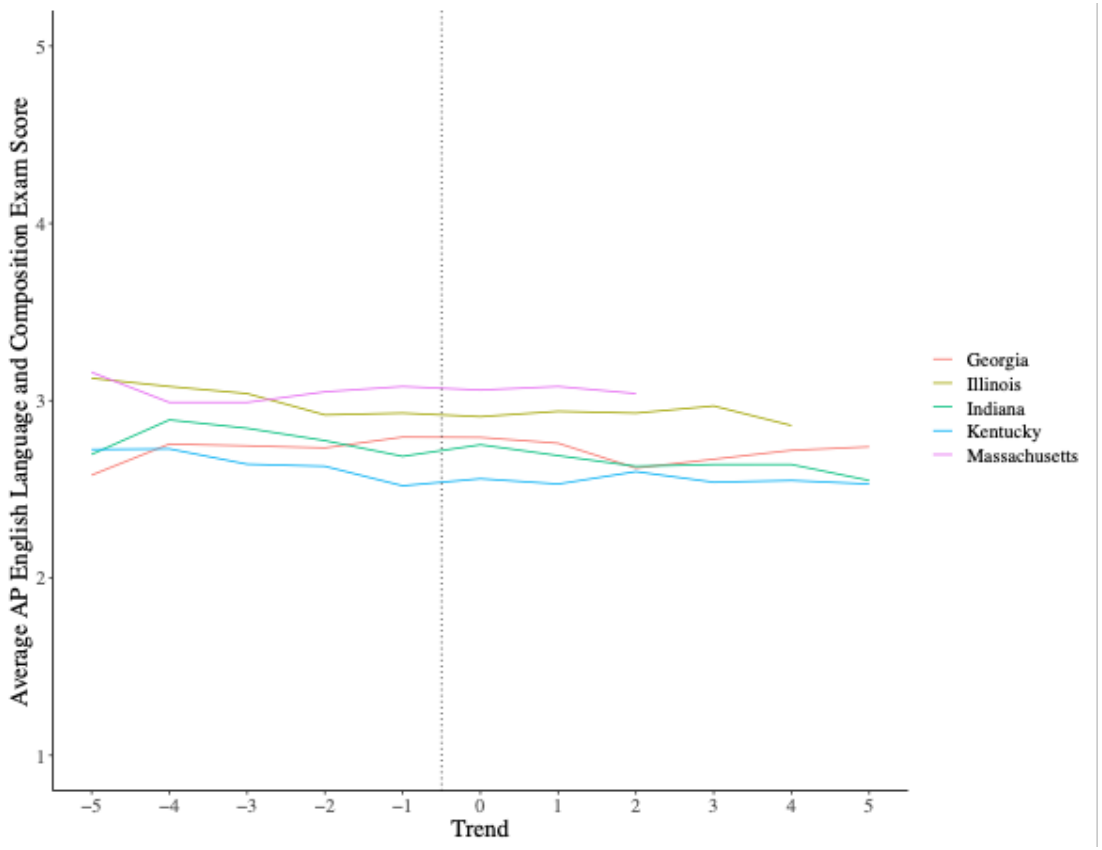


Figure K3

State by State Trajectories for AP English Language and Composition: Group 3

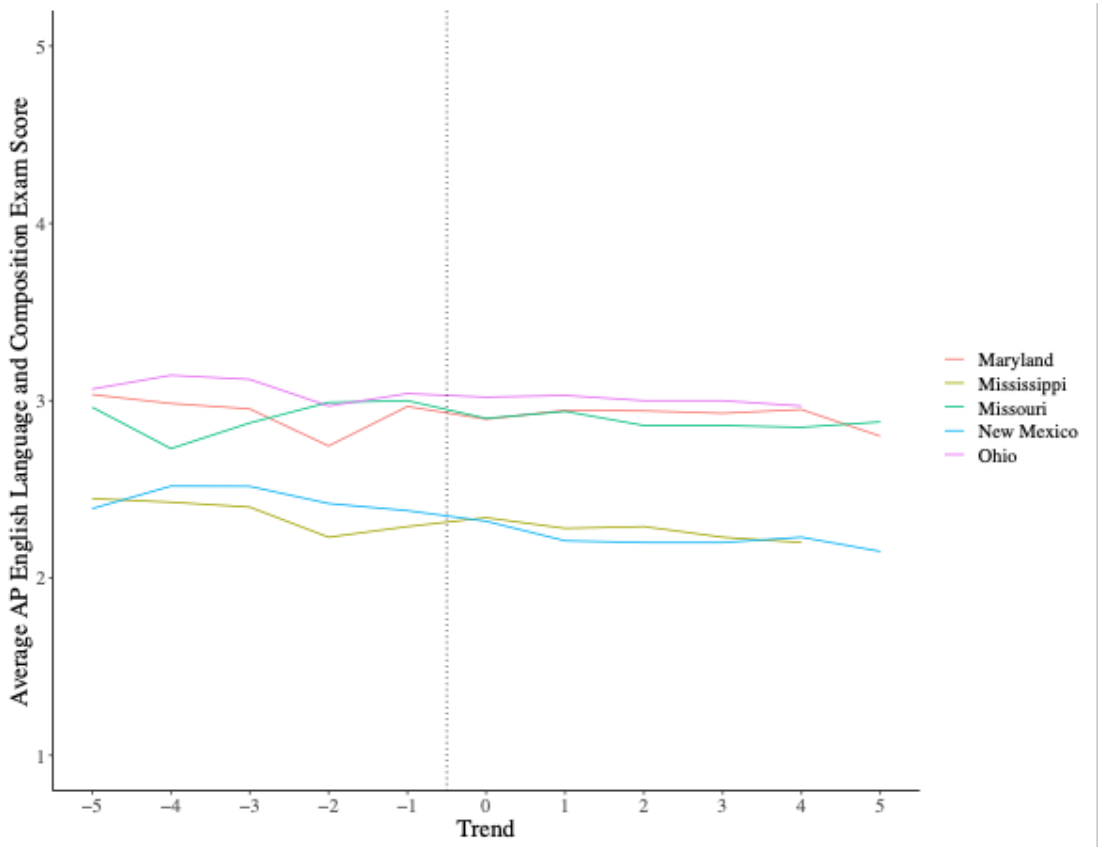


Figure K4

State by State Trajectories for AP English Language and Composition: Group 4

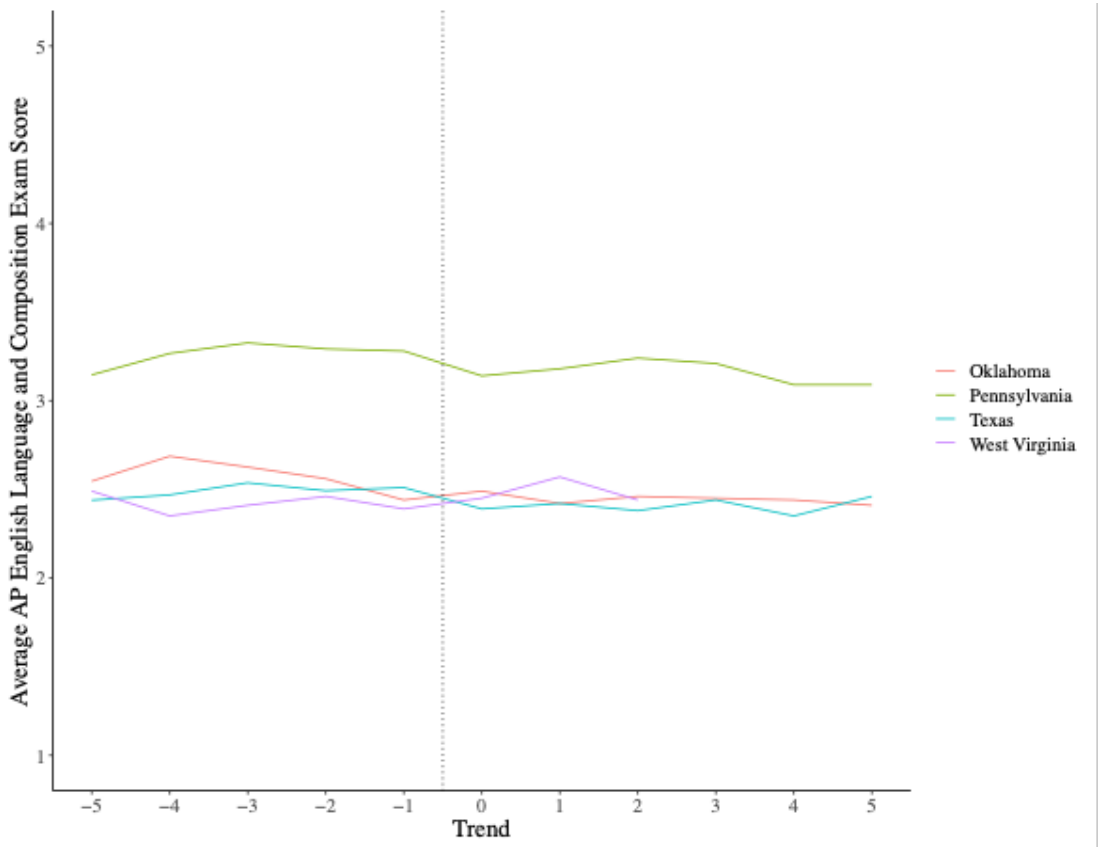


Figure K5

State by State Trajectories for AP Environmental Science: Group 1

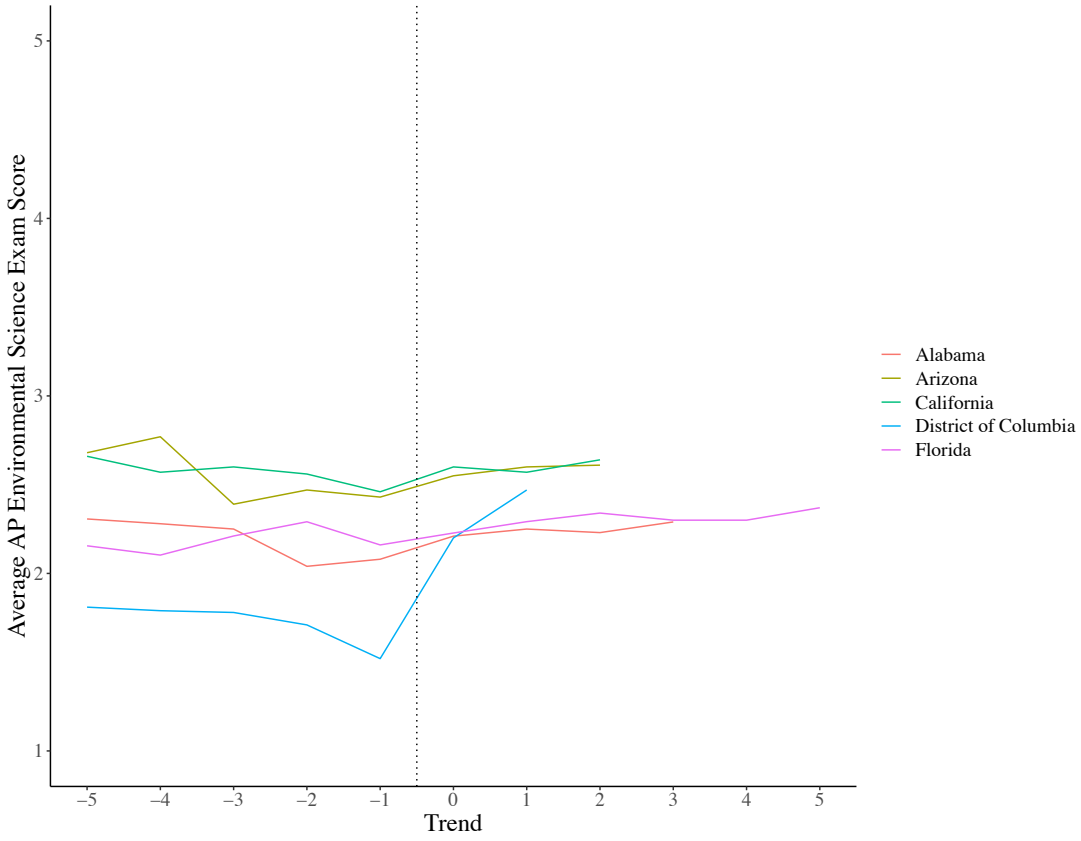


Figure K6

State by State Trajectories for AP Environmental Science: Group 2

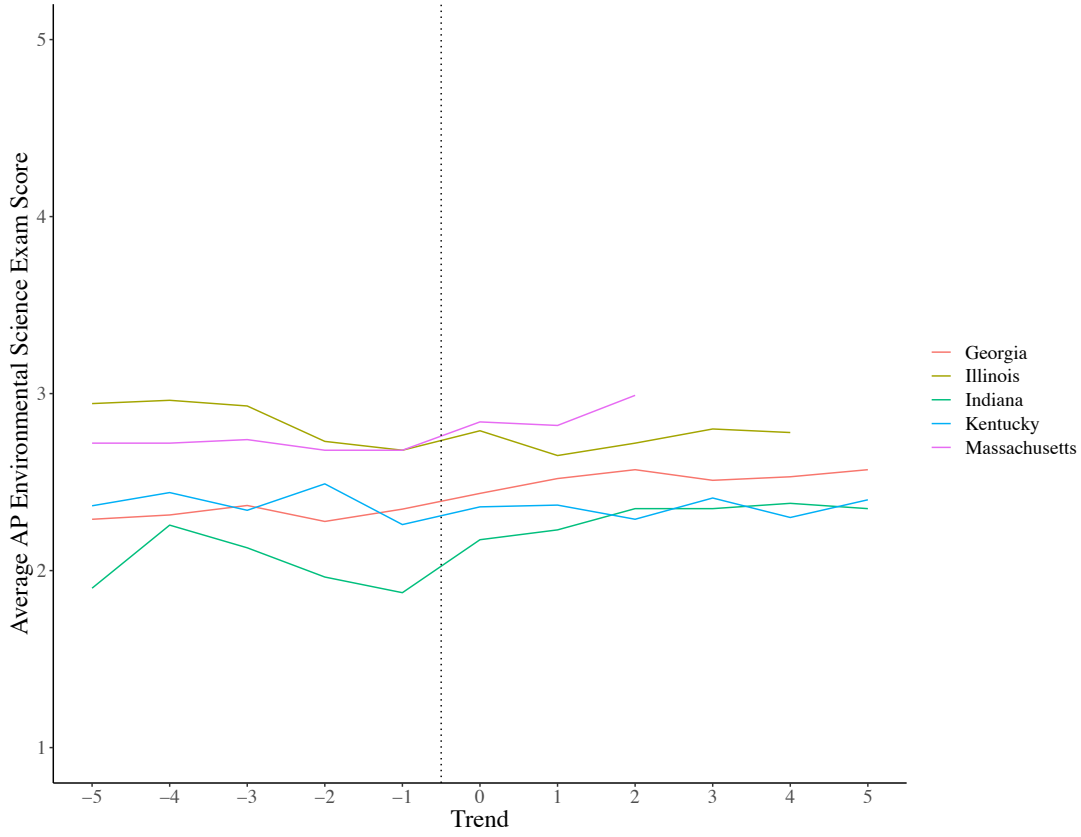


Figure K7

State by State Trajectories for AP Environmental Science: Group 3

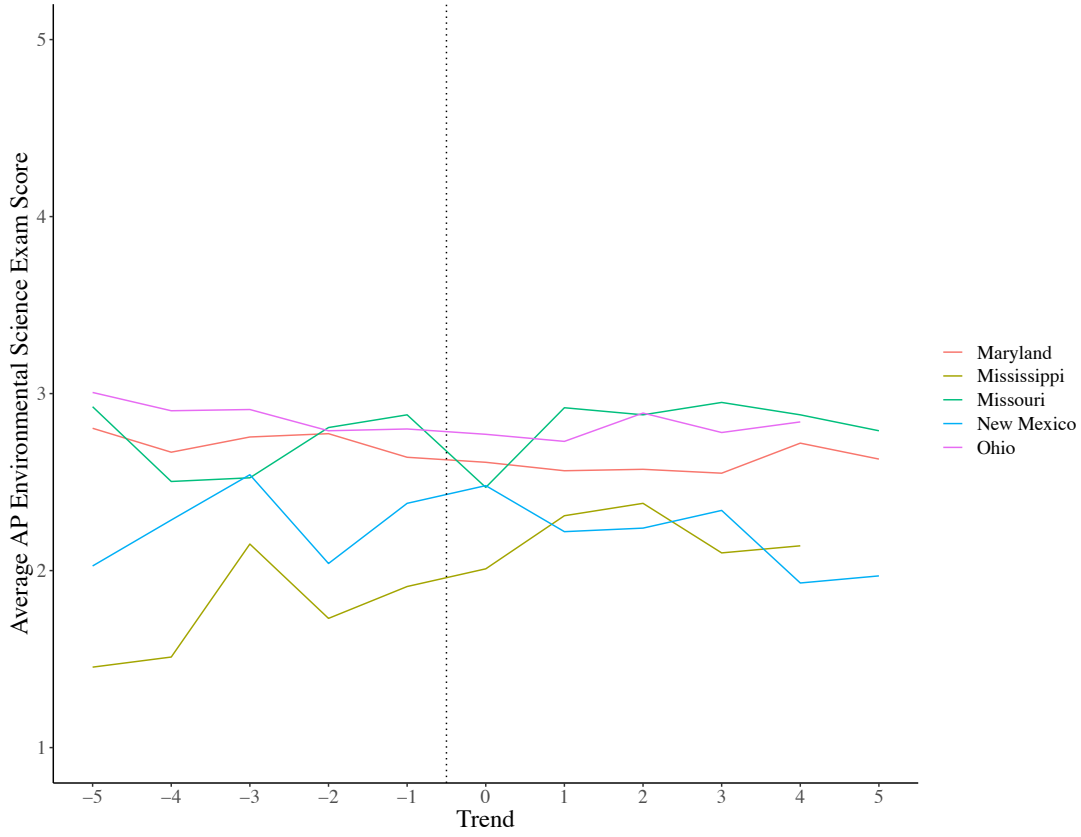


Figure K8

State by State Trajectories for AP Environmental Science: Group 4

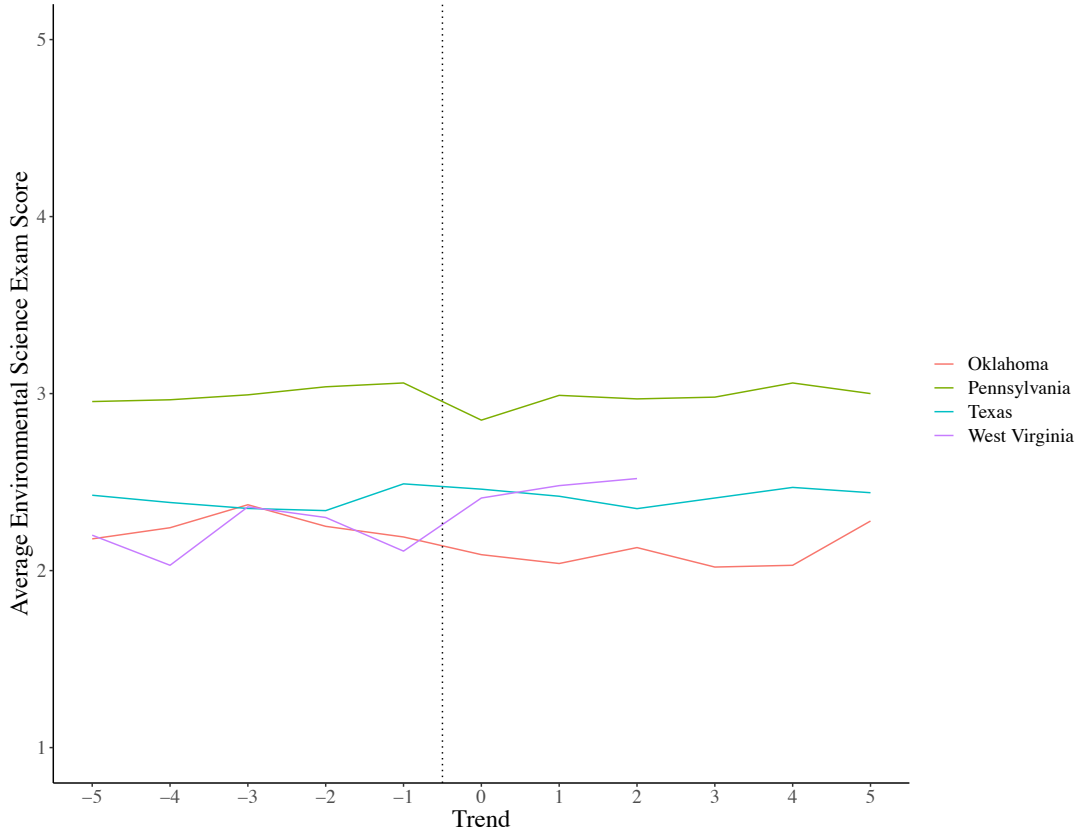


Figure K9

State by State Trajectories for AP Psychology: Group 1

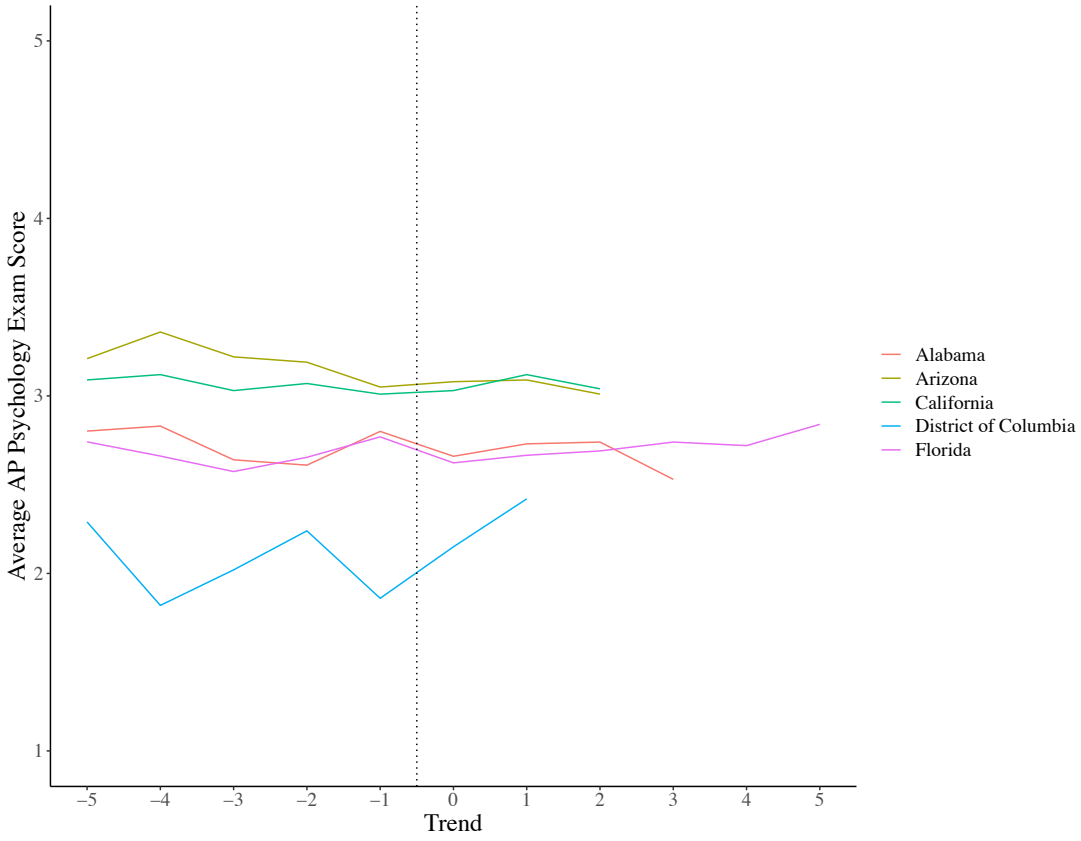


Figure K10

State by State Trajectories for AP Psychology: Group 2

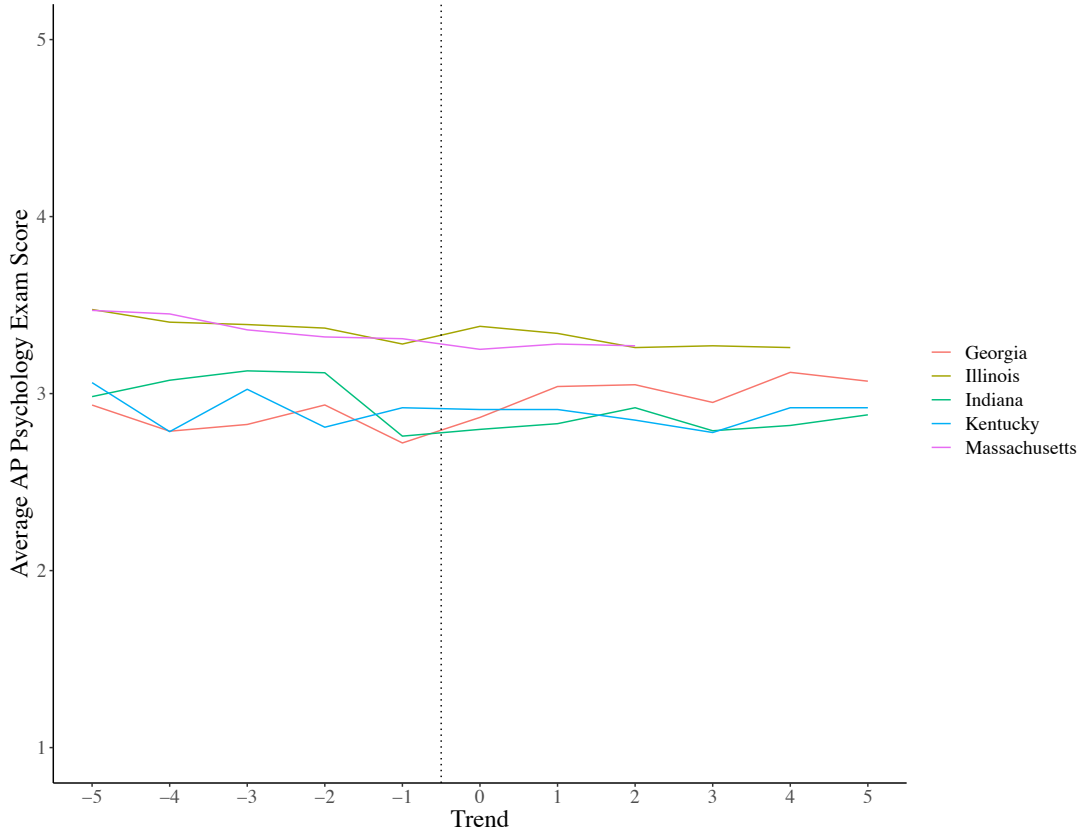


Figure K11

State by State Trajectories for AP Psychology: Group 3

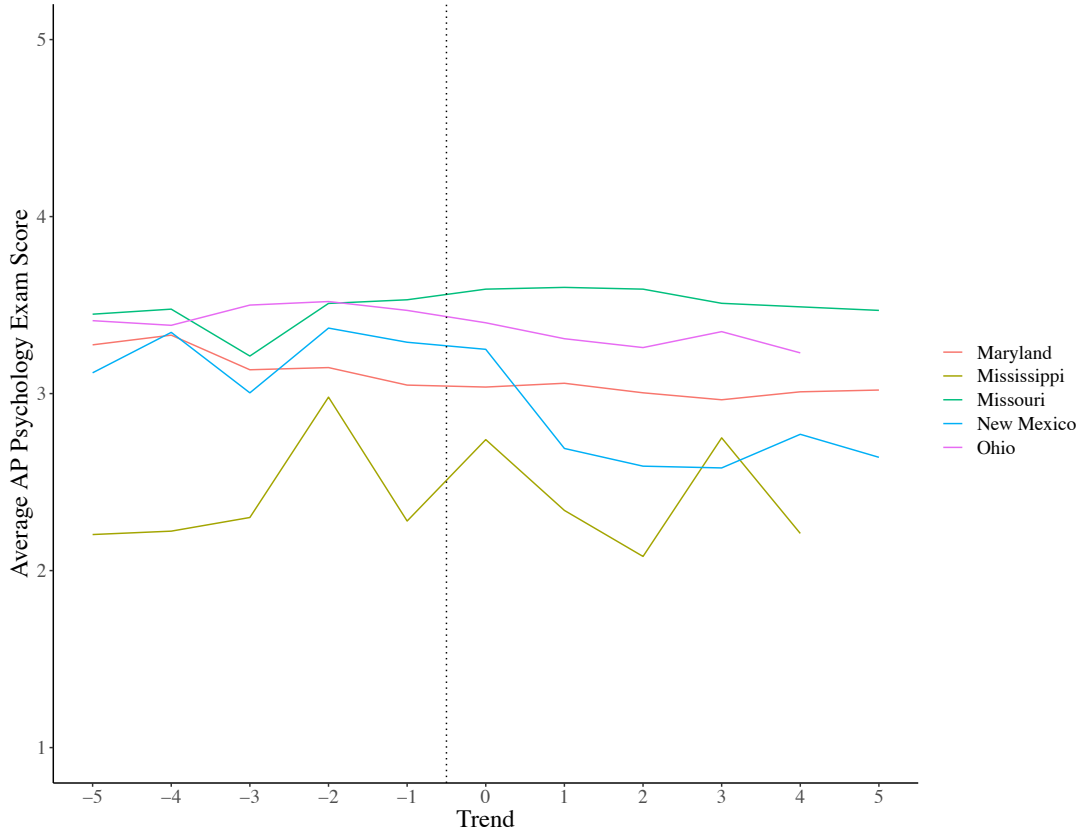


Figure K12

State by State Trajectories for AP Psychology: Group 4

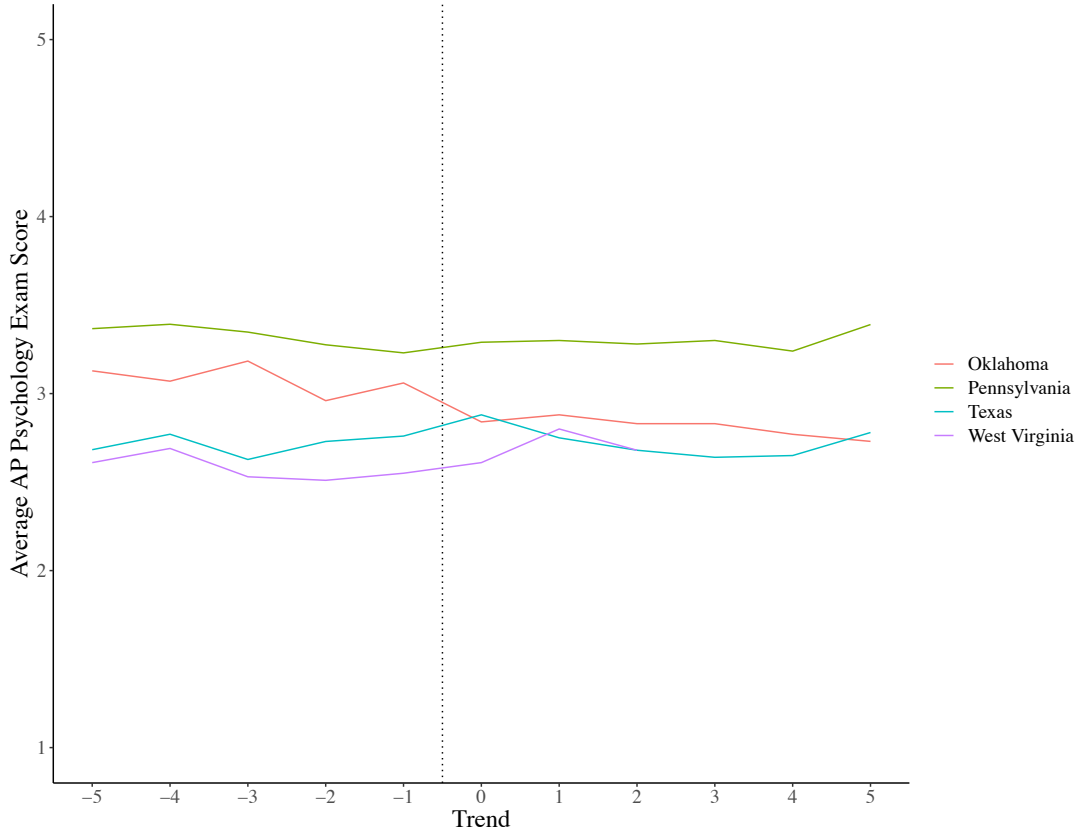


Figure K13

State by State Trajectories for AP Statistics: Group 1

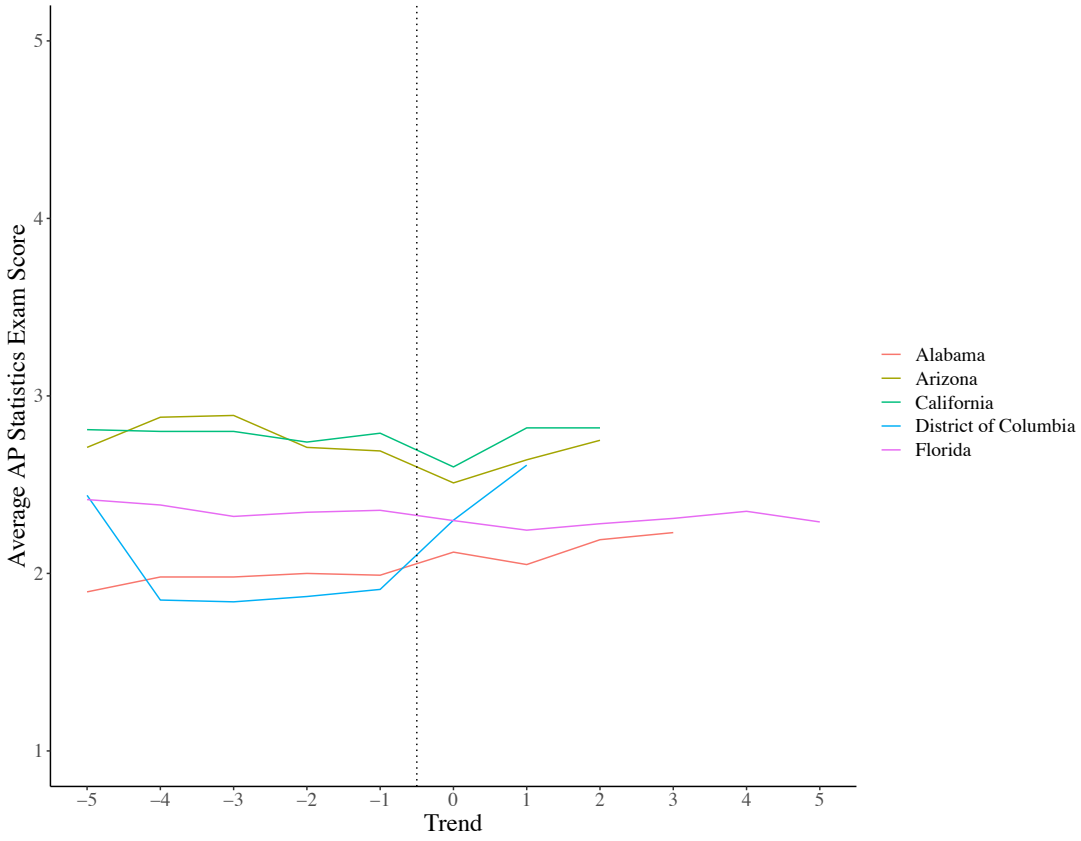


Figure K14

State by State Trajectories for AP Statistics: Group 2

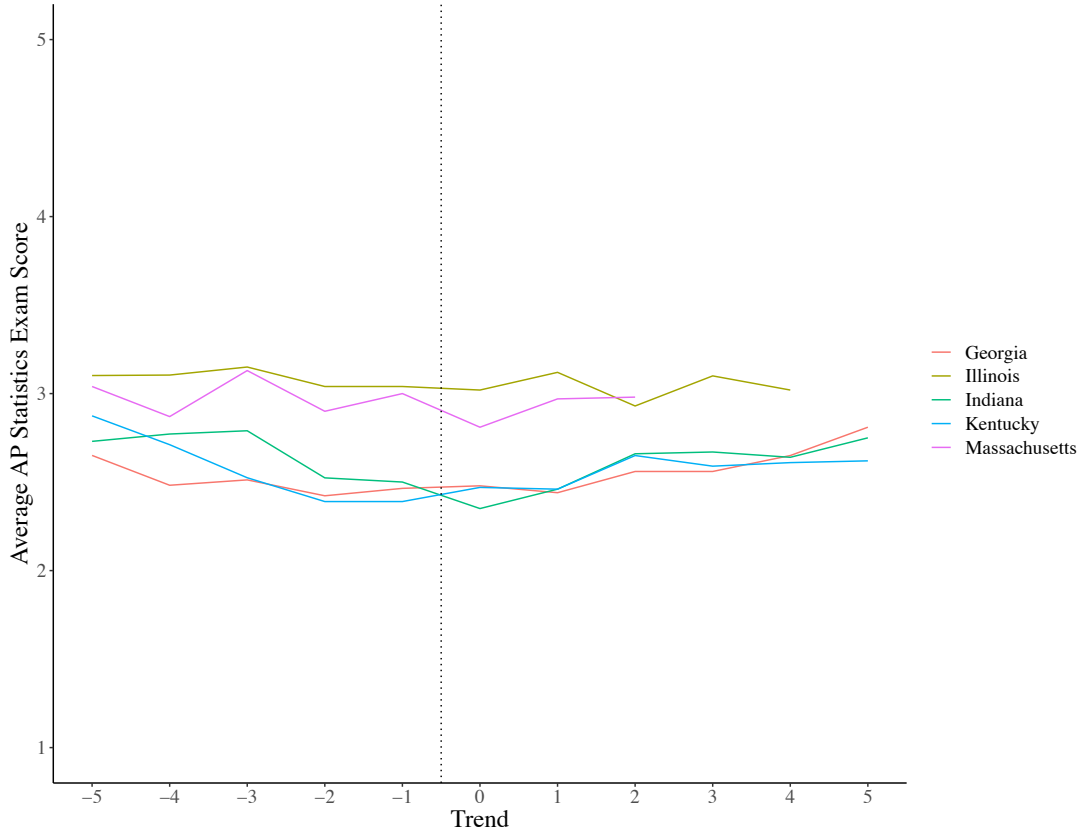


Figure K15

State by State Trajectories for AP Statistics: Group 3

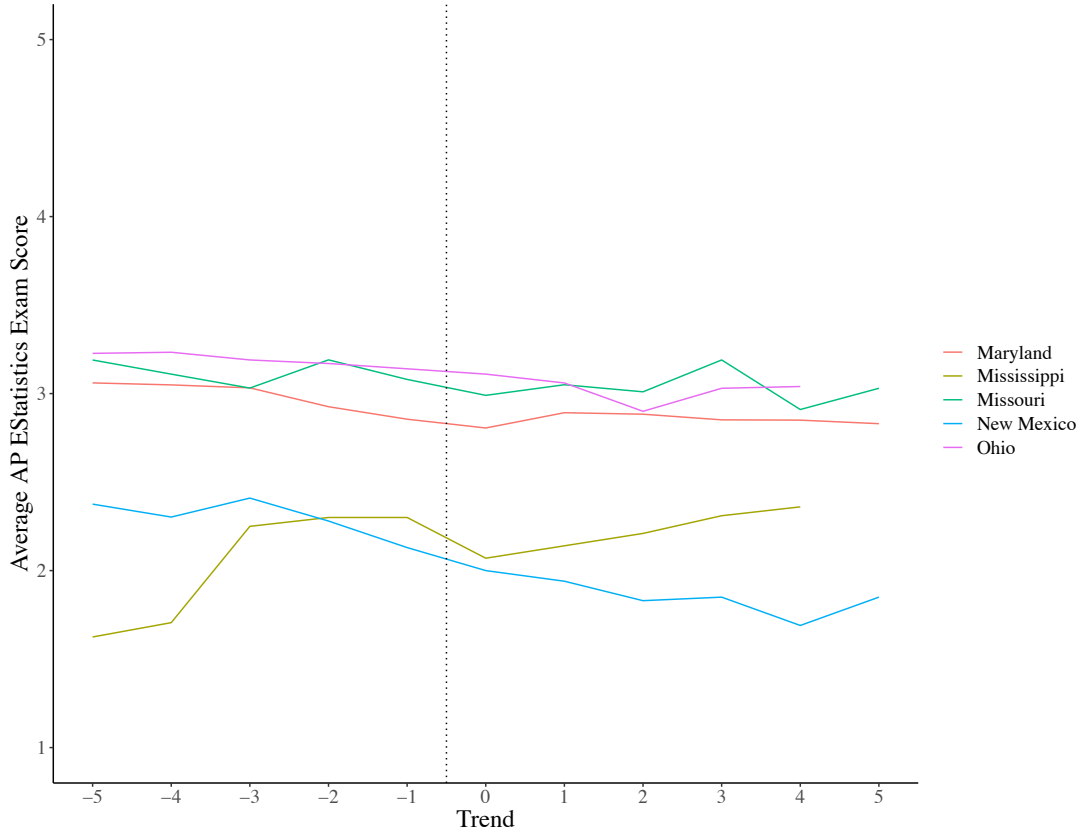


Figure K16

State by State Trajectories for AP Statistics: Group 4

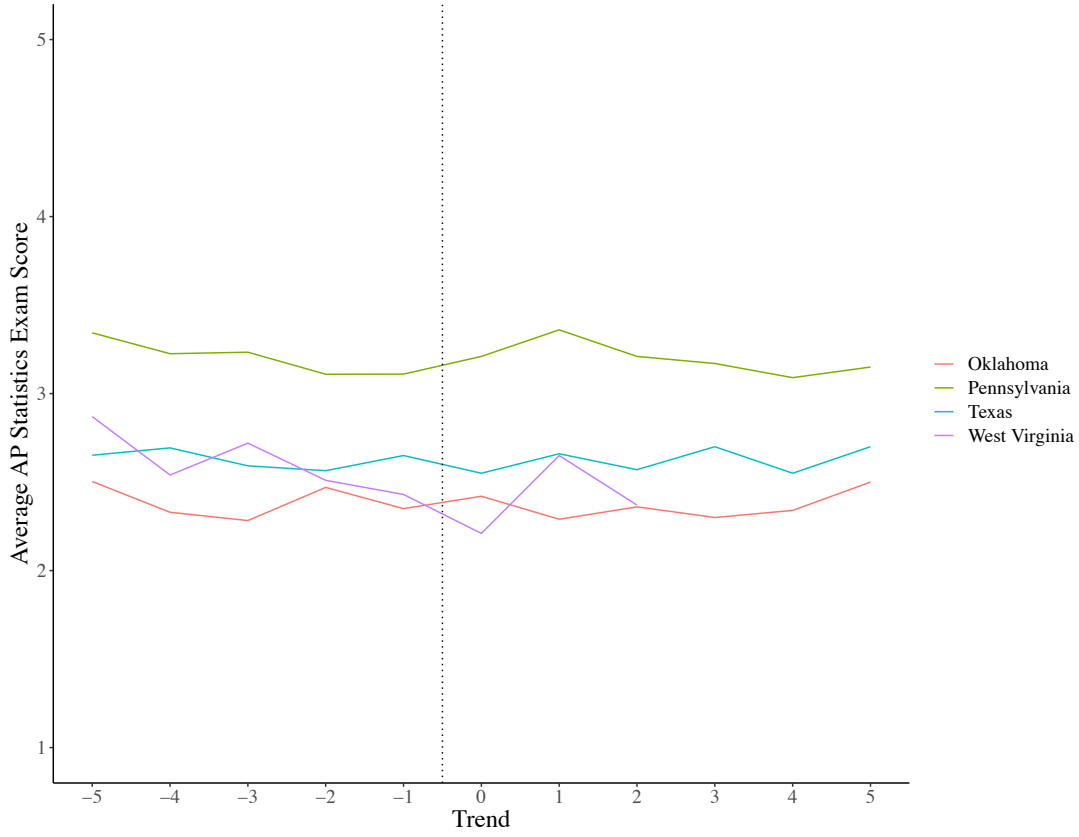


Table K1

Variance in AP Exam Scores across Time, Number of Implementation Years for Policy-Adopting States, and State Rank According to Overall AP Exam Participation

	ELC	ES	Psych	Stat	Imp. Yrs.	Particip. Rank
Policy-Adopting States						
Alabama	0.29	0.62	0.39	0.68	4	27
Arizona	0.20	0.25	0.25	0.34	3	20
California	0.10	0.09	0.07	0.07	3	1
D.C.	0.30	0.43	0.55	0.48	2	48
Florida	0.10	0.11	0.11	0.08	10	3
Georgia	0.08	0.25	0.16	0.13	9	7
Illinois	0.13	0.14	0.22	0.15	5	5
Indiana	0.08	0.27	0.28	0.25	9	18
Kentucky	0.09	0.17	0.11	0.25	6	23
Maryland	0.17	0.17	0.19	0.19	12	9
Massachusetts	0.10	0.19	0.12	0.13	3	14
Mississippi	0.15	0.47	0.59	0.67	5	39
Missouri	0.21	0.54	0.14	0.14	7	29
New Mexico	0.10	0.30	0.36	0.31	6	35
Ohio	0.07	0.23	0.12	0.16	5	12
Oklahoma	0.17	0.19	0.25	0.21	6	28
Pennsylvania	0.12	0.09	0.07	0.14	7	11
Texas	0.12	0.16	0.09	0.12	7	2
West Virginia	0.14	0.22	0.20	0.30	3	37
Comparison States						
Alaska	0.12	0.38	0.22	0.36		46
Arkansas	0.32	0.33	0.36	0.20		25
Colorado	0.12	0.35	0.11	0.12		16
Connecticut	0.10	0.16	0.17	0.14		21
Delaware	0.29	0.44	0.19	0.19		42

Table K1 (continued)

	ELC	ES	Psych	Stat	Imp. Yrs.	Particip. Rank
Idaho	0.15	0.43	0.26	0.29		38
Iowa	0.15	0.46	0.12	0.27		32
Kansas	0.22	0.72	0.28	0.32		34
Louisiana	0.41	0.72	0.29	0.50		33
Maine	0.12	0.24	0.34	0.33		36
Michigan	0.09	0.15	0.12	0.12		13
Minnesota	0.09	0.22	0.13	0.17		19
Montana	0.13	0.63	0.20	0.32		47
Nebraska	0.17	0.69	0.18	0.33		40
Nevada	0.11	0.37	0.20	0.33		30
New Hampshire	0.15	0.33	0.26	0.33		41
New Jersey	0.10	0.14	0.14	0.14		10
New York	0.12	0.13	0.07	0.12		4
North Carolina	0.12	0.15	0.18	0.17		8
North Dakota	0.20	0.60	0.21	0.75		50
Oregon	0.19	0.42	0.11	0.25		31
Rhode Island	0.31	0.43	0.27	0.49		44
South Carolina	0.09	0.18	0.20	0.17		22
South Dakota	0.12	0.53	0.27	0.39		49
Tennessee	0.11	0.23	0.17	0.24		26
Utah	0.08	0.15	0.10	0.15		24
Vermont	0.14	0.34	0.35	0.40		45
Virginia	0.09	0.18	0.17	0.19		6
Washington	0.11	0.35	0.24	0.21		15
Wisconsin	0.09	0.15	0.09	0.13		17
Wyoming	0.27	0.38	0.37	0.41		51

Note. ELC = AP English Language and Composition; ES = AP Environmental Science; Psych = AP Psychology; Stat = AP Statistics; Imp. Yrs. = Years of implementation for policy-adopting states; Particip. Rank = State rank based on average number AP exam participants from 1997 to 2019

REFERENCES CITED

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (No. w24003). National Bureau of Economic Research.

Adams, C. J. (2014, December). Colleges vary on credit for AP, IB, and dual classes. Education Week. https://www.edweek.org/ew/articles/2014/12/10/ colleges-vary-on-credit-for-ap-ib.html

Aldeman, C., Hyslop, A., Marchitello, M., Schiess, J. O., & Pennington, K. (2017a). *An independent review of ESSA state plans: June 2017*. Bellwether Education Partners.

Aldeman, C., Hyslop, A., Marchitello, M., Schiess, J. O., & Pennington, K. (2017b). *An independent review of ESSA state plans: December 2017.* Bellwether Education Partners.

Atchison, D. (2020). The impact of priority school designation under ESEA flexibility in New York State. *Journal of Research on Educational Effectiveness*, *13*, 121–146.

Athey, S., & Imbens, G. W. (2018). *Design-based analysis in difference-in-differences settings with staggered adoption* (No. w24963). National Bureau of Economic Research.

Au, W. (2007). High-Stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*, 258–267.

Bae, S. (2018). Redesigning systems of school accountability: A multiple measures approach to accountability and support. *Education Policy Analysis Archives*, *26*(8), 1–32.

Baker-Bell, A. (2013). "I never really knew the history behind African American language:" Critical language pedagogy in an Advanced Placement English language arts class. *Equity & Excellence in Education*, *46*, 355–370.

Ballou, D., & Springer, M. G. (2017). Has NCLB encouraged educational triage? Accountability and the distribution of achievement gains. *Education Finance and Policy*, *12*, 77–106.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (R package version 1.1-7). [Computer software]. https://doi.org/10.18637/jss.v067.i01

Beach, P., Zvoch, K., & Thier, M. (2019). The influence of school accountability incentives on Advanced Placement access: Evidence from Pennsylvania. *Education Policy Analysis Archives*, *27*(138), 1–29.

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education*, *41*, 287–302.

Bernier, J., Feng, Y., & Asakawa, K. (2011). Strategies for handling normality assumptions in multi-level modeling: A case study estimating trajectories of Health Utilities Index Mark 3 scores. *Health Reports*, *22*, 45–51.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, *119*, 249–275.

Boal, A., & Nakamoto, J. (2020). *School change: How does IB Primary Years Programme implementation impact school climate?* WestEd.

Bonilla, S., & Dee, T. (2020). The effects of school reform under NCLB waivers: Evidence from Focus Group schools in Kentucky. *Education Finance and Policy*, *15*, 75–103.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*, 231–268.

Borusyak, K., & Jaravel, X. (2017). *Revisiting event study designs* (SSRN Working Paper). SSRN Working Papers. https://doi.org/10.2139/ssrn.2826228

Brewer, D. J., Killen, K. M., & Welsh, R. O. (2013). The role of politics and governance in educational accountability systems. *Education Finance and Policy*, *8*, 378–393.

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2018). *Educational goods: Values, evidence, and decision making.* The University of Chicago Press.

Brookhart, S. M. (2009). The many meanings of "multiple measures." *Educational Leadership*, *67*, 6–12.

Campbell, D. T. (1966). Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik*. Holt, Rinehart, & Winston.

Campbell, D. T. (1976). *Assessing the impact of planned social change* (Occasional Paper Series No. 8). Western Michigan University Evaluation Center.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*, 305–331.

Center on Education Policy. (2007). *Has student achievement increased since No Child Left Behind?* Author.

Clotfelter, C., Ladd, H., Vigdor, J., & Wheeler, J. (2007). *High-poverty schools and the distribution of teachers and principals.* North Carolina Law Review, University of North Carolina.

College Board. (2020a). AP Course Audit**.** https://apcentral.collegeboard.org/courses/ap-course-audit

College Board. (2015). AP Potential™: Discover the possibilities. https://apcentral-y2ffarwztgd04mxuwfj25nrfrvo-origin.collegeboard.org/pdf/ap-potential-brochure-schools-and-districts.pdf?course=ap-art-history

College Board. (2020b). Discover the benefits of AP. https://apcentral.collegeboard.org/about-ap/discover-benefits

College Board. (2020c). ESSA and low-income AP students. https://reports.collegeboard.org/archive/ap-program-results/2016/essa-low-income-ap-students

College Board. (2020d). Exam fees. https://apstudents.collegeboard.org/exam-policies-guidelines/exam-fees

College Board. (1997). National report. https://research.collegeboard.org/programs/ap/data/archived/1997

College Board. (1998). National report. https://research.collegeboard.org/programs/ap/data/archived/1998

College Board. (1999). National report. https://research.collegeboard.org/programs/ap/data/archived/1999

College Board. (2000). National report. https://research.collegeboard.org/programs/ap/data/archived/2000

College Board. (2001). National report. https://research.collegeboard.org/programs/ap/data/archived/2001

College Board. (2002). National report. https://research.collegeboard.org/programs/ap/data/archived/2002

College Board. (2007). National report. https://research.collegeboard.org/programs/ap/data/archived/2007

College Board. (2018). National report. https://research.collegeboard.org/programs/ap/data/archived/ap-2018

College Board. (2019). National report. https://research.collegeboard.org/programs/ap/data/participation/ap-2019

College Board. (2014). The 10th annual AP report to the nation. https://research.collegeboard.org/programs/ap/data/nation

College and Career Readiness and Success Center at the American Institutes of Research (2019). State profile comparison. https://ccrscenter.org/ccrs-landscape/state-profile/new-state-profile-comparison?checkAll=on&checkAll=on

College Board and the American Council on Education. (2020). AP credit-granting recommendations. https://aphighered.collegeboard.org/setting-credit-placement-policy/credit-granting-recommendations

Conley, D. (2003). *Who governs our schools? Changing roles and responsibilities*. Teachers College Press.

Conley, D. T., Beach, P., Thier, M., Lench, S. L., & Chadwick, K. L. (2014). *Measures for a college and career indicator: Final report.* Educational Policy Improvement Center.

Cornett, L. M., & Gaines, G. (1997). *Accountability in the 1990s: Holding schools responsible for student achievement* (ED406739). Southern Regional Education Board.

Cornman, S. Q., Zhou, L., Howell, M. R., & Young, J. (2018). *Revenues and expenditures for public elementary and secondary education: School year 2014–15 (fiscal year 2015).* United States Department of Education, Institute of Education Sciences.

Cronin, J., Kingsbury, G. G., McCall, M. S., Bowe, B. (2005). *The impact of the No Child Left Behind Acton student achievement and growth, 2005 edition.* Northwest Evaluation Association, Northwest Evaluation Association Technical Report.

Cuban, L. (2004). Looking through the rearview mirror at school accountability. In K. Sirotnik (Ed.), *Holding accountability accountable* (pp. 18–34). Teachers College Press.

Cullen, J., & Reback., R. (2006). Tinkering towards accolades: School gaming under a performance accountability system. In T. J. Gronberg & D. W. Jansen (Eds.), *Advances in applied microeconomics* (Vol. 14, pp. 1–34). Emerald Group Publishing Limited.

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives*, *22*(86), 1–37.

Dee, T. S., & Dizon-Ross, E. (2019). School performance, accountability, and waiver reforms: Evidence from Louisiana. *Educational Evaluation and Policy Analysis*, *41*, 316–349.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*, 418–446.

Dixit, A. (2002). Incentives and organizations in the public sector: An interpretative review. *Journal of Human Resources*, *37*, 696–727.

Dodd, B. G., Fitzpatrick, S., & De Ayala, R. J. (2002). *An investigation of the validity of AP grades of 3 and a comparison of AP and non-AP graduate groups* (Report No. 2002–9). College Entrance Examination Board.

Dorn, S. (2007). *Accountability Frankenstein: Understanding and taming the monster*. Information Age Publications.

Dougherty, S. M., Goodman, J. S., Hill, D. V., Litke, E. G., & Page, L. C. (2017). Objective course placement and college readiness: Evidence from targeted middle school math acceleration. *Economics of Education Review*, *58*, 141–161.

Dougherty, S. M., & Weiner, J. M. (2019). The Rhode to turnaround: The impact of waivers to No Child Left Behind on school performance. *Educational Policy*, *33*, 555–586.

Duncan, A. (2011). Letters from the Education Secretary or Deputy Secretary. U.S. Department of Education. https://www2.ed.gov/policy/gen/guid/secletter/110923.html

Education Commission of the States. (2006). Advanced Placement policies: All state profiles. http://ecs.force.com/mbdata/MBQuestRT?Rep=AP03

Education Commission of the States. (2016). Advanced Placement policies: All state profiles. http://ecs.force.com/mbdata/mbprofallrt?Rep=APA16

Education Commission of the States. (2007). Exit exams: State requires passage of exit exam for high school graduation. http://ecs.force.com/mbdata/mbquestsnr?Rep=EE01

Education Commission of the States. (2018). Math and English language arts assessments and vendors for grades 3-8 (2017-18): April 2018. http://ecs.force.com/mbdata/mbquestrt?rep=SUM1801

Epstein, N. (2004). *Who's in charge here? The tangled web of school governance and policy.* Brookings Institution Press.

Evans, B. J. (2019). How college students use Advanced Placement credit. *American Educational Research Journal*, *56*, 925–954.

Ferris, J. M. (1992). School-based decision making: A principal-agent perspective. *Educational Evaluation and Policy Analysis*, *14*, 333–346.

Figlio, D. (2006). Testing, crime and punishment. *Journal of Public Economics*, *90*, 837–851.

Figlio, D., & Getzler, L. (2007). Accountability, ability and disability: Gaming the system? In T. J. Gronberg & D. W. Jansen (Eds.), *Advances in applied microeconomics* (Vol. 14, pp. 35–50). Emerald Group Publishing Limited.

Figlio, D. N., & Kenny, L. W. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics*, *93*, 1069–1077.

Figlio, D., & Ladd, H. F. (2015). School accountability and student achievement. In H. F. Ladd & M. Goertz (Eds.), *Handbook of research in education finance and policy* (2nd ed., pp. 194–210). Routledge.

Figlio, D., & Loeb, S. (2011). School accountability. In E. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 383–421). Elsevier.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver & Boyd.

Forte, E. (2010). Examining the assumptions underlying the NCLB federal accountability policy on school improvement. *Educational Psychologist*, *45*, 76–88.

Gafoor, K. A., & Kurukkan, A. (2016). Self-regulated learning: A motivational approach for learning mathematics. *International Journal of Education and Psychological Research*, *5*, 60–65.

Gailmard, S. (2012). Accountability and principal-agent theory. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford handbook on public accountability* (pp. 90–105). Oxford University Press.

Gibbons, R. (1998). Incentives in organizations. *The Journal of Economic Perspectives*, *12*, 115–132.

Goertz, M. E., & Duffy, M. C. (2001). *Assessment and accountability systems in the 50 states: 1999–2000.* CPRE Research Report RR-046. Consortium for Policy Research in Education.

Goldhaber, D., & Özek, U. (2019). How much should we rely on student test achievement as a measure of success. *Educational Researcher*, *48*, 479–483.

Goodman-Bacon, A. (2018). *Difference-in-differences with variation in treatment timing* (No. w25018). National Bureau of Economic Research.

Gordon, R. A., & Heinrich, C. J. (2004). Modeling trajectories in social program outcomes for performance accountability. *American Journal of Evaluation*, *25*(2), 161–189.

Groen, M. (2012). NCLB—The educational accountability paradigm in historical perspective. *American Educational History Journal*, *39*(1), 1–14.

Gunzenhauser, M., & Hyde, A. (2007). What is the value of public school accountability? *Educational Theory*, *57*, 489–507.

Hallberg, K., Williams, R., Swanlund, A., & Eno, J. (2018). Short comparative interrupted time series using aggregate school-level data in education research. *Educational Researcher*, *47*, 295–306.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*, 297–327.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*, 267–271.

Hemelt, S. W., & Jacob, B. (2017). *Differentiated accountability and education production. Evidence from NCLB waivers* (Working paper 23461). National Bureau of Economic Research.

Hess, F. M., & McShane, M. Q. (2018). *Bush-Obama school reform: Lessons learned*. Harvard Education Press.

Holmstrom, B., & Costa, J. R. (1986). Managerial incentives and capital management. *The Quarterly Journal of Economics*, *101*, 835–860.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, *7*, 24–52.

Jackson, C. K. (2010). A little now for a lot later: A look at a Texas Advanced Placement Incentive Program. *Journal of Human Resources*, *45*, 591–639.

Jacob, B. A. (2005). Accountability, incentives and behavior: Evidence from school reform in Chicago. *Journal of Public Economics*, *89*, 761–796.

Jacob, B. A. (2017). The changing federal role in school accountability. *Journal of Policy Analysis and Management*, *36*, 469–477.

Jacob, R., Somers, M. A., Zhu, P., & Bloom, H. (2016). The validity of the comparative interrupted time series design for evaluating the effect of school-level interventions. *Evaluation Review*, *40*, 167–198.

Jacobsen, R., Snyder, J., & Saultz, A. (2014). Informing or shaping public opinion? The influence of school accountability data format on public perceptions of school quality. *American Journal of Education*, *121*(1), 1–27.

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*, 381–389.

Jennings, J. L., & Lauen, D. L. (2016). Accountability, inequality, and achievement: The effects of the No Child Left Behind act on multiple measures of student learning. *Russell Sage Foundation Journal*, *2*, 220–241.

Jensen, M. C., & Meckling, W. C. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*, 305–360.

Jeong, D. W. (2009). Student participation and performance on Advanced Placement exams: Do state-sponsored incentives make a difference? *Educational Evaluation and Policy Analysi*s, *31*, 346–366.

Judson, E., & Hobson, A. (2015). Growth and achievement trends of Advanced Placement (AP) exams in American high schools. *American Secondary Education*, *43*, 59–76.

Kaufman, T. E., Grimm, E. D., & Miller, A. E. (2012). *Collaborative school Improvement: Eight practices for district-school partnerships to transform teaching and learning.* Harvard Education Press.

Kramer, D. (2016). Examining the impact of state level merit-aid policies on Advanced Placement participation. *Journal of Education Finance*, *41*, 322–343.

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. (2004). The Advanced Placement expansion of the 1990s: How did traditionally underserved students fare? E*ducation Policy Analysis Archives*, *12*(68), 1–15.

Klopfenstein, K., & Thomas, M. K. (2009). The link between Advanced Placement experience and early college success. *Southern Economic Journal*, *75*, 873–891. https://scholar.harvard.edu/files/mkraft/files/kraft_et_al._teacher_evaluation_-_updated_feb_2019.pdf

Klugman, J. (2013). The Advanced Placement arms race and the reproduction of educational inequality. *Teachers College Record*, *115*(5), 1–34.

Kolluri, S. (2018). Advanced Placement: The dual challenge of equal access and effectiveness. *Review of Educational Research*, *88*, 671–711.

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. University of Chicago Press.

Koretz, D. M., & Hamilton, L. S. (2003). *Teachers' responses to high-stakes testing and the validity of gains: A pilot study* (pp. 127-147). Center for Research on Evaluation, Standards, and Student Testing, CSE Technical Report 610.

Kraft, M. A., Brunner, E. J., Dougherty, S. M., & Schwegman, D. J. (2019*). Teacher evaluation reforms and the supply and quality of new teachers* (Brown University Working Paper).

Krieg, J. M. (2008). Are students left behind? The distributional effects of the No Child Left Behind Act. *Education Finance and Policy*, *3*, 250–281.

Labaree, D. (1997). Public goods, private goods: The American struggle over educational goals. *American Educational Research Journal*, *34*, 39–81.

Lacy, T. (2010). Examining AP: 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.

Ladd, H. F. (1996). *Holding schools accountable: Performance-based reform in education.* Brookings Institution.

Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly*, *38*, 494–529.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*, 465–491.

Laffont, J. (1990). Analysis of hidden gaming in a three-level hierarchy. *Journal of Law, Economics, & Organization*, *6*, 301–324.

Lauen, D. L., & Gaddis, S. M. (2016). Accountability pressure, academic standards, and educational triage. *Educational Evaluation and Policy Analysis*, *38*, 127–147.

Lee, J. (2006). Input-guarantee versus performance-guarantee approaches to school accountability: Cross-state comparisons of policies, resources, and outcomes. *Peabody Journal of Education*, *81*, 43–64.

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research*, *78*, 608–644.

Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources state NAEP 1990–2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis*, *34*, 209–231.

Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal*, *41*, 797–832.

Liebowitz, D. D., Porter, L., & Bragg, D. (2019). *The effects of higher-stakes teacher evaluation on office disciplinary referrals.* (EdWorkingPaper: 19-159). Annenberg Institute at Brown University.

Lichten, W. (2010). Whither Advanced Placement—now? In P. M. Sadler, G. Sonnert, R. H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program* (pp. 233–244). Harvard Education Press.

Loveless, T. (2009). *Tracking and detracking: High achievers in Massachusetts middle schools.* Thomas B. Fordham Institute.

Lucas, S. (2001). Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American Journal of Sociology*, *106*, 1642–1690.

Macartney, H., McMillan, R., & Petronijevic, U. (2018). *Teacher performance and accountability incentives* (No. w24747). National Bureau of Economic Research.

Malkus, N. (2016). *The AP peak: Public school offerings Advanced Placement, 2000-2012*. American Enterprise Institute.

McEachin, A. (ND). *Incentives, information, and ideals: The use of economic theory to evaluate educational accountability policies*. http://www.invalsi.it/invalsi/ri/improving_education/papers/mceachin/91.doc

McGuinn, P. (2005). The national schoolmarm: No Child Left Behind and the New educational federalism. *Publius*, *35*, 41–68.

McMurrer, J. (2007). *NCLB Year 5: Choices, changes, and challenges: Curriculum and instruction in the NCLB era*. Center on Education Policy.

Milgrom, P. R. (1988). Employment contracts, influence activities, and efficient organization design. *Journal of Political Economy*, 96, 42–60.

Milgrom, P., & Roberts, J. (1990). The economics of modern manufacturing: Technology, strategy, and organization. *The American Economic Review*, *80*, 511–528.

Moe, T. M. (1984). The new economics of organization. *American Journal of Political Science*, *28*, 739–777.

Moe, T. M. (2003). The politics and practice of accountability. In P. E. Peterson & M. R. West (Eds.), *No Child Left Behind? The politics and practice of school accountability.* Brookings Institution Press.

National Center for Education Evaluation and Regional Assistance. (2009). *Technical methods report: Using state tests in education experiments.* Author.

National Center for Education Statistics. (2018). Fast facts: Back to school statistics. https://nces.ed.gov/fastfacts/display.asp?id=372

National Research Council. (2011). *Incentives and test-based accountability in education: Committee on incentives and test-based accountability in public education*. M. Hout, & S.W. Elliott (Eds.). National Academies Press.

Neal, D., & Schanzenbach, D. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*, 263–283.

Oakes, J. (1985). *Keeping track.* Yale University Press.

Oakes, J. (2005). *Keeping track* (2nd ed.). Yale University.

Oakes, J. (2018). 2016 AERA presidential address: Public scholarship: Education research for a diverse democracy. *Educational Researcher*, *47*, 91–104.

O'Day, J. A. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review*, *72*, 293–329.

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

Pederson, P. V. (2007). What is measured is treasured: The impact of the No Child Left Behind Act on nonassessed subjects. *Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *80*, 287–291.

Polikoff, M. S., McEachin, A. J., Wrabel, S. L., & Duque, M. (2014). The waive of the future? School accountability in the waiver era. *Educational Researcher*, *43*, 45–54.

Ravitch, D. (2002). Testing and accountability, historically considered. In W. M. Evers & H. J. Walberg (Eds.), *School accountability*. Hoover Institution Press.

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

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*, 1394–1415.

Rosenbaum, P. R. (2009). *Observational studies* (2nd ed.). Springer-Verlag.

Rosenbaum, P. R. (2011). Some approximate evidence factors in observational studies. *Journal of the American Statistical Association*, *106*, 285–295.

Rothstein, R., Jacobsen, R., & Wilder., T. (2008). *Grading education: Getting accountability right*. Economic Policy Institute and Teachers College Press.

Rowland, M. L., & Shircliffe, B. J. (2016). Confronting the "acid test:" Educators' perspectives on expanding access to Advanced Placement at a diverse Florida high school. *Peabody Journal of Education*, *91*, 404–420.

Ryan, K., & Shepard, L. A. (2008). *The future of test-based educational accountability*. New York: Routledge.

Saw, G. K., & Schneider, B. (2015). Challenges and opportunities for estimating effects with large-scale education data sets. *Contemporary Educational Research Quarterly*, *23*, 93–119.

Schneider, J. (2009). Privilege, equity, and the Advanced Placement program: Tug of war. *Journal of Curriculum Studies*, *41*, 813–831.

Severson, K. (2011, July 5). Systematic cheating is found in Atlanta's school system. *The New York Times*. https://www.nytimes.com/2011/07/06/education/06atlanta

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.

Simon, N. S., & Johnson, S. M. (2015). Teacher turnover in high-poverty schools: What we know and can do. *Teachers College Record*, *117*(3), 1–36.

Singer, J. D., & Willet, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford University Press.

Somers, M. A., Zhu, P., Jacob, R., & Bloom, H. (2013). *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation*. MDRC.

Springer, M. G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*, *27*, 556–563.

The Education Trust. (2020). New school accountability systems in the states: Both opportunities and peril. https://edtrust.org/new-school-accountability-systems-in-the-statesboth-opportunities-and-peril/

Warne, R. (2017). Research on the academic benefits of the Advanced Placement program: Taking stock and looking forward. *SAGE Open*, *7*(1), 1–16.

West, M. (2009). Public choice and the political economy of American education. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 362–371). Routledge.

West, M. R., & Peterson, P. E. (2003). The politics and practice of accountability. In P. E. Peterson & M. R. West (Eds.), *No Child Left Behind? The politics and practice of school accountability*. Brookings Institution Press.

Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding design elements to improve time series designs: No Child Left Behind as an example of causal pattern-matching. *Journal of Research on Educational Effectiveness*, *8*, 245–279.

Wrabel, S. L., Saultz, A., Polikoff, M. S., McEachin, A., & Duque, M. (2018). The politics of Elementary and Secondary Education Act waivers. *Educational Policy*, *32*, 117–140.

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