USC Dornsife
Center for Economic and Social Research

Knowledge in Action Efficacy Study Over Two Years

FEBRUARY 22, 2021

USC DORNSIFE CENTER FOR ECONOMIC AND SOCIAL RESEARCH
Anna Rosefsky Saavedra, *Principal Investigator*
Ying Liu
Shira Korn Haderlein

GIBSON CONSULTING GROUP
Amie Rapaport, *Co-Principal Investigator*
Marshall Garland
Danial Hoepfner

PENN STATE UNIVERSITY
Kari Lock Morgan, *Co-Principal Investigator*
Alyssa Hu

cesr.usc.edu

PREFACE

Knowledge in Action is a project-based learning approach to Advanced Placement (AP). Developers designed KIA intervention—comprised of curriculum, instructional materials, and robust professional development supports for teachers—to bolster students' deeper understanding of content and skills by encouraging their active engagement through projects. With funding from the George Lucas Educational Foundation between 2008-15, University of Washington professors, in collaboration with local teachers, designed the KIA curriculum to cover an entire subject-specific AP curriculum framework through five project units taught over the course of an academic year. Lucas Education Research, the research division of the George Lucas Educational Foundation, asked the University of Southern California's Dornsife Center for Economic and Social Research to conduct a randomized controlled trial efficacy evaluation of the Knowledge in Action intervention in 2016-17 with a follow-up study of RCT schools in 2017-18, funding this work with grants from March 2015 through March 2021. This report describes the study motivation, context, research methods, results, and implications.

ACKNOWLEDGEMENTS

The authors* extend our deep gratitude to the students, teachers, school leaders, and district staff who voluntarily participated in Knowledge in Action data collection. We are grateful to Lucas Education Research staff and leadership, AnneMarie Baines, Britte Cheng, and Nathan Warner, and for their comprehensive support. We thank Jill Carle, Sarah Jean Johnson, Janet Li, and Elizabeth Marwah for their many research contributions, and PBLWorks staff and coaches for their participation in data collection activities. We recognize the College Board, including Trevor Packer, Jeff Olson, Daryl Morris, Sherby Jean-Leger, and many other colleagues, for their support through sharing de-identified student assessment data. We are thankful to Beth Tipton and anonymous reviewers for two rounds of feedback to our methodology, results, and reporting. We thank, for their advice at varying points throughout the course of the study, Howard Everson, Jennifer Hamilton, Jane Lo, Richard Murnane, Susan Nolen, Walter Parker, Juan Saavedra, and Sheila Valencia. Also, we are grateful to Charlotte Holms and Karlen Lusbourgh for their support to district staff, Azucena Carabez and Tricia Goebel for serving as student survey and College and Work Readiness Assessment test proctors, and Mike Branom for writing support. We thank the staff of Council for Aid to Education for facilitating assessment administration and providing assessment data and analytic support. Finally, we credit John Engberg, Laura Hamilton, Tiffany Tsai, and Jennifer Tamargo for their initial contributions to the study. Any flaws are solely the authors' responsibility.

We performed computations for this research on the Pennsylvania State University's Institute for Computational and Data Science's Roar supercomputer.

We gratefully acknowledge financial support for this research from the George Lucas Educational Foundation. Findings and conclusions are those of the authors' and do not necessarily reflect Foundation views.

* *Principal Investigator Anna Rosefsky Saavedra is the corresponding author. She can be reached at asaavedr@usc.edu.*

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Background. Knowledge in Action (KIA) is a project-based learning (PBL) approach to Advanced Placement (AP). Developers designed the KIA intervention—comprised of curriculum, instructional materials, and robust professional development supports for teachers—to bolster students' deeper understanding of content and skills by encouraging their active engagement through projects. With funding from the George Lucas Educational Foundation between 2008-15, University of Washington professors, in collaboration with local teachers, designed the KIA curriculum to cover a subject specific AP curriculum framework over the course of an academic year.

The Knowledge in Action Intervention. KIA curriculum and instructional materials are available with open access through the Sprocket portal, hosted by Lucas Education Research, for AP U.S. Government, AP Environmental Science, and AP Physics; this study addressed the first two. The KIA intervention supports are course-specific, designed to align to the College Board's course curriculum frameworks for AP U.S. Government (APGOV) and AP Environmental Science (APES). However, the same design principles apply to both courses, so both versions of the KIA intervention include similar resources. AP exams administered and scored by the College Board serve as ideal measures of student outcomes. Developers designed both courses with the goals of developing students' civic, political, and/or environmental awareness and engagement. Ongoing job-embedded professional learning provided by PBLWorks included a four-day summer institute, four full days during the year, and on-demand virtual coaching support.

Research Design. After one year, we evaluated the efficacy of the KIA intervention using a randomized controlled trial (RCT) with school-level randomization. Our research design featured a staggered roll-out, such that Year One (2016-17) impact analyses compared student outcomes of teachers with one versus zero years of KIA experience, while Year Two (2017-18) impact analyses compared student outcomes of teachers with two years of KIA versus one. Since the staggered rollout did not allow for a pure Year Two control group, we used two methods to estimate differences in AP outcomes between students of teachers with two and zero years of KIA experience. Our study of teachers' KIA classroom implementation accompanied the impact analyses.

Sample. Teachers and their students were from five large school districts across the country. A higher proportion of the student sample, compared to typical AP exam-takers, was from low-income households. Four of five participating districts serve majority Black and Hispanic students.

Results. Results have notable implications for practitioners and policymakers. Under optimal conditions of teacher support, the Year One pattern of results suggests a positive KIA impact on students' propensity to earn qualifying scores on AP examinations and underlying continuous AP scores. Earning qualifying AP scores can translate into college credit, and relates to enrolling and persisting in college. At the end of the first year, we observed positive results both within each course and pooled across courses. The pattern also was positive within respective groups of students from lower- and higher-income households; in districts serving a majority of students from lower-income households; in districts serving a majority of students from higher-income households; and within each of the five participating districts. Thus, one subgroup of students did not drive overall results.

Strengthening causal claims is the RCT design, as well as the statistical significance, magnitude, and robustness of estimated effect sizes across multiple covariate-adjusted sensitivity analyses. Weakening causal claims are high school-level attrition post-randomization, and differences between unadjusted results compared to those statistically adjusted to address observed baseline differences.

Results contribute to a body of evidence, currently narrow, on whether teachers' proficiency implementing inquiry-based pedagogy, including PBL, changes after their first year. Year Two results were less conclusive due to a lack of significance in second- versus first-year estimates, and substantial limitations to estimates of two years of KIA experience relative to none. Though student AP exam performance may continue to benefit from a teacher's second year of experience with KIA, AP performance gains occurred primarily in teacher's first KIA year. The only outcome with a different second-year trend was students' propensity to take the AP exam, for which we observed nearly all the effect in a teacher's second KIA year.

Investigation of implementation in Year One revealed teachers felt KIA was more engaging for students, offering the opportunity for them to develop real-world skills. Though KIA treatment teachers found the change considerable, with challenges in pacing and groupwork facilitation, the majority recommended the approach, citing benefits for themselves and students. Treatment students voiced benefits related to civic engagement, group work, engagement with learning, and examination preparation. For students of KIA teachers to outperform control students is even more practically meaningful given both teachers and students perceived benefits beyond examination performance.

Implications. Particularly in a high-stakes AP setting, shifting from primarily transmission to PBL instruction is a substantial change for teachers, suggesting the need for high levels of professional development and coaching support. During teachers' first year of KIA, professional learning supports included four days in the summer, four days during the school year, and on-demand coaching. In a teacher's second year, KIA supports were optional and did not include access to on-demand coaching; few second-year teachers participated. That the impact on student AP exam performance occurred primarily in teachers' first KIA year aligns with the ongoing, job-embedded professional learning occurring during that time. The lack of observed erosion of impact on student outcomes in teachers' second year of KIA suggests costs to shift to PBL do not require annual professional development expenses for teachers.

Related to scaling KIA beyond the KIA RCT study are treatment teachers' self-reported perception that KIA aligned to the AP curriculum framework and examinations, and students' feelings of learning more deeply and being prepared for the AP examinations. Also critical to scaling are KIA teachers' positive perceptions of the approach across courses and their recommendations of KIA to others.

In conclusion, the results of the Knowledge in Action Efficacy Study support teacher-driven adoption of the KIA approach in both APGOV and APES courses, among districts with open-enrollment AP policies that support project-based learning, and for students from both lower- and higher-income households.

STUDY OVERVIEW

Knowledge in Action (KIA) is a project-based learning (PBL) approach to Advanced Placement (AP). Developers designed the KIA intervention—comprised of curriculum, instructional materials, and robust professional development supports for teachers—to bolster students' deeper understanding of content and skills by encouraging their active engagement through projects. With funding from the George Lucas Educational Foundation between 2008-15, University of Washington professors, in collaboration with local teachers, designed the KIA curriculum to cover an entire subject-specific AP curriculum framework through five project units taught over the course of an academic year. KIA curriculum and instruction materials are available through the online Sprocket portal hosted by Lucas Education Research. Ongoing job-embedded professional learning, provided by PBLWorks during the study, included a four-day summer institute, four full days during the year, and on-demand virtual coaching support throughout the year.

KIA curriculum and instructional materials are available with open access for AP U.S. Government, AP Environmental Science, and AP Physics; this study addressed the first two. The KIA intervention supports are course-specific, designed to align to the College Board's course curriculum frameworks for AP U.S. Government (APGOV) and AP Environmental Science (APES). However, the same design principles apply to both courses, so both versions of the KIA intervention include similar resources. Serving as ideal measures of student outcomes for both courses are well-defined, well-known, end-of-year tests with strong psychometric properties: AP exams administered and scored by the College Board. In addition, designed both courses with the goals of developing students' civic, political, and/or environmental awareness and engagement, which are areas of needed focus (e.g., Duncan & Ryan, 2021; Levine & Kawashima-Ginsberg, K. 2017; Valant & Newark, 2017).

After one year, we evaluated the efficacy of the KIA intervention using a randomized controlled trial (RCT) with school-level randomization and student-level outcomes. The study design featured a staggered roll-out such that teachers in control schools could participate in KIA in the second year. In the second year, we harnessed the RCT design to compare AP outcomes between students of teachers with one year of KIA versus two. Also, to estimate differences in AP outcomes between students of teachers with two years of KIA experience versus those with none, we used a novel method of leveraging the sample of teachers who volunteered in the RCT. We complemented this two-year estimate with a propensity-score matched analysis comparing students of teachers with two years of KIA to students of teachers in the same districts who did not volunteer for the RCT. Our study of teachers' classroom implementation of the KIA approach accompanied the impact analyses.

Sample students, in five predominantly urban school districts across the country, were composed of a higher proportion of students from low-income households than the typical AP exam-taking community. This report summarizes our research approaches, and impact and implementation results across the Year One (2016-17) and Year Two (2017-18) school years, concluding with discussion of the implications of findings.

STUDY MOTIVATION

The traditional "transmission" model of instruction, in which teachers transmit knowledge to students through lectures and assigned readings, may be suboptimal for supporting students' ability to think and communicate in sophisticated ways, demonstrate creativity and innovation, and transfer their skills, knowledge, and attitudes to new contexts (Gardner, 1999; Perkins, Jay, & Tishman, 1993). In contrast, through project-based learning (PBL), teachers primarily play a facilitator role while students actively engage in teacher- and student-posed learning challenges, working alone and in groups on complex tasks organized around central questions leading to a final product (Hmelo-Silver, 2004; Thomas, 2000).

A large body of observational studies has revealed positive associations between student exposure to PBL instruction and outcomes including: transfer of knowledge and skills (Barron & Darling-Hammond, 2008; Dochy et al., 2003; Gijbels et al., 2005); student engagement with learning (Boaler, 1997; Cognition and Technology Group at Vanderbilt, 1992; Maxwell, Bellisimo & Mergendoller, 2001; Strobel & Van Barneveld, 2009; Wieseman & Cadwell, 2005); standardized test performance (Geier et al., 2008; Schneider, Krajcik, Marx & Soloway, 2002); civic engagement (Saavedra, 2014; Youniss, 2012); metacognition (Kolodner et al., 2003); long-term retention (Strobel & Van Barneveld, 2009); problem-solving (Drake & Long, 2009); disciplinary thinking (Hernandez-Ramos, P., & De La Paz, S. 2009); and collaboration (Wieseman & Cadwell, 2005).

In a meta-analysis of 82 observational studies comparing PBL to other instructional approaches, Walker and Leary (2009) found PBL was most strongly associated with student gains in strategic thinking and designing solutions to complex challenges. A synthesis of eight meta-studies of observational studies from the past four decades showed transmission instructional approaches lagged behind PBL in the areas of students' long-term retention and skill development, as well as satisfaction among both teachers and students. Transmission approaches were, on average, more effective for short-term retention, as measured through standardized exams (Strobel and van Barneveld, 2009). However, causal conclusions are difficult to draw from observational studies because teachers who choose to teach using a PBL approach may differ from those who do not—and those differences, rather than the PBL approach, may drive observed associations between students' exposure to PBL and their outcomes.

More limited is causal evidence of PBL instruction's impact on student outcomes (Condliffe et al., 2017). Using an RCT design, Finkelstein et al. (2010) demonstrated that a PBL economics course increased high school students' performance on measures of economic literacy and problem-solving, with effect sizes of approximately 0.3 standard deviations. In an RCT evaluation of the effects of a middle school PBL-based science program on end-of-unit earth and physical science tests, Harris et al. (2015) estimated impacts of 0.22-0.25 standard deviations. Evaluations also detected effects for social studies and informational reading at the elementary school level using an RCT design (Duke et al., 2020). In a related RCT study, Jackson and Makarin (2018) demonstrated inquiry-based instructional materials improved middle school students' state standardized mathematics scores by 0.06 standard deviations. When those instructional materials were paired with online professional development support, the improvement was 0.09 standard deviations.

No evaluations of PBL interventions harnessing an RCT design have documented program effects for more than one year. Multi-year studies of curricular interventions have demonstrated that teachers may not be able to fully implement new curriculum during the first year (e.g., Fullan and Miles 1992; Fullan 1993), with impacts on student outcomes lower in the first year than in subsequent years (e.g., Balfanz et al. 2006; Freer-Alvarez 2016; Rimm-Kaufman et al. 2007).

Building on earlier studies, we estimated the impact of the Knowledge in Action approach upon students' AP scores—outcomes with concrete relevance to students' post-secondary enrollment and success—after one and two years of KIA. The study also documents teachers' experiences using the KIA approach and students' experiences in those AP classes.

AP courses are intended to serve as rigorous preparation for college. When students earn qualifying AP exam scores—often a 3 or higher, though determined by individual post-secondary institutions—they earn credit accepted by many colleges as work towards graduation, which can help lower tuition costs. Earning qualifying AP exam scores also relates to other critical college outcomes, such as enrolling and persisting (Sadler, 2010; Smith, Hurwitz, & Avery, 2017).

The AP program has grown exponentially over the last two decades due to a concerted effort on the part of the College Board and school districts nationwide to expand AP course enrollment beyond already higher-performing and advantaged students—including relaxing prerequisites and encouraging more students to enroll (Finn & Scanlan, 2019; Sadler, 2010). Over the past 15 years, the number of 12th-graders who took at least one AP exam nearly doubled, from approximately 20 to 40% (College Board, 2020). Equity has been one of the College Board's core growth objectives, which they have realized with "impressive gains in access" (Kolluri, 2018, 1) among low-income (College Board, 2019) and Hispanic students (College Board, 2020). Over the past 15 years, the proportion of AP exam-takers from low-income families nearly tripled, from 11% in 2003 to 30% by 2018 (College Board, 2019).

AP is an especially challenging setting for shifting away from transmission instruction to a student-centered approach (e.g., Dole, Bloom, and Kowalske, 2016). AP teachers may be particularly inclined to follow the transmission model because of the prevailing AP culture emphasizing breadth of coverage over depth of learning, and exam preparation over any other outcome (Parker et al., 2013). Also, students have expressed concerns that anything other than transmission may not be the most effective means of preparation for exam success (Barron & Darling-Hammond, 2008; Parker et al. 2011, 2013).

Thus, the impetus for this study is to contribute to a growing body of evidence on the efficacy of PBL, this time in the AP setting and with our primary outcomes of interest the widely-known AP scores. To date, no studies have harnessed an RCT approach to determine the impact of a PBL intervention on classroom practice in an AP setting (Condliffe et al., 2017), nor examined impacts on outcomes as widely familiar—and policy-relevant—as AP examination scores. In addition, to our knowledge none have attempted to estimate PBL effects after the first year of an intervention's use within the context of an RCT study.

THE KNOWLEDGE IN ACTION INTERVENTION

KIA is a PBL approach to AP designed to support students' deeper learning of content and skills. The University of Washington researchers and local teachers who collaboratively developed the KIA curriculum envisioned a means to realize the potential yet under-realized impact AP courses could have on deeper learning for all students (Parker et al., 2011, 2013).

Six closely related concepts drawn from the learning sciences serve as KIA's theoretical foundation (Parker et al., 2013).

1. Accelerated coverage of material at a rapid pace is not strongly associated with learning depth.
2. Depth of transferable learning is preferable to breadth of temporary, non-transferable learning.
3. Critical to aligning the focus of instruction with deeper learning are assessments requiring students to demonstrate transferable deeper learning, rather than just knowledge and skills.
4. Courses should include transmission and participatory approaches, and their sequencing is critical.
5. Students' receptivity to a forced shift in the balance between transmission and inquiry approaches matters, particularly in the context of a high-stakes examination.
6. Students can develop knowledge through active approaches to learning.

KIA's "engagement first" principle—based on Schwartz and Bransford's (1998) research on instructional sequence—states that initiating learning about a topic through project work will prime students' interest and create a context for learning content through reading or lecture. KIA employs transmission instruction, but in a way intended to maximize its benefits. Other KIA design principles grounded in research on deeper learning include projects as central to the curriculum rather than supplementary; the curriculum looping "quasi-repetitively" over content and skills throughout the course of the year; teachers as co-designers, adapting the curriculum in response to their students' needs and their own; and a scalable approach.

AP U.S. Government and AP Environmental Science Courses

Though the intervention's curriculum and instruction course materials, as well as professional development (PD) supports, are course-specific, designed to respectively align to the College Board's APGOV and APES course curriculum frameworks, the same design principles apply to both courses with similar types of support resources. KIA's original developers and funders chose to focus on these AP courses for two reasons. First, well-defined, well-known, end-of-year examinations with strong psychometric properties—the AP exams administered and scored by the College Board—serve as ideal measures of student outcomes. Impact, as demonstrated through AP scores, is universally easy to interpret as practically meaningful. Second, the courses are designed to develop students' civic, political, and/or environmental awareness and engagement, areas of needed focus (e.g., Valant & Newark, 2017). Aligning with several best practices of civics education (Gould et al., 2011; Levine &

Kawashima-Ginsberg, K. 2017), KIA includes classroom instruction in civics and government; simulations of adult civic roles; student voice; "action civics;" and discussion of current events and controversial issues. Through KIA, students should learn how and why to engage civically rather than simply absorbing facts about citizenship, as is more typical in civics education (Hillygus & Holbein, 2020). Given the KIA APGOV and APES content focus, in combination with their pedagogical approaches, the courses provide opportunities for students to develop civic engagement skills.

Knowledge in Action Curriculum and Instructional Materials

The developers designed KIA to cover an entire subject-specific AP curriculum framework through project units taught over the course of an academic year. Curricula for both KIA APGOV and APES consist of five units designed to address the knowledge, concepts, and skills included in the College Board's AP curriculum frameworks for those respective courses and examinations. Projects are the core structure within each unit and are strategically sequenced to build upon the course's driving question. Example APGOV projects include student debates over historical and contemporary constitutional issues, mock presidential elections, and, for the culminating project, creating a political action plan intended to move an agenda item (e.g., immigration policy) through the political system (Parker & Lo, 2016). Example APES projects include students considering the ecological footprints of themselves and their families, environmental resources for farming, and, for their culminating project, assuming the role of a delegate to an international climate accord convention (Tierney et al., 2020).

Teachers access KIA's curriculum and instruction materials through the Sprocket online portal hosted by Lucas Education Research. Sprocket offers teachers a number of tools: they can look at curriculum pages; download curriculum and/or instructional materials; upload materials to share with others; adapt existing materials to share with others; participate in an online forum discussion; request support; and organize their calendar. Though Sprocket is now open access, during the study, teachers only had access to the portal if they participated in the PD program.

Knowledge in Action Professional Development

The overarching objectives of KIA's PD are to familiarize teachers with the KIA design principles, curriculum, and resources; support teachers' planning for KIA curriculum and instruction; and develop teachers' PBL instructional abilities. During the RCT, PBLWorks provided KIA's PD and included a four-day summer institute, four full in-person days of during the year, virtual coaching, and on-demand support throughout the year.

Critical to the first and second objectives, KIA PD emphasizes to teachers the approach's lack of curriculum "scripting." Rather, successful KIA teaching and learning depends upon adapting the approach to specific classroom contexts. To inform the third objective of developing teachers' PBL instructional practices, PBLWorks integrated their "Gold Standard" definitions of PBL design and practice into all aspects of the professional development program. The Gold Standard Design Elements are a comprehensive, research-based model for best practices in PBL instruction (Larmer, Mergendoller, & Boss, 2015).

The KIA PD model aligns with PD best practices by stipulating that support should be: curriculum- or content-specific (Correnti, 2007; Murray, Savin-Baden, 2000; Supovitz & Turner, 2000); provided on

an ongoing basis for substantial amounts of time (Yoon et al., 2007); aligned with standards (Geier et al., 2008); and featuring active learning and collective participation of teachers from the same communities (Garet et al., 2001). In addition, coaching can help teachers transfer the classroom skills and knowledge gained through PD, and develop norms of collaboration, experimentation, development, and collegiality (Joyce & Showers, 2002). KIA also seeks to build teacher community through its online curriculum portal and in-person PD convenings. Professional learning communities, essential for capacity-building within schools (Stoll et al., 2006), can be valuable supports for instructionally-focused reforms (Andrews & Louis, 2007; Little, 2002). We provide further details about the Knowledge in Action intervention in Appendix A.

BUSINESS-AS-USUAL SUPPORT FOR CONTROL TEACHERS

Participating districts continued to provide teachers randomized to the control condition—i.e., those offered the opportunity to participate in the KIA program the following year—business-as-usual professional development. Supports primarily included College Board-administered PD and school-level AP PD grants. Several districts provided additional support to AP teachers, including those with and without access to the KIA program, above and beyond the College Board's annual 30-plus hour summer institutes. As examples: one district provided PD each August for all its teachers of AP, International Baccalaureate, honors, and dual enrollment courses; In another, teachers could attend monthly trainings on Saturdays through a university partnership program; another offered an AP Institute day each fall; one offered PD not specific to AP for all environmental science teachers; finally, another provided documents highlighting overlap between state standards and AP objectives. With the exception of the PBLWorks PD provided to treatment-group teachers as part of the study, no KIA districts offered PBL PD specific to AP courses. Further details about the business-as-usual support for control teachers can be found in Appendix A.

RESEARCH QUESTIONS

Our research questions address the impact of the offer to use the KIA intervention, or the intent-to-treat (ITT). Insight into the extent to which the offer to use the KIA intervention impacted students' AP performance and other outcomes can inform whether other teachers, schools, and districts should adopt the program, and justify time and resource investments.[1]

As the APGOV course addresses social studies and the APES is a science course, it is important to know whether effects were detected separately within each course group. In light of the growth and diversification of students taking the exam over the past 15 years, we also investigated the KIA effect separately within students from lower and higher income households, using student eligibility for free or reduced-price lunch as a proxy for family income. Our first research question asked:

1 We also estimated the impact of participating in the KIA intervention, or treatment-on-the-treated (TOT). As compliance rates were high, ITT and TOT estimates were essentially equivalent.

RQ1: What is the impact after one year of the offer of the Knowledge in Action intervention—including curriculum, instructional materials, and professional development supports—on student AP examination-taking and performance, and other student outcomes?

A. After one year overall among the full sample

B. After one year within each course subsample

C. After one year within students of lower- or higher-income households

Since the effects of KIA after only one year of use may not have been representative of KIA's potential after teachers have a year to familiarize themselves with the full sequence of KIA curriculum and instructional approaches, we added a second year of data collection in 2017-18. During this time, we followed teachers in schools randomized to the treatment condition into their second year of the KIA offer as well as teachers in control schools into their first year of KIA. Our second research question asked:

RQ2: Did the effect of the KIA offer on student AP examination-taking and performance differ after two years relative to after one?

In the 2017-18 year, all teachers, no matter their original assignment to treatment or control, had received the KIA offer and therefore we did not have a pure experimental comparison group with no KIA experience. Our third research question addressed whether the effect of the KIA intervention offer differed after two years relative to no access to the KIA intervention.

RQ3: What is the effect of the second year of the KIA offer on AP examination-taking and performance relative to no access?

Insight into the extent to which AP teachers who accessed the KIA intervention changed their pedagogy, as well as teachers' and their students' experiences with KIA, substantiates impact results while also informing decisions about program adoption. With qualitative and quantitative research methods, we thus addressed the fourth research question:

RQ4: How did teachers and students experience the KIA intervention?

RANDOMIZED CONTROLLED TRIAL RESEARCH DESIGN

We evaluated the efficacy of the KIA intervention using an RCT, randomizing schools to the conditions of treatment or control. Recruiting districts and teachers into the study depended upon factors such as approval from districts' research review boards and teachers' willingness to enroll. To encourage both, we, in collaboration with our funder Lucas Education Research (LER), intentionally designed the KIA evaluation as a one-year "delayed entry" RCT, in which consented teachers within randomized schools received their first offer to use the KIA intervention in either 2016-17 or the next year (i.e., a staggered rollout design).

Teachers volunteered in spring and summer 2016 to participate in the KIA RCT. One of LER's guiding principles is that instructional shifts must be teacher-driven. For that reason—and due to the

intensity of the study's curriculum and instruction, professional development, and research-related requirements—we recruited teachers directly. Though principals' roles varied by district, for the most part they had minimal involvement in teachers' decisions other than providing consent as per district requirements. Appendix B includes specific criteria for the inclusion of teachers and schools. We did not recruit students to take part in the study, nor did we advertise or announce to parents or school staff any changes in APES or APGOV curriculum.

Randomization of schools within the participating districts took place prior to the 2016 Summer Institute held locally in each school district. We used the re-randomization approach proposed in Morgan and Rubin (2012) to randomly assign 74 schools, with their 86 volunteering/consented teachers, to the treatment or control condition.[2] Participating treatment teachers were in schools randomized to receive the KIA intervention offer starting in 2016-17; control starting in 2017-18.

STUDY SCHOOL DISTRICTS

We partnered with five primarily urban districts, distributed geographically throughout the nation and all serving more than 50,000 students. As we show in Table 1, four of the five are in large cities. Four districts serve majority Black and Hispanic students. In three districts, approximately three-quarters or more of enrolled students live in lower-income households, as measured by eligibility for free or reduced-price meals. Each district sought to develop teachers' capacity for PBL instruction, and required open-access AP course enrollment, such that students did not have to meet any course or prior academic achievement benchmarks if they wanted to enroll.

TABLE 1: Approximate characteristics of KIA RCT participating districts.

	DISTRICT A	DISTRICT B	DISTRICT D	DISTRICT E	DISTRICT E
Locale	City, large	City, large	City, large	Suburb, large	City, large
Enrollment > 50,000	Yes	Yes	Yes	Yes	Yes
Proportion Black and Hispanic students	76%	62%	83%	37%	82%
Proportion students qualifying for free or reduced-price meals*	72%	35%	76%	29%	83%
Participating course(s)*	APGOV & APES	APGOV & APES	APGOV & APES	APGOV	APES
Count (proportion) of randomized school sample	11 (15%)	6 (8%)	12 (16%)	12 (16%)	33 (45%)

*Sources: National Council for Education Statistics Common Core of Data; district administrative data, district websites. Year of statistic provided (2014-2019) varies by source.**

[2] *We originally stated in our summer 2017 RCT preregistration with the American Economic Association that we randomized 76 schools; however, we later determined two schools had been deterministically randomized. We did not include those two schools, or teachers and students within, in any analyses. In spring 2019, we also preregistered with the Society for Research on Educational Effectiveness.*

Three districts participated in the study with both APGOV and APES teachers and students, one with only APGOV, and one with only APES. District D participated only with the APGOV course, and District D students were from higher-income households than the other districts. Districts A, C, and D each composed approximately 15% of the school sample, while District B schools represented 8%, and District E 45%. We further describe the districts, including their motivation to participate in the study and district AP policies, in Appendix C.

STUDENT OUTCOME MEASURES

Student outcome measures included AP performance outcomes from the May 2017 and May 2018 APGOV or APES exam, depending on the course in which they were enrolled and in which year. We examined KIA impacts on students' AP performance in four ways:

- Exam-taking outcome: Whether students took the APES or APGOV exam.
- Qualifying score outcome (full sample): This outcome encompasses all students of sample APES or APGOV teachers, no matter whether they took the exam. Students who took the AP exam and scored a 3, 4, or 5 "earned a qualifying score;" students who either did not take the exam or took it but failed to obtain at least a score of 3 did not earn a qualifying score.
- Qualifying score outcome (exam-takers only): This outcome includes only sample teachers' students who took the APGOV or APES exam. Students who took the AP exam and scored a 3, 4, or 5 "earned a qualifying score;" students who took the exam but scored a 1 or 2 did not earn a qualifying score. Students who did not take the exam are excluded from this analysis.
- Continuous scale score outcomes: These are the scale scores underlying students' ordinal scores, including a total score, multiple choice sub score, and free-response sub score. Like the qualifying score outcome among the exam-taker sample, this outcome is conditional on taking the examination.

The fourth measure was available in four districts, excepting District D, that provided the necessary permissions for the College Board to share de-identified student records with the research team. District D composed approximately 70% of the APGOV exam-taking sample.

In the first year, we administered an end-of-year student survey measuring students' inter- and intra-personal skills, as well as their inclination for civic engagement. We also administered the College and Work Readiness Assessment (CWRA+), which measures students' skills in critical thinking and written communication.[3] In Appendix D, we provide further details on all study data sources; in Appendix E, we describe transformation of student achievement variables.

[3] *This summary document focuses exclusively on AP outcomes because of high student-level attrition and lack of baseline equivalence on other student outcomes. We provide other student outcome results in the accompanying appendices.*

RESEARCH QUESTION 1: SAMPLE, ANALYTIC METHODS, AND RESULTS

RQ1: What is the impact after one year of the offer of the Knowledge in Action intervention—including curriculum, instructional materials, and professional development supports—on student AP examination-taking and performance, and other student outcomes?

We first addressed the study's primary research question about the impact of the offer of the Knowledge in Action intervention on student outcomes after one year by comparing Year One student outcomes of teachers in treatment schools with student outcomes of teachers in control schools. We explored the answer to this question overall, and also within each course, and within groups of students from lower- and higher-income households.

Research Question 1 Sample

We included complier and non-complier teachers, and their students, in our ITT causal analyses of the impact of KIA on student outcomes.

Without consideration for missing student outcomes, six schools, all treatment, attrited from the randomized sample before the end of September 2016. Following What Works Clearinghouse (WWC) version 4.1, this school-level attrition from randomization to the Year One sample exceeded acceptable thresholds at 8% overall, with a 16-percentage point differential between treatment and control conditions.

School attrition from the experimental sample is an important consideration relevant to districts' school-level implementation of the KIA intervention. A key driver of attrition was the lack of schools offering the APGOV or APES course every year, sometimes determining whether to offer the course based on student enrollment as of a September cut-off date, other times offering APES and AP Biology every other year, or APGOV and AP U.S. History every other year. The latter applied to teachers who taught in schools offering APGOV or APES in 2015-16 and consented to enroll in the study with the hopes that their school would again offer the course in 2016-17, and to second-year attrition described later in this report. Another driver of school-level attrition was principals not assigning consented teachers to teach APGOV or APES across consecutive years, even if the school offered the course across consecutive years. Other drivers included teachers switching schools within districts or leaving the district altogether. All school-level attrition from the randomized sample was due to the consenting teacher either not teaching APGOV or APES in 2016-17 or no longer teaching in their randomized school even if they continued to teach APGOV or APES.

School-level attrition also affected interpretation of impact results by potentially changing the composition of the schools, teachers, and students within the randomized treatment group.

After losing six schools post-randomization by the end of September of Year One (i.e., 2016-17), 74 teachers across 68 schools—and their 3,645 students—participated in the study by teaching APGOV or APES during the school year in their randomized school (intent-to-treat sample). Among the 74 teachers' students, 43% were from lower-income households and 47% of the first-year KIA student sample was Black or Hispanic.

With one school each attriting from Districts A, C, and D, and three from District E, the post-randomization school-level sample loss did not change the district composition of the school sample in a meaningful way (Appendix F). We provide details about Year One school and teacher characteristics in Appendix G, showing that changes in observed sample composition over time were quite minimal from randomization to Year One. Located in Appendices H and I are respective tables describing teachers and their students, overall and by treatment status and course, at randomization and Year One.

Research Question 1 analytic methods

We followed standard protocol for RCT analysis, assessing baseline equivalence, estimating ITT effects, and addressing sensitivity to modeling choices. Our primary analytic method for properly accounting for nested data was Hierarchical Linear Modeling (HLM). With school-level randomization and student-level outcomes, we fit two-level HLM models, grouping students within schools, with district-fixed effects to account for blocked randomization by school within districts.

Regression-adjusted models[4] included student-, teacher-, and school-level covariates to account for remaining imbalance between treatment and control groups after randomization, and to improve statistical precision. Drawing from education literature, we chose for consideration in our impact models 22 substantively important covariates across students (e.g., prior achievement, race/ethnicity), teachers (e.g., course, baseline students' AP performance, years teaching APGOV/APES), and schools (e.g., proportion of school eligible for FRLP), as well as all student-level covariates averaged at the school level (Altonji and Mansfield, 2018). Our primary approach to covariate selection was to include in impact models all covariates with absolute baseline standardized mean differences of greater than 0.05 as well as those determined through automated covariate selection to improve model fit.

We used multiple imputation for missing covariate data, multiply imputing 20 data sets and combining results following Rubin's rules (Rubin, 1987).

Though student-level attrition did not exceed WWC thresholds for any AP performance outcomes, what did was the combination of overall and differential school-level attrition. Thus, we conducted baseline equivalence analysis to investigate the extent to which the groups differed after attrition. To meet WWC baseline equivalence standards, baseline differences between treatment and control groups on respective relevant student-level covariates must be less than an absolute value of 0.25 standard deviations. In addition, we included any relevant student-level covariates with absolute effect sizes greater than 0.05 in the impact model. Though baseline equivalence analysis of teacher- and school-level covariates is not required per WWC, the most substantial forms of attrition were at these levels, so we included those covariates as part of our baseline equivalence analysis. Therefore, baseline equivalence analysis was a necessary step to selecting which covariates to include in impact models, and informing interpretation of impact estimates.

We conducted several Year One robustness checks. To check model type, we fit our two-level HLM model as a three-level model, grouping students within schools within districts, and as an ordinary least squares model (i.e., not accounting for nesting with random effects). In addition to fitting

[4] We fit linear regression models for continuous outcomes and logistic regression models for dichotomous outcomes.

unadjusted models without covariates, we also fit our primary two-level HLM model with several combinations of covariates. In addition, as a check to the homogeneity of overall results by district, we estimated the magnitude of treatment effects separately by district.[5]

We also addressed the possibility that non-random sorting of students into KIA treatment or control classrooms could have biased results.

We conducted pre-registered exploratory subgroup analyses within courses and student household income groups. In addition to defining household income subgroups at the student level, as a robustness check we separately estimated effects within the two districts with lower proportions of students from lower-income households (Districts B and D) and within the three districts with higher proportions of students from lower-income households (Districts A, C, and E). We did not design the RCT sample size with enough power to detect significant effects within subgroups should they truly exist. Appendix J includes further details about these analytic steps.

Research Question 1 results

YEAR ONE OVERALL BASELINE EQUIVALENCE

For the qualifying score (exam-takers only) outcome (n=2,963), student-level attrition was 19%, with a differential of 4 percentage points, meeting the WWC "cautious boundary" threshold. Within the subgroup of exam-takers in schools with at least one continuous score outcome (n=1,599), overall attrition was 29% with a 6-percentage point differential, meeting the WWC optimistic threshold. We detail student-level analytic sample attrition across years and outcomes in Appendix K.

To inform our overall impact analyses, we analyzed baseline equivalence for three AP outcome analytic samples: the full sample of 3,645 students, all with an exam-taking outcome and a qualifying score outcome; the sample of students with a qualifying score outcome (exam-taking sample); and the sample of students with continuous score outcomes. For the first two samples, for all covariates baseline equivalence standardized mean differences fell within the 0.25 threshold defined by the WWC standards. For the third sample, including students who took the AP exam in four of five districts with which we examined the KIA impact on continuous AP outcomes (i.e., total score, multiple choice sub-score, and free-response sub-score), baseline imbalance does not exceed the WWC threshold for any student-level covariates. The covariates for which the WWC requires baseline balance, student-level prior achievement and socio-economic status, meet that threshold. Baseline imbalance for this sample exceeded the WWC threshold on two school-level covariates, both with values higher for treatment teachers: proportion of a teachers' students who took at least one AP examination in May 2016, and proportion of a teachers' students who were Asian-American.

No p-values associated with standardized mean differences at baseline fell below the 0.05 significance threshold. Despite school-level attrition exceeding WWC thresholds, these p-values indicated balance between treatment and control groups as expected in an RCT. To the extent we could calculate given missing data, post-attrition baseline equivalence on key covariates describing the teachers'

[5] *Though we are unable to report district-specific effects due to agreements with each participating district's research review board, we conducted these analyses to determine whether results were concentrated in one or more district or district type (e.g., proportion of district students from lower-income households).*

baseline students' AP performance did not get worse from the randomization sample to the Year One sample (Appendix G).

We included, per WWC requirements, all covariates with baseline equivalence standardized mean differences of greater than 0.05 in absolute value in the corresponding impact models. Because treatment groups were not perfectly balanced, either due to random chance or attrition, covariate inclusion in our impact models "adjusted" these baseline imbalances. For example, the prior students of teachers assigned to the treatment group had lower average AP outcomes than the prior students of teachers assigned to the control group—so without covariate adjustment, we would expect a naïve comparison of 2016-17 treatment versus control AP outcomes to be biased against the treatment group. We adjusted our impact estimates accordingly by including in the impact model imbalanced baseline covariates, such as baseline students' 2015-16 AP averages. However, a consequence is our estimated effect sizes depend on modeling the relationship between covariates and outcomes. (For details, see baseline equivalence analysis tables and figures informing Research Question One overall analyses [i.e., non-subgroup] in Appendix L.)

YEAR ONE SUBGROUP BASELINE EQUIVALENCE

We also looked at baseline standardized mean differences for AP performance outcomes within course and student household-income subgroups. With the sample sizes smaller for subgroups, we expected less balance here than for the overall sample. Focusing on the qualifying score outcome (full sample), with a few exceptions described below, we found balance within WWC thresholds on most student-level covariates for APGOV and APES subgroups, and for lower- and higher-household income students. Among the imbalanced covariates, APGOV treatment students were more likely to have taken at least one AP exam the prior year (i.e., in May 2016, SMD=0.658) and scored higher on their eighth-grade English Language Arts (ELA) tests—though not on their tenth-grade national ELA (i.e., PSAT, SAT, or ACT) exam. Within APES, the treatment group was composed of a higher proportion of White (SMD=0.255) and Asian (SMD=0.304) students. Within the lower-income group, White students composed a higher proportion of the treatment group (SMD=0.279) and treatment students scored higher on their eighth-grade math and ELA tests, though not on their 10th-grade national math or ELA exams. All student-level covariates were balanced within the higher-income group.

Within subgroups, most though not all teacher- and school-level covariate standardized mean differences were within the 0.25 threshold. Arguably the most important was the teacher-level covariate describing the proportion of teachers' baseline APGOV or APES students who earned qualifying scores on the exam out of all of their baseline APGOV or APES students. There were sizable standardized mean differences on this covariate in APGOV (SMD=-0.434) yet minimal in APES (SMD=-0.031), and sizable in the higher-household income student group (SMD=-0.391) yet minimal in the lower-household income group (SMD=0.028).

Due to smaller sample sizes, we could not have expected complete baseline equivalence within subgroups. Subgroup baseline equivalence results suggest the within-group impact results are more robust for lower-household income and APES students, and less robust for higher-income and APGOV students. (See all baseline equivalence analysis tables and figures for Research Question One subgroups in Appendix M).

YEAR ONE IMPACT ESTIMATES

OVERALL RESULTS

In covariate-adjusted models (Table 2), students of KIA teachers significantly outperformed students of control teachers on their May 2017 APES and APGOV exams, as measured by several indicators of AP performance. This held true among all 3,645 students of the sample teachers, as well as among the 2,963 students who took the exam.

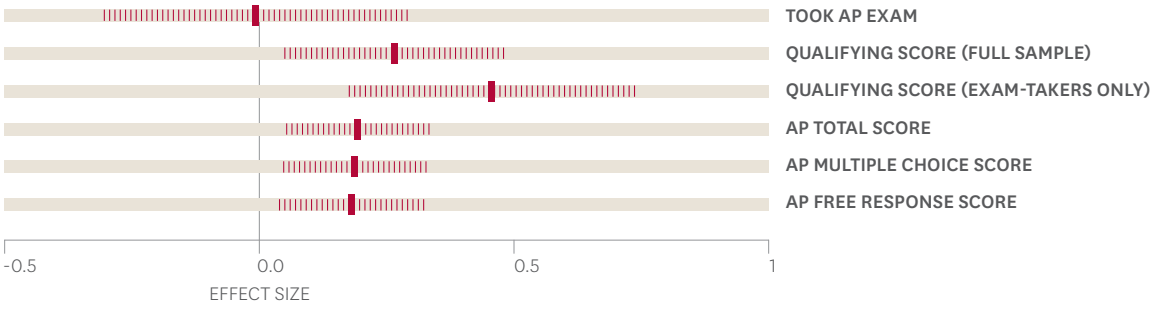
TABLE 2: Covariate-adjusted estimates of the overall Year One AP KIA impact

	EFFECT SIZE (SE)	95% CI	p-value	n
Took AP exam	-0.009 (0.15)	(-0.305, 0.286)	0.950	3645
Qualifying score (full sample)	0.264 (0.11)*	(0.049, 0.478)	0.016	3645
Qualifying score (exam-takers only)	0.457 (0.14)**	(0.177, 0.737)	0.002	2963
AP Total Score	0.192 (0.07)**	(0.054, 0.331)	0.006	1599
AP Multiple Choice Score	0.188 (0.07)**	(0.048, 0.328)	0.009	1599
AP Free Response Score	0.181 (0.07)*	(0.039, 0.323)	0.012	1599

*Notes: Table columns show standardized effect sizes, standard errors, confidence intervals, p-values, and sample sizes. Asterisks denote statistical significance: *p<0.05, ** p<0.01, ***p<0.001.*

We present the results from Table 2 visually in Figure 1.

FIGURE 1: Covariate-adjusted estimates of the overall Year One AP KIA impact



Notes: Figure shows standardized effect sizes and 95% confidence intervals.

We found no overall exam-taking effect (ES=-0.009, p=0.95), indicating no discernable overall relationship between KIA and whether students took the exam.

At the end of Year One, for all students in the Year One sample (n=3,645) we estimated the percentage of students who would earn a qualifying score with KIA is 4 percentage points higher than the percentage that would earn these scores without KIA (ES=0.264, p=0.016); dependent on the particular students in our sample, we estimated 31.1% of the students in our sample who took the AP exam would have earned a qualifying score without KIA, as compared to 35.1% with the KIA intervention.

Among students who took the exam (n=2,963), the equivalent difference was 7.6 percentage points (ES=0.457, p=0.002); again, dependent on the students, we predicted that about 37.2% of those taking the AP exam would have passed without KIA, as compared to 44.8% with the KIA intervention.

In the four districts with continuous AP score outcomes, the estimated effect sizes are significant for total scores (ES=0.192, p=0.009), as well as the multiple-choice (ES=0.188, p=0.009) and free-response (ES=0.181, p=0.012) subsection scores.

In the first column of Table 3, we show covariate adjustment was necessary to reported treatment effects with unadjusted models estimating effect sizes close to zero. Once we conditioned upon including in the impact model all covariates with absolute baseline equivalence effect sizes greater than 0.05, as required per WWC, the direction of estimated effect sizes was robust to method of selecting which covariates to include. Robustness checks also showed results were consistent across model type (i.e., 2-level HLM with district fixed effects, 3-level HLM, ordinary least squares), see Appendix N.

TABLE 3. Robustness of overall Year One KIA impact estimates to covariates.

	NO COVARIATES	PRIMARY MODEL	ALL COVARIATES WITH BE > 0.05	ALL COVARIATES
Took AP exam	0.125 (0.18)	-0.009 (0.15)	-0.007 (0.15)	0.056 (0.16)
Qualifying score (full sample)	-0.005 (0.21)	0.264 (0.11)*	0.16 (0.13)	0.351 (0.13)**
Qualifying score (exam-takers only)	-0.043 (0.21)	0.457 (0.14)**	0.145 (0.13)	0.353 (0.14)*
AP Total Score	0.035 (0.17)	0.192 (0.07)**	0.196 (0.07)**	0.188 (0.08)*
AP Multiple Choice Score	0.049 (0.17)	0.188 (0.07)**	0.196 (0.07)**	0.17 (0.08)*
AP Free Response Score	0.016 (0.16)	0.181 (0.07)*	0.185 (0.07)**	0.191 (0.07)*

*Notes: Table columns show standardized effect sizes and standard errors. Asterisks denote statistical significance: *p<0.05, ** p<0.01, ***p<0.001.*

In addition to the robustness tests, we also investigated the possibility of systemic sorting of students into KIA classrooms in a way that could have threatened the internal validity of our estimates. As we show in Appendices O and P, we did not find evidence of this threat.

Additionally, we estimated treatment effects separately within each district, finding estimated effect sizes on the qualifying score (full sample) outcome were substantively consistent, positive, and of meaningful magnitude.

SUBGROUP RESULTS

By course, 32 APGOV teachers across 29 schools taught 1,693 students (46%) and 42 APES teachers across 42 schools taught 1,952 students (54%); teachers of both courses taught in three schools.

Whereas across courses, the estimated exam-taking effect was zero, within courses we detected exam-taking effects. Though the exam-taking effects were not significant for either course, we did not power the study to detect significant effects within subgroups. Within the APGOV subsample, the effect was negative such that in covariate-adjusted models, KIA students took the exam at lower rates than control (ES=-0.4, p=0.09). In contrast, KIA students within the APES subsample took the exam at higher rates than control (ES=0.225, p=0.22).

Due to differences within courses concerning the characteristics of students who did and did not take the exam, with exam-takers demonstrating higher prior academic performance (Appendix Q), the KIA exam-taking effect affects interpretation of estimated impacts on the qualifying score outcome (exam-takers only) and continuous AP outcomes. The KIA exam-taking effect does not affect interpretation of the qualifying score outcome among the full subgroup samples, as we estimated this outcome for all students regardless of whether they took the AP examination. As reported above, across courses on the qualifying score (full sample) outcome, KIA students outperformed control students (ES=0.264, p=0.016) in covariate-adjusted models. When we disaggregated this estimate by course, the APGOV treatment subgroup outperformed control students, although the difference was not significant (ES=0.227, p=0.13). Within the APES subgroup, KIA students also outperformed control (ES=0.304), and this latter effect size was significant at the 0.05 level (p=0.048).

Among the full sample of 2016-17 students, 43% were from low-income households, as were 38% of exam-taking students. KIA students significantly outperformed control when we estimated effects only within the group of lower-household income students, and when we estimated effects only within the group of higher-household income students.

KIA effect sizes generally were comparable between students from lower-income households and those from higher-income households. We did not detect exam-taking effects within either student group: lower-income (ES=-0.016, p=0.92); higher-income (ES=-0.002, p=0.99). Among lower-income students, KIA students significantly outperformed control on the qualifying score outcome for exam-takers (ES=0.386, p=0.028), total score (ES=0.206, p=0.006), multiple-choice score (ES=0.191, p=0.013), and free-response score (ES=0.208, p=0.009). Among higher-income students, KIA students significantly outperformed control students on the AP qualifying score outcome in both the full sample and among those who took the exam (respectively, ES=0.319, p=0.007; ES=0.496, p=0.002), as well as the outcomes for total score (ES=0.173, p=0.03) and multiple-choice score (ES=0.183, p=0.024).

Robustness checks also showed positive KIA effects within both the lower-household income Districts A, C, and E and the higher-household income Districts B and D, and with effect sizes of comparable magnitude within district groups and, overall, across districts. (Figures and tables describing all Research Question One subgroup impact results can be found in Appendix R.)

OTHER YEAR ONE STUDENT OUTCOMES

In addition to examining the impact of KIA on AP outcomes, after the AP examination period we also administered the online College and Work Readiness Assessment (CWRA+), designed to measure students' skills in critical thinking, program solving, and written communication; and an end-of-year paper survey to students measuring their inter- and intra-personal skills, as well as their inclination for civic engagement. Due to challenges administering the 90-minute online CWRA+ during the final two weeks of school, when computers also were being used for state standardized tests, the sample of students with valid CWRA+ outcomes suffered from student-level attrition exceeding WWC thresholds in addition to the aforementioned school-level attrition. The CWRA+ and survey outcome samples lacked equivalence on measures of students' prior achievement as well as other covariates. Given these limitations, we do not highlight the results from either measure in this report, though we do share details on the accompanying technical apendices. None of the estimates for the survey or CWRA+ outcomes were significant. (Data collection, samples, and results for these outcomes are summarized in Appendices S [CWRA+] and T [survey measures].)

Research Question 1 limitations

School-level attrition exceeded WWC thresholds for the sample used to address Research Question 1, raising concerns about the extent to which attrition changed the composition, both observed and unobserved, of treatment schools relative to control schools. By performing baseline equivalence analyses, we assessed the extent to which such differential attrition could have impacted results. Baseline equivalence analysis demonstrated acceptable balance for observed covariates—partly mitigating this concern, although not completely. The differential cluster-level attrition may have changed the composition of treatment schools relative to control schools—and teachers—in ways we cannot measure or statistically adjust, and that also relate to student performance. In addition, the presented results depend on including covariates in the corresponding impact models, and modeling the relationship between covariates and outcomes.

RESEARCH QUESTION 2: SAMPLE, ANALYTIC METHODS, AND RESULTS

> ***RQ2: Did the effect of the KIA offer on student AP examination-taking and performance differ after two years relative to after one?***

Our second research question asked whether the effect of the KIA offer on student AP examination-taking and performance differed after two years relative to after one. We compared student outcomes of teachers in their second year of the KIA offer (i.e., those in schools randomized to the treatment group in Year One) with student outcomes in their first year of the KIA offer (i.e., those in schools randomized to the control group in Year One). In treatment teachers' second KIA year, they continued to have full access to curriculum and instructional materials through Sprocket. For treatment teachers in their second KIA year, professional development supports were optional, and on-demand coaching was not offered; few second-year teachers participated. In 2017-18, control teachers in their first year of the KIA offer had access to the full KIA intervention, including all the same professional development supports received by treatment teachers in 2016-17.

Research Question 2 sample

Schools and teachers from the Year One sample were included in the Year Two sample if teachers taught APGOV or APES in both 2016-17 and 2017-18 in their randomized school. Of the 74 teachers across 68 schools in the Year One sample, 53 treatment and control teachers across 50 schools persisted into the Year Two sample.

Therefore, after losing six schools post-randomization by the end of September of Year One, an additional 18 schools—10 treatment, 8 control—left the sample by September of Year Two. A school left the sample if it did not include a volunteering teacher of APGOV or APES teaching the course in 2017-18. From Year One to the Year Two, school-level attrition was 27% overall with a 11-percentage point differential between treatment and control schools, showing higher attrition in the treatment group. From randomization to the Year Two sample, school-level attrition was 32% overall with a 22-percentage point differential. Both levels of attrition exceeded WWC thresholds.

School-level attrition was not uniform across districts; rather, it was concentrated in Districts A and E, with 14 of 18 schools dropping from those two districts (seven from each; see Appendix F). According to districts' historical records of when schools offered APGOV or APES, 33% of District A schools offered APGOV or APES every year while 17% of District E schools offered APES every year.[6] Thus, the observed high overall level of school attrition from Year One to Year Two in these two districts was unavoidable. Because both districts served high proportions of students from low-income households—71% in District A, participating in the study with APGOV and APES, and 83% in District E, which participated with APES only—the composition of schools serving primarilty students from lower-income households decreased across the five districts between Years One and Two (Appendices F and G). As a result, a lower proportion of Year Two schools were Title 1 (approximately 65%) relative to Year One schools (approximately 72%).

Among the 53 teachers' 2017-18 students (n=2,946) included in the Research Question 2 sample, across treatment conditions 47% were eligible for free or reduced-price lunch (46% among exam-takers) and 45% were Black or Hispanic (43% among exam-takers).

Research Question 2 analytic methods overview

We used the same analytic methods to address Research Question 2 used to address our first research question (Appendix J). Due to effects of attrition on the sample's composition by district-level socio-economic status—as well as concerns about power—we do not report sub-group analyses in our Year Two analyses (i.e., research questions 2 and 3). Our Year Two robustness checks included addressing sensitivity to covariates and investigating the possibility that non-random sorting of students into KIA treatment or control classrooms could have biased results.

[6] *Of the five districts, three (Districts A, C, and E) provided historical course offerings in the three academic years prior to KIA implementation (2013-14 to 2015-16), whereas District B provided information for two years prior (2014-15 and 2015-16) and District D provided for one prior year (2015-16). In District A, 33% of schools offered APGOV or APES every year. The corresponding percentages for Districts B, C, D, and E were 45%, 18%, 81%, and 17%.*

Research Question 2 results

BASELINE EQUIVALENCE

Attrition on the qualifying score outcome among exam-takers only (n=2,311) was within WWC thresholds. However, student-level attrition on the continuous score outcome samples exceeded the WWC thresholds, so we do not report results for this outcome (Appendix K).

For the analysis of Research Question 2 in Year Two, no student-level covariates exceeded WWC baseline equivalence thresholds, and no p-values were smaller than 0.05 on any standardized mean differences at the student, teacher, or school level. However, two teacher-level variables exceeded the WWC threshold for the qualifying score outcome in both the full sample and that of exam-takers only. Notably, baseline (2015-2016) students of treatment teachers in their second year of KIA had a lower propensity to earn qualifying scores on the APES/APGOV exam than did the baseline students of the control teachers in their first year of the KIA offer. This difference implies that, in the absence of any treatment, we would expect fewer qualifying scores among students of teachers in their second year of KIA compared to what we would expect of first-year KIA teachers' students. In addition, teachers in their second year of the KIA offer had fewer years of experience teaching APGOV or APES relative to teachers in their first year of the KIA offer, and this difference also exceeded the WWC threshold. (In Appendix U, we further discuss Research Question 2 baseline equivalence.)

IMPACT ESTIMATES

When fitting covariate-adjusted models, estimated effect sizes are all positive but non-significant (Table 4). The positive nature of the effect sizes suggests students of second-year KIA teachers may take the exam at higher rates and may perform better on the exam than do students of first-year KIA teachers. The lack of significance means these results may have been due simply to chance. We cannot determine with certainty that students of teachers in their second year of the KIA offer performed better than students of teachers in their first KIA-offer year, though we did not observe evidence of erosion of effects.

TABLE 4. Covariate-adjusted estimates of the differences in AP performance between students of teachers with two years of the KIA offer relative to one year of the KIA offer

	2 years vs. 1 year of the KIA offer			
	EFFECT SIZE (SE)	**95% CI**	**p-value**	**n**
Took AP exam	0.222 (0.18)	(-0.125, 0.568)	0.33	2946
Qualifying score (full sample)	0.212 (0.13)	(-0.05, 0.473)	0.34	2946
Qualifying score (exam-takers only)	0.087 (0.14)	(-0.196, 0.371)	0.57	2311
AP Total Score	0.163 (0.11)	(-0.057, 0.383)	0.13	1424
AP Multiple Choice Score	0.151 (0.11)	(-0.067, 0.37)	0.17	1424
AP Free Response Score	0.126 (0.11)	(-0.085, 0.337)	0.20	1424

Notes: Table columns show standardized effect sizes, standard errors, confidence intervals, p-values, and sample sizes.

Like for Year One, in Year Two we did not find evidence of systematic sorting of students into treatment or control classrooms (Appendices P and O).

Research Question 2 limitations

As was the case for Research Question 1, limitations to Research Question 2 results included school-level attrition and model dependence, with results contingent upon covariate adjustment. An additional limitation to the results of Research Question 2 was lack of baseline equivalence on key measured teacher-level characteristics, such that in the year prior to the KIA offer (2015-16), students of teachers in their second year of the KIA offer had notably lower APGOV/APES performance outcomes compared to students of teachers in their first KIA offer year. Adjustment for this baseline difference was crucial to our estimated impacts, with most of the effect size difference between unadjusted and adjusted results explained by the teachers' baseline students' May 2016 APGOV/APES exam performance.[7]

RESEARCH QUESTION 3: SAMPLE, ANALYTIC METHODS, AND RESULTS

> ***RQ3: What is the effect of the second year of the KIA offer on AP examination-taking and performance relative to no access?***

In the 2017-18 school year, all teachers, no matter their original assignment to treatment or control, had received the KIA offer. Therefore, we did not have a pure experimental comparison group with no KIA experience. We pursued two approaches, each with its own defined sample of teachers, to estimate differences in student outcomes after two years of the KIA intervention offer relative to zero years. Approach 1 relied on comparisons within the sample of teachers who volunteered to participate in the KIA RCT, while Approach 2 compared students of teachers who volunteered to participate in the RCT against those whose teachers did not. Limitations are more pronounced for results of this third research question relative to the first two.[8]

APPROACH 1: ESTIMATING THE TWO-YEAR EFFECT OF KIA WITHIN THE RCT SAMPLE OF VOLUNTEERING TEACHERS

First, we used a novel approach to estimate the difference between two and zero years of the KIA offer, using the sample of teachers who volunteered to participate in the RCT and continued to teach APGOV or APES in 2017-18. To estimate the effect, we broke down the estimate into two components:

---

[7] *To explore the consequence of the baseline difference, we fit models with the teachers' baseline students' May 2016 APGOV/APES exam performance and treatment status as the sole predictors of outcomes.*

[8] *Two potential alternatives to Approach One may have been to compare treatment teachers' students in Year Two (2017-2018) to their baseline (2015-2016) pre-KIA students, or to directly compare treatment teachers' Year Two students to control teachers' Year One students. However, comparisons over time, including before-after comparisons, are well known to be inferior as causal estimands. Any natural changes over time—in students, teaching, schools, context, or outcome measures themselves—become completely confounded with the estimated treatment effect. For this reason, we restricted our analyses to approaches directly comparing treatment and control students within the same year.*

the effect of one year of the KIA offer compared to none, and then any additional effect from the first year of KIA to the second.

APPROACH 1 SAMPLE

Our Research Question 3 Approach 1 sample was comprised of the 53 teachers across 50 schools who volunteered to participate in the RCT, taught APGOV or APES during the 2016-17 and 2017-18 school years, and kept teaching the course in their originally randomized school. The teachers are the same as those examined in Research Question 2—but in this approach, we included the students of the 53 teachers across both their 2016-2017 (n=3,100) and 2017-2018 (n=2,946) cohorts.

APPROACH 1 ANALYTIC METHODS OVERVIEW

We used most of the same analytic methods to address Research Question 3 as we used to address our first two research questions (Appendix J), with an important exception. To estimate the impact of two years of KIA as compared to no KIA, we first estimated the effect of one year relative to zero years of the KIA offer (like Research Question 1, but limited to students of teachers who persisted into Year Two). Then we estimated the effect of two years of the KIA offer relative to one year (Research Question 2). Finally, we added the two estimated coefficients and translated them into effect sizes as to estimate the difference between two and zero years of the KIA offer.

APPROACH 1 RESULTS

The 2017-18 student-level attrition (Appendix K) and baseline equivalence (Appendix U) results for the outcome samples used to address Research Question 2 also apply to the Research Question 3 2017-18 student sample (n=2,946). Baseline equivalence patterns for the 2016-17 student sample (n=3,100) informing Research Question 3 followed a similar pattern. Teacher-level covariates describing first-year KIA teachers' 2015-16 (baseline) students' propensity to earn a qualifying score on the APES/APGOV examination and years of experience teaching APGOV/APES did not exceed WWC thresholds for the full sample outcome (they did for the exam-takers only sample), though we observed the same direction of differences as Year Two, with standardized mean differences close to the WWC threshold.

By adding the covariate-adjusted estimates from Year 1 and Year 2 models, we observed students of teachers with two years of the KIA offer were significantly more likely to earn a qualifying score among the full sample and among only exam-takers, relative to students of teachers without the offer (Table 5).

TABLE 5. Covariate-adjusted estimates of differences in AP performance outcomes between students of teachers with one and zero years of the KIA offer, one and two years of the KIA offer, and two and zero years of the KIA offer (n=53 teachers across 50 schools).

	1 VS. 0 YEARS OF KIA OFFER			2 VS. 1 YEARS OF KIA OFFER			2 VS. 0 YEARS OF KIA OFFER	
	EFFECT SIZE (SE)	p-value	n	EFFECT SIZE (SE)	p-value	n	EFFECT SIZE (SE)	p-value
Took AP exam	0.025	0.910	3100	0.222	0.33	2946	0.246	0.410
Qualifying score (full sample)	0.374*	0.017	3100	0.212	0.34	2946	0.586*	0.046
Qualifying score (exam-takers only)	0.440*	0.016	2537	0.087	0.57	2311	0.527*	0.047
AP Total Score	0.186	0.091	1318	0.163	0.13	1424	0.349	0.052
AP Multiple Choice Score	0.204	0.081	1318	0.151	0.17	1424	0.355	0.057
AP Free Response Score	0.154	0.130	1318	0.126 0.33	0.20	1424	0.279	0.084

**p<.05 We calculated all 2-year vs. 0-year p-values via randomization-based inference, harnessing the original randomization scheme. We also used randomization-based inference to calculate p-values on models with singularity across one or more of 20 imputed datasets—the qualifying score (full sample) and qualifying score (exam-takers only) impact models.*

While not significant, the magnitude of the effect sizes describing the KIA exam-taking effect for two versus zero years of KIA was positive and relatively high in magnitude (ES=0.246, p=0.410), suggesting more students of teachers with two years of the KIA offer took the exam than students of teachers with no KIA offer. Notably, 90% of the two-year exam-taking effect stems from increases in exam-taking comparing second-year KIA teachers' students to those of first-year KIA teachers (ES=0.222), relative to 10% coming from first-year teachers' students relative to no KIA (ES=0.025). This implies that if the KIA two-year exam-taking effect is a function of KIA, as opposed to just chance, the increases in exam-taking happen predominantly in a teacher's second year of their KIA offer.[9]

We estimated the proportion of KIA students who would earn a qualifying score on the exam with second-year KIA teachers—among all sample students not just those who actually took the exam—would be greater than the proportion of control students by approximately 10 percentage points[10] (ES=0.586, p=0.046). Depending on the unique characteristics of the students in the sample, we predicted 34.1% of the full sample would have earned a qualifying score on the exam without KIA,

[9] *Research Question One results showed no exam-taking effect among 3,645 students of 74 teachers across 68 randomized schools (ES=-0.009, p=0.95).*

[10] *For each of the 53 teachers' 3,100 students in the Year One sample, based on their covariates and the model comparing one year of the KIA offer to zero years of the offer, we estimated their probability of earning a qualifying score on the AP exam with no KIA. We then obtained a predicted probability of earning a qualifying score with a second-year KIA teacher by scaling up each of the estimated control probabilities according to the effect size describing two-years of the KIA offer relative to no KIA offer.*

compared to 44.1% for students of teachers in their second year of KIA experience. Of the total two-year qualifying score (full sample) effect, 64% of the effect size stems from the first year and 36% from the second year. On the qualifying score (exam-takers only) outcome, conditional upon exam-taking, the effect size (ES=0.527, p=0.047) translates to an increase of 9.7 percentage points; 41.6% of the exam-taking students in our sample would earn qualifying scores without KIA (conditional upon exam-taking), as compared to 51.3 % with two years of the KIA intervention. Of this total two-year effect for exam-takers only, 83% is from the first year and 17% from the second year—meaning that the vast majority of improvement in exam scores themselves comes with a teacher's first year of the KIA offer. (See Approach 1 sensitivity analyses in Appendix V.)

ADVANTAGES OF AND LIMITATIONS TO APPROACH 1 ESTIMATES

The key advantages of our first approach are randomization and lack of self-selection bias. Each component of our Approach 1 estimates come from a comparison of groups originally randomly assigned, albeit with subsequent attrition. Moreover, all teachers included in this approach taught in randomized schools and volunteered to participate in the RCT study, thus sharing similar unobserved characteristics, like motivation to try a new instructional approach.

Approach 1 suffers from the same limitations as Research Question 2: School-level attrition exceeded acceptable thresholds, and the analytic outcome samples lacked baseline equivalence on key measured teacher-level covariates (i.e., measures of their baseline students' APGOV/APES performance), such that all estimates required covariate adjustment and were model-dependent.

In addition, Approach 1 assumes the impact of KIA for teachers in their first year in 2016-17 was the same as the impact of KIA for teachers in their first year in 2017-18. This assumption was likely violated for various reasons; for example, because of year-to-year contextual changes, changes in the composition of student groups from year to year, and/or because the KIA program designers implemented changes designed to improve to the KIA curriculum and professional development supports over time. Another limitation was inflated variance from summing two different estimates, resulting in decreased statistical power and greater uncertainty around estimated effect sizes (Appendix W).

APPROACH 2: ESTIMATING THE TWO-YEAR EFFECT OF KIA BY COMPARING YEAR TWO RCT TEACHERS TO A MATCHED GROUP OF TEACHERS WITH NO KIA EXPOSURE

Our second approach to estimating the difference in student outcomes between two and zero years of the KIA offer used propensity-score matching to select a comparison group of teachers with no KIA exposure from within the five participating KIA districts and comparing their students' May 2018 AP performance to that of treatment teachers with two years of the KIA offer.

APPROACH 2 SAMPLE AND METHODS OVERVIEW

Non-volunteering teachers were among the universe of eligible teachers who did not enroll in the KIA Efficacy Study, either because they chose not to, could not attend the 2016-17 Summer Institute (a prerequisite for enrolling), or did not know about it. They did not have any access to the KIA intervention. Like volunteering teachers, non-volunteering teachers had to have taught APGOV or APES in 2016-17 and 2017-18.

Pre-match, volunteering RCT treatment teachers differed markedly from non-volunteering teachers. Notably, 29% of volunteering treatment teachers' students earned a qualifying score on the baseline (i.e., May 2016) APGOV or APES exam, relative to 47% of non-volunteering teachers' baseline students. Other measures of teachers' baseline students' academic achievement followed this same pattern. Determining which non-KIA teachers were most similar to KIA teachers at baseline presented challenges requiring multiple decision points (e.g., whether or not to allow matches within experimental schools) with tradeoffs to every decision. Ultimately, we used two different matching algorithms to choose two different matched groups, then examined the quality of both sets of matches and compared student outcomes between students of teachers with two years of the KIA offer to each of the two matched groups.[11]

APPROACH 2 RESULTS

Even after matching, each set of comparison groups suffered from persistent baseline imbalance on student-, teacher-, and/or school-level covariates (Appendices BB and CC). Subsequent impact models fit with the two sets of matches resulted in estimates in opposite directions—one negative, one positive (Appendices DD and EE).

The marked discrepancy in baseline balance on observed measures pre-match—suggesting similar discrepancy in unobserved measures that cannot be addressed through matching or statistical adjustments—persistent baseline imbalance in the matched set of teachers (after two matching attempts), and opposing results from the two sets of models deem results from Approach 2 uninterpretable.

ADVANTAGES AND LIMITATIONS TO APPROACH 2 ESTIMATES

The key advantage of Approach 2, relative to Approach 1, was that it did not require assuming a constant first-year treatment effect from Year One to Year Two. Also, without having to sum two estimates, we would have expected greater precision of estimates from Approach 2.

The major limitation of this approach was bias due to self-selection into study conditions. The AP performance of volunteering teachers' prior students was dramatically different from non-volunteering teachers' students prior to matching. While matching and subsequent covariate adjustment can improve upon observed differences, they cannot resolve unobserved differences that may exist between teachers who volunteered to participate in an innovative and intensive PBL intervention and those who did not. The measured and potential unmeasured differences between volunteering and non-volunteering teachers after matching and covariate adjustment injects doubt into either set of results from Approach 2.

In summary, Approach 2 results were inconclusive. Future research may show the distribution of effect sizes across a range of approximately 30,000 possible matches.

[11] *Appendix X provides the details of the first matching algorithm, in which we did not permit matched non-experimental teachers to have taught in the same school as experimental teachers, In Appendix Y are tables describing teachers in their second year of the KIA offer compared to unmatched and matched teachers, as well as the teachers' students, and schools. In the second matching algorithm, we permitted matched non-experimental teachers to have taught in the same school as experimental teachers (Appendix Z). Appendix AA describes teachers in their second year of the KIA offer compared to unmatched and matched teachers, as well as the teachers' students, and schools.*

RESEARCH QUESTION 4: SAMPLE, ANALYTIC METHODS, AND RESULTS

RQ4: How did teachers and students experience the KIA intervention?

Our fourth research question addressed how teachers and students experienced the KIA intervention, including their perceived challenges and benefits. To address this question, we collected implementation-related data in both years. In Year One, we collected data from students, teachers, instructional coaches, school leaders, and district staff. In Year Two, we collected data only from teachers. The Year Two samples were limited, and we do not include those results in this document. (In Appendices FF and GG, respectively, we provide details about the Year One and Year Two implementation analyses.)

Implementation sample

We limited our implementation analyses to teachers who "complied" with their schools' assigned treatment status. This is a subset of the RCT teacher sample. We define compliance as having participated in the KIA professional development program for at least one day or logging into Sprocket at least once.[12]

In Year One, 31 of 35 (89%) treatment teachers taught APGOV or APES during the 2016-17 school year and complied with their assigned treatment status. The 31 teachers across 27 schools were treatment "compliers," and all 39 control teachers across 37 control schools also were compliers. Of the 70 teachers, all 31 treatment teachers used materials accessed through Sprocket while none of the 39 control teachers accessed Sprocket. KIA Summer Institute attendance among complier treatment teachers ranged from 90 to 97% (one did not participate at all). KIA Professional Development Session attendance ranged from 90% for the first session to 68-74% attendance across the three subsequent sessions.[13]

For our Year One implementation analysis, we included at least one form of collected data from all complier teachers. Collected data included surveys, instruction logs, and interviews with students, teachers, and instructional coaches, with most response rates by teacher data type above 80% and balanced between treatment and control groups.

Implementation analysis methods

To address our Year One implementation analysis research questions, we analyzed first-year field data describing teachers' implementation of the KIA program—including surveys, instruction logs, and interviews with students, teachers, and instructional coaches—using appropriate quantitative and qualitative methods and triangulated results across participant types.

---

[12] *Appendix II provides details about teachers' use of Sprocket in Year One. Appendix JJ gives details about teachers' participation in, and provision of, the KIA professional development program in Year One. In Appendix KK we share a summary of Year One professional development observations.*

[13] *In Year 2, for treatment teachers in their second year of KIA implementation, we carried over complier status from Year One, as this indicator reflected their participation in the intervention when it was offered. That is, a non-complier in 2016-17 could not become a complier in 2017-18, and vice versa. For control teachers, all "complied" with control status in Year One because none received the treatment. In Year Two, we created a complier flag based upon teachers' participation in the KIA intervention in 2017-18. Of the 23 KIA second-year teachers in 2017-18, 20 were compliers (87%). Of the 30 first-year KIA teachers in 2017-18, 26 were compliers (87%).*

Year One implementation results

Drawing from triangulation of Year One data collected from complying teachers, their students, and KIA instructional coaches, this study demonstrated that teachers who used KIA's curriculum and supports changed their APGOV and APES pedagogy. Across courses, and under conditions of optimal professional learning supports and compared to randomly assigned, business-as-usual control teachers, treatment teachers in their first year of KIA implementation placed greater emphasis on deeper learning objectives, more frequently used student-centered pedagogy and in ways their students felt were authentic, and less frequently lectured or relied upon explicit exam preparation. For the most part, teachers across courses sustained their use of the KIA approach throughout the year and felt the KIA approach aligned to the AP curriculum framework and examinations. Their students reported feeling prepared for their relevant AP examinations, such as one student who explained, "I feel like the exam was a lot easier than any test we've ever taken in this class." Another student elaborated:

> ***"I take a couple of AP classes, and I think this is one of the easier ones for me to understand and grasp more because I am a hands-on learner. All the other ones, it's kind of like you sit in a class and you take notes, and then you don't understand those notes, and then you fail the test and so on. I think this class made us more involved in what we were learning, so it was easier to grasp."***

The study also illustrated a duality in which teachers and students perceived KIA as beneficial and helpful, while also challenging. Both teachers and students found the transition difficult in the area of students' comfort with responsibility for driving their own learning—a new experience for many students. Students experienced "project fatigue," wanting to rest and listen to lectures between back-to-back project units. In a group interview, a student expressed desire for balance between projects and more traditional, direct transmission modes of instruction:

> ***"Projects were more fun than sitting and watching someone, but I feel like there needs to be a better balance between just sitting and watching someone lecture and projects. I think it needs to be almost 50/50, because you'd be able to retain the information more. Like, the projects were fun—it was just too much, too many, and not enough time."***

Teachers found it difficult to balance transmission and student-centered approaches, while also experiencing issues with pacing and facilitating group work. Additionally, they voiced concern over whether the curriculum sufficiently prepared students to earn a qualifying score on the AP examinations.

The AP context is particularly challenging because of the sheer amount of content covered in the course-specific AP curriculum frameworks and the looming end-of-year, high-stakes examination. The unique AP contextual demands—on top of the challenge of shifting from a transmission to PBL approach in any context—meant the observed, sustained shift in KIA teachers' AP practice in virtually all critical aspects of instruction was not assured. Yet despite the challenges and changes in

pedagogy, including less reported time spent on explicit AP examination-taking skills, students across the two courses still felt prepared for the examinations, and the majority of teachers recommended the approach, citing benefits for themselves and their students.

Limitations to Year One implementation analysis

All data from interviews, focus groups, instruction logs, and surveys was reported from the perspectives of individual principals, teachers, and students, and so was subject to the drawbacks of self-reported data. Teachers' self-reported responses about their KIA implementation may have been particularly subject to potential over-reporting bias, as teachers may have said they implemented more KIA practices than they actually did. For this reason, we cross-referenced, or "triangulated," data across sources to verify self-reported responses from one group against other groups' responses. Though we triangulated responses, highlighting any discrepancies, response bias was a concern.

STUDY-WIDE LIMITATIONS

This study has several limitations, some of which apply across all presented analyses, while others are specific to one analysis type. We presented analysis-specific limitations throughout the report, and here state study-wide limitations.

Limitations related to external validity apply to all study results. All results apply to teachers within the participating districts who chose to enroll in the KIA RCT. The KIA Efficacy Study was, by definition, a test of the efficacy of KIA under ideal conditions. The five participating districts were not representative of all districts offering AP courses. Rather, they are districts supporting a teaching and learning approach philosophically aligned with KIA; offering AP courses at a great enough number of individual high schools to warrant inclusion in the RCT; interested enough in KIA to agree to participate; and requiring open-access AP course enrollment.

The teachers and schools volunteering to participate in the KIA Efficacy Study were "early adopters," and may not have delivered the courses in a way representative of large-scale implementation. As our complementary analysis including non-volunteering teachers demonstrates, those who volunteered to enroll in the KIA RCT were not representative of teachers who did not volunteer. Non-volunteering teachers' baseline students' APGOV/APES exam performance exceeded those of volunteering RCT treatment teachers by 18 percentage points. For the second-year analyses, results are applicable to a select group: teachers who volunteered to enroll and persisted to teach APGOV or APES in 2016-17 and 2017-18 in their randomized school.

In addition, the inextricability of the PBL approach's effect from the effect of the professional development and coaching part of the intervention means that in this study it is impossible to disentangle the separate influences of professional development and PBL on students' AP performance outcomes. Also, due to the level of professional learning supports provided, "efficacy" study conditions for teachers in their first year of the KIA offer in 2016-17 and 2017-18 were optimal compared to conditions that would be realistic in an "effectiveness" study.

IMPLICATIONS AND CONCLUSION

Despite the limitations, the results contribute to the PBL research base in several ways, with notable implications for practitioners and policymakers. Under optimal conditions of teacher support, the Year One pattern of results suggests a positive KIA impact on students' propensity to earn qualifying scores on AP examinations and on underlying continuous AP scores. Based on our primary Year One covariate-adjusted models, we estimated the percentage of all students in our sample earning a qualifying score would be about 4 percentage points higher with KIA, and those earning a qualifying score among exam-takers would be about 8 percentage points higher. Earning qualifying scores on AP examinations can earn students college credit, and relates to enrolling and persisting in college (Finn & Scanlan, 2019; Sadler, 2010; Smith, Hurwitz & Avery, 2017). At the end of the first year, we observed the positive pattern of results within both courses as well as pooled across courses. The pattern also was positive within respective groups of students from lower- and higher-income households, in districts serving a majority of students from lower-income households, in districts serving a majority of students from higher-income households, and within each of the five participating districts.

Strengthening causal claims is the RCT design, as well as the statistical significance, magnitude, and robustness of estimated effect sizes across multiple covariate-adjusted sensitivity analyses. Weakening causal claims are high school-level attrition post-randomization, and differences between unadjusted results and those statistically adjusted to address observed baseline differences.

Results also contribute to a narrow body of evidence on whether teachers' proficiency implementing PBL curriculum and instructional approaches—and, more generally, complex curricula—changes after their first year. Year Two results were less conclusive due to a lack of significance in second- versus first-year estimates, and substantial limitations to estimates of two years of the KIA offer relative to none. Though the impact of a teacher's continued experience with KIA may benefit their students, we observed the majority of gains in AP exam performance within the first year of the KIA offer. Teachers did not require more than one year of KIA experience before their students' AP performance benefitted. The only outcome with a different second-year trend was students' propensity to take the AP examination. In contrast to student AP exam performance for which we observed the majority of the effect in Year One, we observed nearly all of the exam-taking effect in Year Two. Notably, we did not observe erosion of student AP performance benefits in a teacher's second year of the KIA offer after teachers discontinued participation in KIA professional development. We estimate that the percentage of students in our sample who would earn a qualifying score is about 10 percentage points higher for students taking the course from a second-year KIA teacher, as opposed to students learning without KIA. This estimate is limited to students of teachers who persisted into our Year Two sample and applies to both the full sample of these students and only exam-takers.

Investigation of Year One implementation revealed teachers felt KIA was more engaging for students, offering the opportunity for them to develop real-world skills. Though KIA treatment teachers found the change considerable, with challenges in pacing and groupwork facilitation, the majority recommended the approach, citing benefits for themselves and students. Treatment students voiced benefits related to civic engagement, group work, engagement with learning, and examination preparation. The take-away message—students of KIA teachers outperformed students of control

on AP exams—is even more practically meaningful given both teachers and students perceived other benefits of KIA beyond exam performance.

The KIA intervention was comprised of a combination of curriculum and instructional materials, and professional development supports. Particularly in a high-stakes AP setting, shifting from primarily transmission to PBL instruction is a substantial change for teachers, suggesting the need for ongoing, job-embedded professional learning and coaching support. In a teacher's second year, KIA supports were optional and did not include access to on-demand coaching; few teachers participated. Impact on student AP performance occurring primarily in teachers' first year of the KIA offer aligns with the intensive provision of professional development in the first year. Lack of observed erosion of the KIA impact on students' AP performance in teachers' second year of the KIA offer suggests the costs to shift to PBL do not require annual payment of professional development expenses.

Another study finding that can inform districts' implementation of the KIA intervention is the high school-level attrition between randomization and Year One, and between the first and second years. Important for district and school leaders to consider as they plan for use of the KIA intervention is the expectation that schools may not offer APGOV or APES courses across consecutive years.

Another consideration for district implementation of KIA is the marked differences between teachers who volunteered to enroll versus those who did not enroll in the KIA RCT. The prior students of teachers who volunteered to enroll performed considerably worse on the May 2016 APGOV or APES examinations than students of teachers who did not.

Of particular importance to scaling the KIA approach beyond the RCT study are treatment teachers' self-reported perception that KIA aligned to the AP curriculum framework and examinations, and students' feelings of learning more deeply and being prepared for the AP examinations. Critical to scaling will be KIA teachers' positive perceptions of the approach across courses and their recommendations of KIA to others.

In conclusion, the results of the Knowledge in Action Efficacy Study support teacher-driven adoption of the KIA approach in both APGOV and APES courses, among districts with open-enrollment AP policies that support project-based learning, and for students from both lower- and higher-income households.

REFERENCES

Altonji, J. G. & Mansfield, R. K. (2018). Estimating Group Effects Using Averages of Observables to Control for Sorting on Unobservables: School and Neighborhood Effects. American Economic Review 108 (10) 2902-46.

Andrews, D., & Lewis, M. (2007.) Transforming Practice from Within: The Power of the Professional Learning Community. In L. Stoll & K.S. Louis (Eds.), Professional Learning Communities: Divergence, Depth and Dilemmas. Maidenhead, UK: Open University Press.

Boaler, J. (1997). Experiencing School Mathematics: Teaching Styles, Sex and Settings. Buckingham, UK: Open University Press.

Balfanz, R., Mac Iver, D. J. and Byrnes, V. (2006). The implementation and impact of evidence-based mathematics reforms in high-poverty middle schools: A multi-site, multi-year study. *Journal for Research in Mathematics Education*, 33-64.

Barron, B., & Darling-Hammond, L. (2008). Teaching for Meaningful Learning: A Review of Research on Inquiry-Based and Cooperative Learning. Powerful Learning: What We Know About Teaching for Understanding. Marin County, CA: Edutopia.

College Board (2020). Archived AP data. Retrieved from https://research.collegeboard.org/programs/ap/data/archived/ap-2019

College Board (2019). Student Participation and Performance in Advanced Placement Rise in Tandem Retrieved from https://newsroom.collegeboard.org/student-participation-and-performance-advanced-placement-rise-tandem

Cognition and Technology Group at Vanderbilt (1992). The Jasper Series as an Example of Anchored Instruction: Theory, Program Description, and Assessment Data. Educational Psychologist, 27(3), 291-315.

Condliffe, B., Visher, M. G., Bangser, M. R., Drohojowska, S., & Saco, L. (2017). Project-Based Learning: A Literature Review. Working Paper. MDRC.

Correnti, R. (2007). An Empirical Investigation of Professional Development Effects on Literacy Instruction Using Daily Logs. Educational Evaluation and Policy Analysis, 29(4), 262-295

Dochy, F., Segers, M. R., Van den Bossche, P., & Gijbels, D. (2003). Effects of Problem-Based Learning: A Meta-Analysis. Learning and Instruction. 13. In Durlak, J., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (Eds). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions Child Development, 82(1), 405-432.

Dole, S., Bloom, L., & Kowalske, K. (2016). Transforming Pedagogy: Changing Perspectives from Teacher-Centered to Learner-Centered. *Interdisciplinary Journal of Problem-Based Learning, 10*(1)

Drake, K. N., & Long, D. (2009). Rebecca's in the Dark: A Comparative Study of Problem-Based Learning and Direct Instruction/Experiential Learning in Two Fourth-Grade Classrooms (Abstract). Journal of Elementary Science Education, 21(1), p 1-16.

Duke, N. K., Halvorsen, A.-L., Strachan, S. L., Kim, J., & Konstantopoulos, S. (2020). Putting PjBL to the Test: The Impact of Project-Based Learning on Second Graders' Social Studies and Literacy Learning and Motivation in Low-SES School Settings. *American Educational Research Journal.*

Duncan, A. & Ryan, A. (2021). After the siege, we need civics education. New York Daily News. January 9, 2021 retrieved from https://www.nydailynews.com/opinion/ny-oped-after-the-siege-civics-education-20210109-w3zfpwu66zhdhp5f2jy3sw67re-story.html

Finkelstein, N., Hanson, T., Huang, C. W., Hirschman, B., & Huang, M. (2010). Effects of Problem Based Economics on High School Economics Instruction (NCEE 2010-4110). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West.

Finn, C. & Scanlan, A. (2019). Learning in the Fast Lane: The Past, Present, and Future of Advanced Placement. Princeton, NJ: Princeton University Press.

Freer-Alvarez, T. A. (2016). *The Impact of a Project-Based Learning Comprehensive School Reform on Student Achievement in a Group of High-Population Bilingual Urban Campuses* (Doctoral dissertation, Texas A&M University).

Fullan, M. (1993). *Change forces: Probing the depths of educational reform* (Vol. 10). New York: Psychology Press.

Fullan, M. G. and Miles, M. B. (1992). Getting reform right: What works and what doesn't. *Phi delta kappan, 73*(10), 745-752.

Gardner, H. (1999). Disciplined Mind: What All Students Should Understand. New York, NY: Simon & Schuster.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What Makes Professional Development Effective? Results from a National Sample of Teachers. American Educational Research Journal, 38(4), 915-945.

Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized Test Outcomes for Students Engaged in Inquiry-Based Science Curricula in the Context of Urban Reform. Journal of Research in Science Teaching, 45(8), 922-939.

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of Problem-Based Learning: A Meta-Analysis from the Angle of Assessment. Review of Educational Research, 75(1), 27–61.

Gould, J., Jamieson, K. H., Levine, P., McConnell, T., & Smith, D. B. (2011). Guardian of Democracy: The Civic Mission of Schools. Philadelphia, PA: Lenore Annenberg Institute for Civics of the Annenberg Public Policy Center and the Campaign for the Civic Mission of Schools.

Harris, C. J., Penuel, W. R., D'Angelo, C. M., DeBarger, A. H., Gallagher, L. P., Kennedy, C. A., Krajcik, J. S. (2015). Impact of Project-Based Curriculum Materials on Student Learning in Science: Results of a Randomized Controlled Trial. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/tea.21263/abstract.

Hernandez-Ramos, P., & De La Paz, S. (2009). Learning History in Middle School by Designing Multimedia in a Project-Based Learning Experience (Abstract). Journal of Research on Technology in Education, 42(2), 151-173.

Hillygus, S. & Holbein, J. (2020) Making Young Voters: Converting Civic Attitudes into Civic Action, Cambridge, UK: Cambridge University Press.

Hmelo-Silver, C. E. (2004). Problem-Based Learning: What and How Do Students Learn? Educational Psychology Review, 16(3), 235-266.

Jackson, K., & Makarin, A. (2018). Can Online Off-the-Shelf Lessons Improve Student Outcomes? Evidence from a Field Experiment. American Economic Journal. Economic Policy, 10(3), 226–254.

Joyce, B., & Showers, B. (2002). Student Achievement Through Staff Development. Retrieved Jan. 16, 2018 from https://www.nationalcollege.org.uk/cm-mc-ssl-resource-joyceshowers.pdf.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., Puntambekar, S., & Ryan, M. (2003). Problem-Based Learning Meets Case-Based Reasoning in the Middle-School Science Classroom: Putting Learning by Design into Practice. Journal of the Learning Sciences,12(4), 495-547.

Kolluri, S. (2018). Advanced Placement: The Dual Challenge of Equal Access and Effectiveness. *Review of Educational Research, 88*(5), 671–711.

Larmer, J., Mergendoller, J. R., & Boss, S. (2015). *Setting the standard for project based learning: A proven approach to rigorous classroom instruction*. Alexandria, VA: ACSD.

Levine, P. & Kawashima-Ginsberg, K. (2015). Civic Education and Deeper Learning. Students at the Center: Deeper Learning Research Series. Boston, MA: Jobs for the Future.

Levine, P., & Kawashima-Ginsberg, K. (2017). The republic is (still) at risk and civics is part of the solution. Washington, D.C.: Democracy at a Crossroads National Summit.

Little, J. W. (2002). Locating Learning in Teachers' Communities of Practice: Opening up Problems of Analysis in Records of Everyday Work. Teaching and Teacher Education 18(8), 917–946.

Maxwell, N., Mergendoller, J. R., & Bellisimo, Y. (2005). Problem-Based Learning and High School Macroeconomics: A Comparative Study of Instructional Methods. The Journal of Economic Education, 36(4), 315-331.

Morgan, K. L. and Rubin, D. B. (2012), "Rerandomization to Improve Covariate Balance in Experiments," The Annals of Statistics, 40, 1263–1282.

Murray, I. & Savin-Baden. (2000). Staff Development in Problem-based Learning. Teaching in Higher Education (5) 107-126.

Parker, W. C., Lo, J. C. (2016). Reinventing the High School Government Course: Rigor, Simulations, and Learning from Text. *Democracy and Education, 24*(1), Article 6.

Parker, W., Lo, J., Yeo, A. J., Valencia, S., Nguyen, D., Abbott, R., Vye, M. (2013). Beyond Breadth-Speed-Test: Toward Deeper Knowing and Engagement in an Advanced Placement Course. American Educational Research Journal, 50(6), 1424-1459.

Parker, W., Mosborg, S., Bransford, J., Vye, N., Wilkerson, J., & Abbott, R. (2011). Rethinking Advanced High School Coursework: Tackling the Depth/Breadth Tension in the AP U.S. Government and Politics course. Journal of Curriculum Studies, 43(4), 533-559.

Perkins, D. N., Jay, E., & Tishman, S. (1993). New Conceptions of Thinking: From Ontology to Education. Educational Psychologist, 28(1), 2867-2885.

Rimm-Kaufman, S. E., Fan, X., Chiu, Y. J. and You, W. (2007). The contribution of the Responsive Classroom Approach on children's academic achievement: Results from a three year longitudinal study. *Journal of School Psychology, 45*(4), 401-421.

Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Surveys, New York: John Wiley & Sons, Inc.

Saavedra, A. (2012). From Dry to Dynamic Civic Education Curricula. In D. Campbell, F. Hess, & M. Levinson (Eds.), Making Civics Count: Citizenship Education for a New Generation. Cambridge, MA: Harvard Education Press.

Sadler, P. M., Sonnert, G., Tai, R. H., & Klopfenstein, K. (2010). AP: A Critical Examination of the Advanced Placement Program. Cambridge, MA: Harvard Education Press.

Schneider, R., Krajcik, J., Marx, R. W., & Soloway, E. (2002). Performance of Students in Project-Based Science Classrooms on a National Measure of Science Achievement. Journal of Research in Science Teaching. 39(5), 410-422.

Schwartz, D., & Bransford, J. (1998). A Time for Telling. Cognition and Instruction. 16(4), 475-522.

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

Strobel, J., & van Barneveld, A. (2009). When is PBL More Effective? A Meta-Synthesis of Meta-Analyses Comparing PBL to Conventional Classrooms (Abstract). The Interdisciplinary Journal of Problem-Based Learning, 3(1), 4

Stoll, L., Bolam, R., McMahon, A., Wallace, M., and Thomas, S. (2006). Professional Learning Communities: A Review of the Literature. Journal of Educational Change, 7(4), 221-258.

Supovitz, J. A.,& Turner, H. M. (2000). The Effects of Professional Development on Science Teaching Practices and Classroom Culture. Journal of Research in Science Teaching, 37(2), 963-980.

Thomas, J. W. (2000). A Review of Research on Project-Based Learning, 1-46. Retrieved from http://bie.org/object/document/a_review_of_research_on_project_based_learning.

Tierney, G., Goodell, A., Nolen, S.B., Lee, N., Whitfield, L. & Abbott, R.D. (2020). (Re)Designing for Engagement in a Project-based AP Environmental Science Course, The Journal of Experimental Education, 88:1, 72-102

Tugend, A. (2017). Who Benefits from the Expansion of AP Classes? New York Times Education Issue, September 7, 2017. Retrieved from https://www.nytimes.com/2017/09/07/magazine/who-benefits-from-the-expansion-of-ap-classes.html.

Valant, J., & Newark, D. A. (2017). My Kids, Your Kids, Our Kids: What Parents and the Public Want from Schools. Teachers College Record, 119(12), n12.

Walker, A. & Leary, H. (2009). A Problem-Based Learning Meta Analysis: Differences across Problem Types, Implementation Types, Disciplines, and Assessment Levels (Abstract). Interdisciplinary Journal of Problem-based Learning, 3(1), 12-43.

Wieseman, K.C. & Cadwell, D. (2005). Local History and Problem-Based Learning. Social Studies and the Young Learner. 18 (1) 11-14.

Yoon, K. S., Duncan, T., Lee, S. W-Y., Scarloss, B., and Shapley, K. (2007). Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. U.S. Department of Education Institute of Education Sciences. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_2007033.pdf

Youniss, J. (2012) How to Enrich Civic Education and Sustain Democracy. In D. E. Campbell, M. Levinson, & F. Hess (Eds.), Making Civics Count. Cambridge, MA: Harvard Education Press.

USC Dornsife
Center for Economic and Social Research

Knowledge in Action Efficacy Study Over Two Years, Technical Appendices

FEBRUARY 22, 2021

USC DORNSIFE CENTER FOR ECONOMIC AND SOCIAL RESEARCH
Anna Rosefsky Saavedra, *Principal Investigator*
Ying Liu
Shira Korn Haderlein

GIBSON CONSULTING GROUP
Amie Rapaport, *Co-Principal Investigator*
Marshall Garland
Danial Hoepfner

PENN STATE UNIVERSITY
Kari Lock Morgan, *Co-Principal Investigator*
Alyssa Hu

cesr.usc.edu

TABLE OF CONTENTS

Figures

Tables

Appendix A: Knowledge in Action Intervention Details

Knowledge in Action is a response designed carefully to meet the challenges of traditional AP instruction. The initiative was based on the potential yet under-realized impact AP courses could have on deeper learning for all students. Knowledge in Action values both transmission and inquiry approaches, seeking to blend them as to maximize their respective benefits. The KIA model is based on a foundation of six concepts regarding the science of learning and five design principles. Implementation includes curriculum and instructional resources delivered through an online portal, intensive and ongoing in-person professional development (PD), in-person and virtual coaching, and community developed through both the portal and PD.

Knowledge in Action Theoretical Foundation

Six concepts in the science of learning serve as a theoretical foundation for responding to the challenges (Parker et al., 2013):

1) Accelerated coverage of material at a rapid pace is not strongly associated with learning depth.
2) Depth of transferable learning is preferable to breadth of temporary, non-transferable learning.
3) Assessments requiring students to demonstrate transferable deeper learning, rather than just knowledge and skills, are critical to aligning the focus of instruction with deeper learning.
4) Courses should include transmission and participatory approaches; there is an optimal balance.
5) Instructional sequencing of the transmission and participatory approaches is critical. "The time for telling" is strategic (Schwartz & Bransford, 1998).
6) Students' receptivity to a forced shift in the balance between the transmission and inquiry approaches is important, particularly in the context of a high-stakes exam.

Knowledge in Action Design Principles

University of Washington researchers and their local teacher partners developed the Knowledge in Action courses based on this theoretical foundation and on five design principles: "Rigorous projects as the spine of the course; quasi-repetitive project cycles where each build on the other; engagement that creates a need to know; teachers as co-designers; a course that can scale" (Parker et al., 2013, 1432). The first three principles address the Knowledge in Action learning theory, while the second two are rooted in design-based curriculum development (Parker et al., 2011). Though not an official design principle, another fundamental feature is KIA's focus on developing students' civic knowledge, skills, and motivation. We describe the five design principles and civic education focus below.

Rigorous projects are the spine of the course

KIA projects are "the spine" of the course, as opposed to the more typical "appendage" notion of projects as a special add-on or end-of-course capstone (Parker et al., 2011). Both APGOV and APES curricula are organized into five units. Each unit has a master question driving the unit project as well as other learning activities, including textbook readings, lectures, and class discussions. KIA projects

are intended to provide authentic, real-world challenges, allowing students to direct and reflect on their learning while engaging with their peers and teacher. KIA student learning objectives include mastery of disciplinary content and skills, as well as development and application of sophisticated thinking and communication skills, including public speaking, critical thinking, collaboration, written and verbal communication, and problem-solving.

Quasi repetitive project cycles

The Knowledge in Action curriculum design employs a project cycle described as quasi-repetitive (Bransford et al., 2006) or "looping" (Parker et al., 2011), with the objective of deeper learning as opposed to superficial breadth of learning typical of most AP courses. The course master question—and the questions, ideas, and problems it encompasses—unites project cycles, which teachers and students revisit as they advance through the different projects. Students revisit and revise their understandings from previous projects—looping—while evolving and accumulating their learning through the introduction of new ideas and novel applications of knowledge in subsequent projects. Formative and summative assessments are placed at critical points within the project cycles, so the feedback loop dovetails with the curricular looping.

Engagement first

A third Knowledge in Action principle is "engagement first," following the assertion of Schwartz and Bransford (1998) that student readiness for learning, through reading or lectures, is maximized after students gain initial understanding of an area of inquiry through more active means. The engagement assumption is that initiating learning about a topic through project work will prime students' interest and create a context for learning through reading or lecture. Engagement coming first reverses the usual "grammar" of schooling (Tyack & Cuban, 1995), in which the "telling"—for example, a teacher-delivered lecture on a given topic—occurs prior to project work.

For example, the fifth and final APES unit's initial task of introduces students to the project's goals by having them choose—individually or as a team—a country they will represent in the culminating "Global Climate Summit." The engagement principle posits that students' chosen roles and responsibilities will substantiate their rationale and prompt their motivation to engage with the unit's objectives, which involve learning about international treaties, the structure and function of the earth's atmosphere, the causes and impacts of climate change, and the complexities of solving global environmental issues. During the project's stages, students develop knowledge and skills through research, reading, lecture, and experiential activities, such as collecting samples from a local water source. By the time students reach the course's culminating summit, ideally, they have developed expertise in APES topics to play their role with authority.

Teachers as co-designers

UW researchers and participating teachers began to collaboratively develop the Knowledge in Action APGOV and APES course models beginning in 2008. They based the models on best practices described in the literature of education, cognitive psychology, PBL, and educator development, and followed a design-based approach to curriculum (Brown, 1992). Underpinning this task was the understanding of teachers as the actors engaging most closely with student learning and, thus, their knowledge and expertise should guide curriculum development. Relatedly, the principle of teacher as

co-designer posits that instructors' curricular adaptations (Fogleman, McNeill, & Krajcik, 2011)—in response to their students' interests and needs, as well as their own—should be considered Knowledge in Action model enactment rather than subversion or non-fidelity. Accordingly, the present study interprets teachers' adaptations as central to KIA implementation.

Scalability

Closely related to the co-designer principle, the fifth Knowledge in Action design principle is scalability. The collaborative researcher-teacher team created KIA as a curriculum model that teachers could, and should, adapt so it both informs their practices and is continually informed by their practices. Teachers have regular opportunities to adapt the Knowledge in Action curriculum, materials, and approach in ways sensitive to local contexts. Also, they have a systematized means through which to document their adaptations and supporting rationales.

Knowledge in Action's civic education focus

The Knowledge in Action focus on gaining knowledge, skills, and dispositions through projects with real-world applications aligns with most best practices of civics education. These include: classroom instruction in civics, government, history, economics, geography, and law; simulations of democratic processes; discussion of current events and controversial issues; service learning; extracurricular activities; and school governance (Gould, Jamieson, Levine, McConnell, & Smith, 2011). Through KIA, students learn how and why to engage civically rather than simply absorbing facts about citizenship, as happens in a traditional civics classroom (Saavedra, 2012). Given the Knowledge in Action APGOV and APES content focus in combination with the pedagogical emphasis on gaining knowledge and skills through simulations and discussion, both courses are ideal vehicles through which to promote students' civic engagement.

The Knowledge in Action Curriculum

APGOV and APES courses are offered as electives in most schools, and so do not reach every student. In fact, schools may not even have the capacity to offer them every academic year. However, for two reasons these courses provide ideal settings in which to test the hypothesis that students can deeply learn rigorous content. First, well-defined, well-known, end-of-year examinations with strong psychometric properties—the AP exams administered and scored by the College Board—serve as ideal measures of student outcomes. Impact, as demonstrated through AP scores, is universally easy to interpret as practically meaningful. When students earn qualifying scores on AP exams, they earn college credit accepted by virtually all U.S. colleges as credits towards graduation, saving tuition costs. Earning qualifying scores on AP exams also relates to other critical college outcomes, such as enrolling and persisting (Kolluri, 2018; Sadler, 2010).

Second, the courses are designed to build students' civic, political, and environmental awareness and engagement—areas in which there is widespread agreement that students should develop (Galston, 2001; Valant & Newark, 2017).

The Knowledge in Action APGOV and APES curricula both consist of five units designed to address the knowledge, concepts, and skills included in the College Board's AP curriculum frameworks for those respective courses and examinations. Projects are the core structure within each unit and are

strategically sequenced to build upon the course master question (also referred to as the driving question). The APGOV driving question is, "What is the proper role of government in a democracy?" while the APES driving question is, "How can we live more sustainably?"

Each project has embedded "tasks" investigating a core content area, and "daily lessons" for building background knowledge and honing skills necessary to successfully execute each project. Students can clearly connect the projects, tasks, and units to the real world. For example, APGOV projects include: student debates over historical and contemporary constitutional issues; presidential elections; congressional and Supreme Court simulations; and for the culminating project, creating a political action plan intended to move an agenda item (e.g., immigration policy) through the political system. Through APES projects, students consider the ecological footprints of themselves and their families, management of their community's sustainable resources, environmental resources for farming, and the impact of their community's choices on the ocean. The culminating project of APES requires students to assume the role of delegates to an international climate accord convention.

Differences between KIA AP U.S. Government and KIA AP Environmental Science

APES and APGOV are elective courses that a school may or may not offer every year, depending on factors like student interest and teacher availability. In 2018 across the United States, approximately 320,000 students took the APGOV exam while approximately 166,000 students took the APES exam, making these the fourth- and 13th-most commonly taken across the 38 AP offerings (College Board, 2018).

Apart from content focus, and the specific knowledge and skills emphasized by the respective APGOV and APES curricula, the same design principles and curriculum features apply to both courses, and both designs address their respective AP content guidelines. But there is a difference in terms of how each course curriculum positions the student. The APGOV curriculum requires students to assume the role of others, such as congressmembers, judges, or lobbyists through simulations (e.g., committee hearings and court cases). The APES curriculum is "place-based," with the student, family, and community at the center of lessons, and the focus on students learning how they can act to improve the environment.

APGOV courses should align with the National Council for the Social Studies (NCSS) Standards, while APES courses should align with the Next Generation Science Standards (NGSS). Both should incorporate the Common Core State Standards (CCSS) for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects.

Sprocket Online Curriculum Portal

Sprocket is an online curricular platform, hosted by Lucas Education Research (LER), that supports three main facets of the KIA model. First, it provides teachers with access to the KIA curriculum and curricular resources while also supporting adaptations. For example, through Sprocket, teachers can customize lesson plans, homework, and assessments, and chart a personalized calendar for pacing the curriculum.

Second, Sprocket serves a community function, providing a supportive learning community by connecting AP teachers from schools within a district and across districts. This online community can

combat the isolation some teachers experience in teaching content and/or teaching with KIA and other PBL approaches, and build shared resources and a common platform for interaction and support. Through Sprocket, teachers can share their adaptations, explain the reasons for their adaptations, and upload materials they created or found from other sources. They also can adopt other teachers' adaptations and materials. The platform hosts community forums where teachers share best practices, encourage each other, problem-solve, celebrate successes, and relate to struggles. Teachers also can meet virtually with their coach within the platform.

Third, Sprocket is key to the vision of Knowledge in Action dissemination to teachers, schools, and districts nationwide and beyond. Teacher adaptations are intended to improve the curriculum and resources over time and to make them applicable to a wider audience, in terms of geographic region, student and teacher demographics, and student ability.

Overview of the KIA Professional Development Model

In partnership with LER, the PBLWorks—in operation since 1999 with school-, district-, and state-level clients nationwide and internationally—provides Knowledge in Action professional development to introduce teachers to the model and support their use of the curriculum.

Summer Institute

The Knowledge in Action Summer Institute takes place over four full days. The objective of the Summer Institute is to train teachers on the Knowledge in Action approach, expose them to the curriculum, and prepare them to teach the first unit. PBLWorks coaches guide teachers' learning about Knowledge in Action through the lens of their "Gold Standard PBL" approach, which models PBL as a critical part of teachers' learning about how to teach KIA and, more generally, PBL.

Professional development Planning Sessions

Throughout the school year, Knowledge in Action teachers have four full-day, PBLWorks-provided instructional classes. These Planning Sessions are strategically timed to align with the curriculum timeline so teachers can reflect on past projects and, with their coach and fellow teachers, prepare for the next of the five units (see Table A1). Each session also focuses on an inquiry topic related to PBL best practices.

Table A1: Alignment between Planning Sessions and improvement cycles

PD Component	Timing
Summer Institute	Plan Unit 1
Planning Session 1	Plan Unit 2
Planning Session 2	Plan Unit 3
Planning Session 3	Plan Unit 4
Planning Session 4	Plan Unit 5

Coaching

PBLWorks provides one-on-one coaching to participating Knowledge in Action teachers using both online and in-person formats. Following the "improvement cycle" model developed by Knight (2007), KIA coaching includes three steps:

1. Together, teacher and coach identify teaching practice areas needing improvement and create a "theory of action" plan to develop these areas
2. Collect data on teaching practice through in-person or video observations
3. Together, teacher and coach analyze the observation data in relation to the theory of action developed in the first step.

The improvement cycle is designed for implementation prior to each Professional Development (PD) session. PBLWorks coaches also provide support at teachers' requests via email, phone, or the Sprocket portal.

Community

LER's greater PBL community includes researchers, students, teachers, and school and district leaders, as well as foundation staff, researchers, and PD providers. Within this larger PBL network and the subsection specific to Knowledge in Action, LER aims to build clarity and consensus around best practices to facilitate teachers' creation and implementation of high-quality projects, improve the curriculum's academic rigor, encourage students' engagement with learning and teacher's engagement with teaching, and build an evidence-based approach for examining the effectiveness of project-based learning (Baines et al., 2015).

Encouraging teachers' inclusion in the PBL and Knowledge in Action support communities is a critical feature of the PD model. Knowledge in Action teachers have three main means through which to interact with the KIA community, each of which we reference above. First, through Sprocket, teachers from districts and schools nationwide can share best practices and support one another while coaches support teachers' evolving early-implementation needs. Second, the Summer Institute kicks off with community-building exercises extending through the week. Third, during the four PD Sessions, teachers reconvene with their peer groups and coaches, providing time for teachers to share their Knowledge in Action successes and challenges and to support each other at critical junctures. As more teachers join LER's KIA and larger PBL community in future years, LER aims to continue promoting community engagement within and between schools and districts, including engaging each generation of Knowledge in Action teachers' provision of mentorship and encouragement to the newcomers.

Stable Unit of Treatment Value Assumption

We made the Stable Unit of Treatment Value Assumption (SUTVA) about the KIA "treatment," as is standard in causal analyses. There are two parts to the SUTVA assumption. The first is that there is only one version of the treatment. The second part of SUTVA, known as "no interference," is that one teacher's group assignment does not affect the outcomes of another teacher's students. Both parts of SUTVA are relevant to all evaluations of educational interventions in which teachers might interact.

SUTVA, Part I: One version of treatment

SUTVA's one version of treatment assumption implies all teachers within the same comparison group (business-as-usual, one year of KIA implementation, or two years of KIA implementation) have access to the same version of their group's treatment. This assumption thus requires teachers in schools assigned to the control condition to receive in Year Two the same treatment that teachers in schools assigned to treatment schools did in Year One. In both years, the KIA treatment included access to

KIA curriculum and instructional resources, as well as an online teacher community, offered through Sprocket; the opportunity to participate in a four-day Summer Institute and four full-day Planning Institutes scheduled throughout the school year; and the opportunity to participate in virtual coaching (in the same quantity in both years). We expect this first part of SUTVA was met for experimental teachers within a treatment status and year. However, there may have been year-to-year contextual changes, and/or the KIA program designers may have implemented changes designed to improve to the curriculum and professional development supports over time.

In addition to requiring that control teachers received the same treatment in Year Two as treatment teachers did in Year One, the "one version of treatment" clause also requires that in Year One, control teachers and non-experimental teachers were both in the same "business-as-usual" condition, without control teachers partially engaging in KIA in anticipation of their future involvement ("anticipation bias"). In our Year Two non-experimental approach, changes in the quality of the KIA intervention over time and anticipation bias are not concerns.

SUTVA, Part II: No interference

In the context of Knowledge in Action, SUTVA's assumption of no interference requires that student outcomes for Teacher A are independent of Teacher B's treatment assignment and actions. A violation of SUTVA would occur if in Year One, treatment teachers shared KIA intervention resources with control teachers, such that those resources might affect the outcomes of control teachers' students. To limit sharing of KIA resources between treatment and control teachers, we employed school-level randomization in the Efficacy Study, so no-interference violations are unlikely given that study's units of randomization are schools and not teachers within schools.

Appendix B: Teacher and School Inclusion Criteria

Experimental teacher recruitment

Pre-determined KIA teacher eligibility requirements were as follows:

1. Minimum of one year of experience teaching APGOV or APES prior to the 2016-17 school year.
2. Scheduled to teach APGOV or APES during the 2016-17 school year in the same school they were teaching in during study recruitment in 2015-16 (i.e., their randomized school).
3. Agreement, if randomized to the treatment group, to attend the four-day 2016 KIA Summer Institute held in each district.
4. Signed the teacher enrollment consent form specifying responsibilities and benefits of participating in the KIA RCT prior to school randomization and the district Summer Institute.
5. Consent of the school principal to participate in the study.

Teachers consenting to take part in the Knowledge in Action Study were willing to participate in professional development, change their teaching practices, and participate in research activities.

Experimental school inclusion criteria

Schools' eligibility for randomization into the KIA Efficacy Study sample was contingent upon the teachers in those schools. A school was eligible if it was in a participating district and had an eligible teacher who signed the KIA consent enrollment form.

Appendix C: District Context

Particularly in an efficacy study—in which an intervention is tested under ideal conditions—understanding context matters. In this Appendix, we describe participating districts' contextually. We drew from interviews conducted during the 2016-17 year with district and school staff, public records, and administrative records to describe district AP policies, AP enrollment and exam participation, and influences on teachers' instruction.

Five large, geographically-distributed, and primarily urban districts participated in the KIA RCT. We designed the study to pool data and analysis across districts. As such, we do not provide any district-specific results. In alignment with our cross-district approach, below we describe contextual features applying across the five districts, such as their motivations to participate in the KIA Efficacy Study, and their AP policies, college and career readiness goals, social and emotional learning goals, curriculum and instruction guidelines, and AP and PBL professional development (as distinct from Knowledge in Action PD). This context may have changed between data collection during the 2016-17 school year and report publication in 2021.

Districts' Motivation to participate in Knowledge in Action

In fall 2016, we asked district staff about their district's motivation to participate in Knowledge in Action and their own beliefs about PBL. Though there was little existing PBL instruction in KIA districts at the secondary level, their responses indicated that the districts were receptive to KIA because they believed in the potential positive effects of PBL approaches on teaching and learning.

Districts' reasons for participating included interest in initiating and/or expanding PBL instructional capacity, learning the results of a PBL approach on student outcomes, alignment with district civic focus, and the attractiveness of the professional development and curriculum resources provided through participation.

At the time of the decision to enroll in the Efficacy Study, only one district that would be involved in KIA implementation had articulated a strong commitment to PBL instruction in its strategic plan—and this was the only district citing KIA's alignment with existing initiatives as a factor in choosing to participate. However, per a staff person in this district, "the groundswell for PBL has not been in high school;" rather, it was in elementary and middle school, with the primary exception of a high school that did not enroll to take part in the study. Another district was attracted originally by KIA's civic focus, which aligned with district civic education priorities.

Staff from three districts explained that a driving factor in the decision to participate was the rigorous evaluation of PBL in an AP context. If the results should prove positive, these districts hope to expand PBL in the district. One staff person explained that PBL can have a stigma associated with it, so they hoped the Efficacy Study could demonstrate positive effects on student outcomes, with such results potentially serving as "speaking points" to convince teachers and other school officials to "expand their efforts around it."

At least one representative in every district shared that a primary motivation to participate was access to the resources provided by Knowledge in Action, including eight PD days at no cost to the district

beyond substitute fees, and a cornucopia of PBL curriculum and instruction resources that had already been designed and used in classrooms. District staff were pleased to provide teachers with what they perceived to be valuable opportunities to improve their practice and, with that, student learning.

AP Policies

All five districts endeavored to increase equitable student access to AP courses—particularly among underrepresented student groups—while also increasing exam participation rates and scores. An additional goal in in two districts was for every student to experience at least one AP course prior to graduation. Strategies used by districts to increase equitable access to AP courses included open enrollment, the AP Potential prediction tool, recommended course sequencing, broad course offerings, AP preparation for teachers and students, policies to encourage examination-taking, and goals for the rate of qualifying scores. We discuss each in turn, below.

Open enrollment

To encourage equitable access, districts do not require prior achievement or teacher recommendation prerequisites for enrolling in AP courses. Though there is a movement in schools and districts nationwide to offer access to AP all students who want to enroll, open access is not universal. Four districts had no district prerequisites at all, such that students who sought advanced coursework could enroll in AP courses and expect to receive necessary academic support. In one district, course prerequisites sometimes limited access to AP courses. According to district staff, opening AP access to a larger pool of students allows schools to offer more AP courses.

AP Potential

The College Board has reported that many students who take the PSAT/SAT but do not enroll in AP courses might have succeeded in those courses (Thissen, 2007). This finding, based on an analysis of the relationship between PSAT/SAT scores and AP scores among a population of students who take both, led to the College Board's development of a tool—AP Potential—that school staff can use to forecast whether a student would earn a qualifying score on an AP exam based on a PSAT/SAT score. Interview data suggested three of the five districts were in the initial stages of exploring use of the tool. Staffers in one district and a school leader in a second reported their districts recommend schools use AP Potential to identify students who may be successful in AP courses and then to encourage those students to enroll. In a third district, staff reported using AP Potential to determine which courses to offer at which schools.

Recommended course sequencing

At least one district encouraged students to start with "easier" AP courses, such as Geography, Spanish for Spanish-speaking students, and—notably—Environmental Science before progressing to more difficult courses, such as U.S. History or Biology.

Broad course offerings

Districts, as a way to increase enrollment, attempted to offer a diversity of AP courses—for example, in computer science, English literature, math, sciences, and world languages. However, finding teachers qualified to teach classes such as computer science, for which there could be high student demand, made expansion of some courses challenging. Related, district policies breaking up larger

comprehensive schools into smaller schools can limit AP course offerings, as smaller schools often do not have the student population to fill numerous AP courses, nor the teachers with the necessary background.

AP teacher preparation and support

Strategies used by Knowledge in Action districts to ensure sufficient numbers of trained AP teachers included offering AP professional development to current and prospective AP teachers, providing school-level AP professional development grants, and during Common Core State Standard (CCSS) professional development, highlighting overlap between its objectives and AP's. One district interviewee described the practice of offering KIA professional development workshops on inquiry and project-based learning in the AP context as a strategy for "getting rid of the stigma that AP is designed for (students) who can 'sit and get.'"

AP student preparation

Districts collaborated with local universities to host events designed to support students' exam preparation and college preparedness. There were "AP Nights," "AP Exam Day" and, in one district, an event held monthly on Saturdays throughout the school year. One district operates a program seeking to identify promising future AP students when they are in middle school.

AP examination-taking

All five districts pay the AP and PSAT exam fees for their students who are eligible for free or reduced-price lunch. One district paid all students' AP exam fees regardless of need, while another district paid exam fees on a case-by-case basis for students who were not eligible for free or reduced-price lunch. District staff believed paying for PSAT exams helps teachers and students, because when students take PSATs, schools then can use AP Potential Reports to identify students who would be good candidates for AP course enrollment.

In one district, prospective AP students and their parents signed contracts at the beginning of an AP course specifying the student's commitment to take the exam. Other districts communicated the expectation in other ways; for example, one exempts students from taking the relevant final course exam if they take the relevant AP exam, and another provides only the "one-point bump" to AP course grade-point averages for students who take the relevant AP exam. Two districts did not expressly communicate the AP exam taking expectation so concretely.

In all five districts, students enrolled to take the AP exams in spring 2017.

Qualifying score-rate goals

Districts' goals for their rates of qualifying scores varied, and included overall qualifying score rates, (for example, a qualifying score rate of 50%) and growth (e.g., increasing the percentage of qualifying scores by 2 percentage points over the prior school year).

Influences on classroom practice

Across the Knowledge in Action districts, we heard from teachers, school leaders, and district staff who aspired to develop students' preparedness for college, career, and citizenship. Influences on classroom practice included district-level goals for student learning in the areas of college and career

readiness, social and emotional learning, and civic engagement, as well as curriculum and instruction resources, and PD supports.

College and Career Readiness goals

All but one of the states in which KIA districts are located had adopted the Common Core State Standards as of the 2016-17 school year. The CCSS defines "the knowledge and skills students should achieve in order to graduate from high school ready to succeed in entry-level, credit-bearing academic college courses and in workforce training programs."

Each KIA district's strategic plan specifies deeper-learning objectives as part of its approach for aligning instruction with the relevant state standards. Though the districts vary in specific skills emphasized, they have the following in common: the "Four Cs" (critical thinking, communication, collaboration, creativity); technology and information literacy; ethical and global citizenship; life-long learning; subject area skills; and vocational and technical skills. The strategic plans also describe goals for closing the college and career "achievement gap" for students from low-income families and for students of color (e.g., advanced studies diplomas, industry credentials, two- and four-year college enrollment).

Social and Emotional Learning goals

Overlapping with deeper learning and CCSS objectives, all the KIA districts specified goals for students' social and emotional learning, as well as laying out strategies for addressing those goals. For example, one district identified a student learning objective as being "goal-directed and resilient," and called for "executive function curriculum" expansion from early childhood programs through high school. Another identified as one of its five focus areas students' social and emotional health, which it was addressing through an anti-bullying campaign, embedding social-emotional learning within the core curriculum, and supporting character-building activities. Another district focused on development of students' self-management, decision-making, and interpersonal relationship skills as necessary to meet CCSS academic expectations. Two districts promoted disciplinary alternatives to suspensions and expulsions, with one using anger-management trainings for teachers. One district described effective learning environments across all grade levels as those meeting students' academic, physical and social-emotional needs—though focused on social and emotional curriculum and development benchmarks in elementary grades.

Curriculum and Instruction

In all five districts, schools had considerable autonomy over curriculum and instruction, with the districts playing a supporting role. District-provided supports included planning and pacing guides aligned with district learning objectives and state standards, subject-area specialists, and textbook recommendations, including solicitation of input from teachers and school leaders. For AP courses, including Knowledge in Action AP courses, the College Board has primary oversight for course content, which it exercises through the requirement that schools submit a subject-specific AP Audit form and course syllabus for all courses designated as AP.

AP and PBL Professional Development

Another district support role was funding and supporting implementation of professional development. Several districts provided additional AP support above and beyond the College Board's annual 30-plus hour summer institutes. For example, one district provided professional development each August for all its teachers of AP, International Baccalaureate, honors, and dual enrollment courses. In another district, teachers could attend monthly trainings on Saturdays through a university partnership program, receiving compensation for their time, while another offered an AP Institute day each fall. In 2016-17, one district offered approximately 20 hours of PD for teachers of environmental science, not specific to AP, and encouraged APES teachers to attend.

With the exception of the PBLWorks professional development provided as part of participation in the Efficacy Study, none of the Knowledge in Action districts offer PBL professional development specific to AP courses. Two districts offered PBL professional development for high school teachers. One district, in partnership with local colleges and universities, offered a one-credit course on PBL. Another offered PBL professional development based on the Understanding by Design framework open to all high school teachers.

Appendix D: Impact Analysis Data Sources

Administrative data sources informing our impact analyses in Years One and Two included records from the school districts and College Board, KIA professional development participation data, and public databases. In Year One, we also collected and used as outcomes College and Work Readiness (CWRA+) and student survey field data.

Administrative district records

Through the execution of data-sharing agreements, each of the five districts provided the research team with de-identified records for all 2015-16 through 2017-18 APGOV and APES students in Knowledge in Action schools, identifying records only for students with consent to share them.[1] We also requested 2014-15 records, though provision of this data was inconsistent across districts with considerable systemic missing data patterns. Variables included prior academic achievement (i.e., eighth-grade Math, ELA, and Science scores, as well as PSAT, SAT and ACT scores), demographics, and AP examination scores.

College Board records

Through a data-sharing agreement, the College Board provided the research team with de-identified records,[2] defined with the district-provided Knowledge in Action study identification, of students from four of the five participating districts. One district did not allow the College Board to share data with the research team. The variables provided by the College Board included all PSAT, NMSQT, SAT, and AP scores on record. The College Board also provided main-form[3] May 2016, 2017, and 2018 APGOV and/or APES examination variables including:

- Scored responses on all items, multiple-choice, and open-ended items
- Total raw (sum of all multiple-choice items) and weighted scores the student received across the multiple-choice items
- Total raw (sum of all open-ended items) and weighted scores the student received across the open-ended items
- Overall composite scores (sum of weighted multiple-choice and open-ended items)
- Ordinal 1-5 AP scores

We relied on composite scores, as well as weighted multiple-choice and free-response section scores, as alternative outcome measures assessing the potential impact KIA may have on student knowledge

[1] Identified records were applicable only to Year One analyses of student survey and College and Work Readiness Assessment outcomes.
[2] The USC research team and College Board developed a matching process by which researchers could access only de-identified College Board data.
[3] AP exams contain multiple forms (usually three) in a single administration. A single main form is administered on the designated date for a particular AP subject test. In rare cases when students could not take the exam on that date, they were offered chances to make up, during which alternative forms are offered. According to the AP program, the vast majority (>95%) of students take the main form every year.

and skills. These "fine-grained" AP scores are the foundation of the College Board's 1-5 overall AP examination scores, and therefore contain finer gradient of one's academic achievement. The fine-grained section scores also permit investigation of whether students of treatment teachers performed better on the writing or multiple-choice sections of the AP exam versus control teachers' students. In sum, fine-grained versions of students' AP scores permitted investigation of whether treatment students demonstrated higher performance than control students, though possibly not high enough to pass the threshold into the qualifying-score level on the coarse metric of 1-5.

Source of AP, PSAT, and SAT covariate and outcome data when records were available from the College Board and districts

Though we requested AP, PSAT, and SAT scores, used as outcomes and covariates, from both districts and the College Board, data fulfillment varied slightly across districts. The College Board was the sole source of fine-grained AP scores for 2016-17 and 2017-18 APGOV and APES student cohorts. For AP scores, out of the five districts, three provided ordinal AP scores and gave permission for the College Board to provide the research team with fine-grained AP scores. One district provided only AP qualifying-score outcomes (exam-takers only), but did grant permission for the College Board to share fine-grained and ordinal AP scores. Finally, one district did not permit the College Board to share data, but did provide the research team with ordinal AP scores. Because the College Board was the sole source for fine-grained outcome data, for this last district we did not have access to fine-grained composite or weighted section scores.

Regarding PSAT and SAT scores, three districts provided data and allowed the College Board to share data with the research team. One district provided neither PSAT nor SAT scores but did grant permission to the College Board, and a fifth district provided PSAT and SAT data but no College Board data-sharing permission. ACT data came solely from the districts, provided by two out of the five. As detailed in Appendix E, we relied on PSAT data if available. Of all students enrolled in APGOV and APES courses in 2015-16, 2016-17 and 2017-18 across the five districts, 92% took at least one of the three tests. We used SAT and ACT data only if a student had no PSAT data available, applicable to 2% and 6% of the sample respectively.

When the same data was available from both the College Board and district, we used the College Board data.

The link rates for the full 2017-18 cohort of APGOV and APES students and non-experimental 2016-17 students (i.e., without any sample exclusions) were 94% and 96%, respectively, across grades 9-12, for the four districts granting us permission to obtain College Board data. This means that the College Board attempted to link 94% of the de-identified student records for our sample. In Appendix E, we describe our approaches to standardization of measures of students' prior achievement, including AP, PSAT, SAT, ACT and eighth-grade scores.

Professional development and coaching participation data

PBLWorks provided the research team with professional development participation data, indicating which teachers participated in any Knowledge in Action professional development activities during the 2016-17 and 2017-18 school year (i.e., Summer Institute, the four PD sessions held throughout

the school year, and/or coaching). We used this participation data, in conjunction with Sprocket log-in data provided by LER, to create flags defining which teachers complied with their treatment status. Impact analyses included complier and non-complier teachers, while implementation analyses included only complier teachers.

Public databases

We collected school-level data from the National Center for Education Statistics (NCES), and state and district websites. We used 2015-16 variables so that all covariates were measured prior to the first KIA implementation year (i.e., 2016-17).

National Center for Education Statistics

The National Center for Education Statistics (NCES) is the federal entity responsible for collection and analysis of education data. The NCES public school database is part of the Common Core of Data (CCD), which is updated annually based on new information collected from state education agencies. The data includes directory information (e.g., address, phone number, principal name); school status (e.g., charter school or Title I); enrollment by grade, ethnicity, and gender; free and reduced-price lunch (FRL) counts; teacher counts; and student-teacher ratios. We used NCES data describing school-level characteristics as of the baseline year, 2015-16. For one district, missing from the 2015-16 file was FRL data, so we used data from 2014-15.

District and state data sources

Unavailable from the NCES was data on school-level proportions of English-language learners and AP performance. We collected these variables from district and state websites.

College and Work Readiness Assessment

The Council for Aid to Education (CAE) developed, administers, and scores the CWRA+[4], a 90-minute online assessment for high school students evaluating their ability to "access, structure and use information" (CAE, 2015). We chose the CWRA+ because it was the only available deeper-learning assessment that, as of the 2015-16 academic year (when we applied for approval from districts' research review boards),: 1) had demonstrated reliability; 2) measured deeper-learning skills without assuming specific content knowledge, and was thus appropriate for APES and APGOV; 3) had a well-functioning online platform and support infrastructure necessary for large-scale administration; and 4) had a reasonable cost of $35 per student.

The CWRA+ is composed of two sections. The performance task (PT) measures students' skills in analysis, problem-solving, and writing mechanics, as well as their writing effectiveness. To optimize the CWRA+ face validity, the research team selected one form of the performance task—featuring content generally familiar to APGOV students—to be administered to APGOV students, and another form similarly selected to be administered to APES students. In the second section, selected response (SR) questions measure students' skills in analysis and problem-solving, including their ability to reason

[4] CWRA+ overview: https://cae.org/solutions/

scientifically, read and evaluate critically, and critique an argument. The SR section consists of a randomly-selected set of 25 questions (not course-specific and varied across students).

Students have 60 minutes to finish the PT and 30 minutes to finish the SR. When an uninterrupted 90-minute block is unavailable, CAE offers a "pause" button after completion of the PT so the exam may be taken over two class periods. CAE's stated rule is that use of the pause button must take place after completion of the PT, resuming with the SR in a subsequent class meeting (within seven days).

CAE scorers rate students' performance on four dimensions: analytical reasoning and evaluation, writing effectiveness, writing mechanics, and problem solving. CAE assigns three scores for each student: a PT score, an SR score, and an overall score determined by an equally-weighted combination of the two section scores. Because the overall score is an unweighted sum of the section scores, a student has an overall score only with valid scores on both sections; otherwise, the overall score is missing. After the PT and SR, students are asked to answer questions about their background characteristics, and levels of engagement and effort during the assessment.

CAE has tested the reliability and validity of CWRA+ scores for more than a decade. School-level internal consistency is greater than 0.90, the total score student-level reliability coefficient is 0.84, the performance task reliability coefficient is 0.78, and selected response question reliability coefficient is 0.76.[5]

CWRA+ outcome variable processing

CAE provided the USC CESR research team with PT scores, SR scores, time spent on each section, and self-reported data on students' engagement and motivation. According to CAE, PT responses were not scored if they completely missed the mark or the answer was irrelevant to the question. SR responses were not scored if a respondent did not answer at least 13 questions out of the 25. The composite score was an unweighted average of the section scores, obtained only if a student received valid scores in both sections.

We examined the time-spent data to understand to what extent the CWRA+ administration was implemented as designed, as well as to flag aberrant responders. Impact analysis results could be "contaminated" if scores were influenced by factors other than students' deeper-learning skills (i.e., the constructs the CWRA+ assesses). While the time spent on the PT section followed a bell-curved shape, as expected, the time spent on the SR section followed a bimodal shape with an early peak at three minutes. Further investigation on the cases spending three minutes or less on the SR section revealed these students rushed through the SR section, averaging less than 10 seconds per question. They also have lower self-reported engagement and motivation than the rest of the sample. Therefore, we decided to define "valid" SR scores as those with a minimum of three minutes spent. We subsequently recoded 107 SR scores as missing and, consequently, recoded 63 total scores as missing. (Of the 107, 44 did not have a valid PT score either, and therefore already did not have an overall score)

---

[5] Technical CWRA+ FAQ document:
http://cae.org/images/uploads/pdf/CWRA+_Plus_Technical_FAQs.pdf

Student Survey

We developed a student survey to measure students' intra- and inter-personal skills, and civic engagement. As Table D1 shows, the survey included items measuring students' attitudes towards learning, adapted from the Consortium on Chicago School Research's "Becoming Effective Learner's" student survey (Farrington et al., 2014); on "collaboration" from the AIR Deeper Learning survey (American Institutes for Research, 2016); and on "leadership" from the "Yes 2.0" survey (Hansen & Larson, 2005cite). We also adapted items on civic skills, attitudes, engagement, and intentions from the California Civic Index and other instruments compiled in Flanagan, Syvertsen & Stout (2007).

Table D1: Intra- and inter-personal, and civic engagement constructs measured in the student survey

Domain	Construct	Meaning	Sources
Intrapersonal skills	Self-efficacy	Belief in one's capacity to behave in a way that results in attainment of specific performances	Bandura, 1977
	Growth mindset	Belief that intelligence and abilities can be developed through dedication and hard work	Dweck, 2000
	Grit	Perseverance and passion for long-term goals	Duckworth, Peterson, Matthews, & Kelly, 2007
Interpersonal skills	Collaboration	"A coordinated, synchronous activity that is a result of a continued attempt to construct and maintain a shared conception of a problem"	Roschelle & Teasley, 1995, 70
	Opportunities for leadership	"Process whereby an individual influences a group of individuals to achieve a common goal."	Northouse, 2013, 6
Civic engagement	Appreciation of diversity	Demonstration of tolerance and interest in engaging with culturally, ethnically, religiously and by gender diverse individuals and groups	Oswald, 2004
	Political efficacy	Belief that a person can affect community or political change, often the impetus for engagement	Flanagan et al., 2007
	Participatory citizenship	Belief that citizens must actively participate and lead within established political and community systems	Westheimer & Kahne, 2004
	Political interest	Interest in politics	Flanagan et al., 2007
	Communication with friends about politics	Measurement of adolescents' discussion of politics with friends serves as a proxy for interest in politics	Flanagan et al., 2007
	Concern for the environment	Belief about the value of protecting the environment	Flanagan et al., 2007

Construction of survey composite measures

We conducted psychometric analyses on data from the student survey, and surveys we administered to teachers prior to and after their participation in the KIA intervention. We refined constructs and computed composite scores with the following steps:

1. The initial set of items forming a construct were either derived from existing measures (e.g., self-efficacy, grit, growth mindset) or from our theory (e.g., quality of groupwork).
2. Using exploratory factor analysis, we examined whether substantively-grouped items formed a single factor. Determining the number of factors were a combination of statistical evidence (e.g., a scree plot showing the location of an "elbow," eigenvalues greater than 1) and structure interpretability (i.e., whether it is conceptually meaningful for an item to load on a specific factor). When the single factor appeared weak, or when we detected multiple factors, we continued with the third step to refine the scale; otherwise, we continued to Step 4.
3. If necessary, we conducted item analysis to examine the performance of individual items and their contributions to the scale. When an item did not perform well in a scale, we considered dropping it from the scale in two scenarios: 1) when an item performed poorly (e.g., factor loading was below 0.4, or dropping greatly improved the reliability), and/or the item content seemed to be particularly prone to self-report bias or was less central to the KIA theory of action; and 2) when an item was alike in content to others in the same scale, so discarding it did not significantly reduce the scale's reliability. When a common set of items had multiple factors, we determined theoretically whether it made sense to split it into multiple subscales. We iterated between steps 2 and 3 until achieving a satisfactory factor.
4. Once we realized a common factor structure consisting of three or more items, we then calculated Cronbach's alpha as a measure of internal consistency. This index often is considered as a lower bound of reliability (McDonald, 1999) and, therefore, a satisfactory value of alpha implies a satisfactory reliability.
5. We calculated the composite scale scores as the unweighted average over a given set's item scores. For the cases in which respondents missed partial items in the same set (the amount of missingness never exceeded 30%), we calculated the scale score as the average of the non-missing item.

Following the steps described above, we retained 19 student survey scales, shown in Table D2.

Table D2: Student survey constructs, number of items, and reliability

	Number of items	Alpha
Interest in politics	3	0.89
Grit	5	0.89
Self-efficacy	4	0.87
Collaboration	10	0.87
Growth mindset	5	0.82
Participatory citizenship	4	0.82
Concern for the environment	4	0.82
Civic/political efficacy	4	0.81
Appreciation of diversity	3	0.77
Opportunities for leadership	3	0.76
Course relevance for the future	5	0.89

Course satisfaction	5	0.88
Student engagement with learning	3	0.8
Quality of groupwork	6	0.85
Quality of classroom discussion	5	0.84
School environment	4	0.81
Opportunity for work with real world relevance	4	0.8
Teacher promotes student agency	3	0.79
Classroom environment	3	0.79

Appendix E: Transformation of Achievement Variables

This Appendix details standardization of student prior achievement variables, the performances on APES and APGOV exams of teachers' classes in the 2015-16 school year, and AP outcome variables across the five participating school districts.

Transformation of eighth-grade achievement variables

We used students' eighth-grade state standardized test scores in Science, Math, and English Language Arts (ELA) as covariates in our experimental and non-experimental arm analyses. Because the five districts are in separate states, each state score was scaled on a different metric, and state assessments were developed to address varying learning standards—such that some were more difficult than others—implying the scores were not directly comparable across different tests.

To put all scores on a common metric, we rescaled the state assessment scores onto the scale of National Assessment of Educational Progress (NAEP). First, we standardized each student's eighth-grade achievement score by the statewide mean and standard deviation (SD) of that particular test administration per test subject, year, and state, obtained from public releases from each state. For example, sample students from District B who were in eighth grade in 2015 all had their 2015 state scores in Math standardized using the state average and standard deviation in 2015. This transformed z score represented each student's relative standing within each test administration.

The second step was to scale up each z score onto the NAEP metric. Specifically, we obtained the means and standard deviations on NAEP scores per test subject, year and state. The summary statistics used were based on the entire NAEP sample, excluding students with disabilities and English-language learners. This was most analogous to the population in which the state eighth-grade achievement data were obtained. Following the District B example above, we obtained state average and SD on the 2015 administration of the NAEP Math, then multiplied the z score from Step 1 by the NAEP SD and added with NAEP mean. This two-step approach rescaled each state assessment score onto the NAEP metric to account for differences between states and across years in students' eighth-grade standardized test scores. Reardon, Kalogrides & Ho (2017) similarly used NAEP scores to realign state assessment distributions onto a common metric.

Two assumptions justified this approach:

1) At the population level, both state and NAEP assessments reflected students' overall performance in the same degree. The difference on mean and SD of state versus NAEP assessments was due primarily to scaling difference.
2) Multiple studies linking state and NAEP assessments suggest the correlation between the two types of assessments is around 0.75 at the student level on the same exam-taking sample (see a review by Thissen, 2007). This correlation is sufficiently high, and similar to typical test-retest reliability.

We acknowledge the content measured in state and NAEP assessments was not always completely aligned, and that the reliability of the state and NAEP assessments were not identical. As information

about state assessment content alignment and reliability was not consistently available across all districts, subjects, and years, we did not account for such factors in our processing.

After re-scaling, all cases with valid assessment data had a rescaled score. We imputed missing scores as described in Appendix J.

Both steps of our approach had their difficulties. To operationalize the first step of standardizing each student's test score by the statewide mean and SD of that particular test administration per subject, year, and state, we referred to technical manuals and state report cards. The manuals usually contained test administration means and SD, and were available for most tests in most years. State report cards contained only means, though they were available for all tests across all years. Reported means did not always correspond between technical manuals and state report cards because sometimes they were informed by slightly different samples. When both the technical report and state report card were available, we investigated differences in reported means, determining all differences were ignorable. Because SD was available only through the technical manual, we used the mean from the technical manual if available; otherwise, we used the report card mean as a proxy. In rare cases when the SD was not available, we used adjacent years' information as proxy as the SDs were usually quite stable across adjacent years.

In a few cases, the technical manual reported only summary statistics on raw scores (rather than scale scores). In these cases, we leveraged the available concordance tables between raw and scaled score, which mapped each possible raw score to a corresponding scale score, and converted the raw-score summary statistics to scale-score using simulation. Specifically, we simulated 100,000 cases using a normal distribution of the reported raw-score mean and SD, converted each simulated raw score to scale score using the raw-to-scale score conversion published in the technical manual, and calculated the mean and SD of the converted scores.

There were two challenges to scaling up each student's z score using each state's NAEP mean and standard deviation in the given test year. First, the NAEP was only administered in 2013 and 2015 for Math and ELA, and in 2011 and 2015 for science; therefore, the state summary statistics were available only in these years. To fill in the missing years, we interpolated and extrapolated the means based on available years, using the average SD. We investigated the across-year difference on the available-year's mean and SD, finding differences negligible. Second, in one district, the state science test administration was for seventh-graders rather than eighth-graders, and NAEP is administered only to fourth, eighth, and 12th-graders. In this case, we used as proxy the eighth-grade NAEP science mean and SD.

Standardization of high-school national prior achievement variables

Key student-level covariates in our baseline-equivalence analysis and impact models included high school national prior achievement variables; namely, scores on the PSAT, SAT, and ACT subsections of Math and ELA. We obtained data from both the district and the College Board on College Board-published tests, including PSAT and SAT, with detailed data availability and sources documented in Appendix D.

Because these three tests were scored on different metrics, and the PSAT and SAT went through major changes in the years in which our sample took the tests, resulting the old and new versions of the PSAT and SAT, we rescaled these five tests onto a common metric. Specifically, first we realigned relevant subjects and combined into the ELA score as shown in Table E1, whereas each test had a single Math section and, therefore, was naturally aligned.

Table E1: English Language Arts sections across the ACT, and old and new PSAT and SAT forms

Test	English Language Arts sections
Old PSAT	Reading + writing*
New PSAT	Evidence-based reading and writing
Old SAT	Reading + writing*
New SAT	Evidence-based reading and writing
ACT	Reading + English*

*=aggregate into a single section

As scores on these five tests were not directly comparable, and no section-to-section concordance was available for some tests, we standardized the test scores against their national norms to put them on a common metric. Specifically, for each section of each test, we computed z scores by subtracting the national mean from a student's score and dividing it by the national standard deviation.[6] We then constructed the ELA composite scores by averaging the relevant sections' z scores.

As our goal was to have one Math and one ELA covariate for each of our base sample students, when a student had taken more than one test among the PSAT, SAT or ACT, we used the PSAT scores because most of our sample took the PSAT. If unavailable, we used the SAT (if administered prior to the intervention year) and otherwise used the ACT.

Aggregation of Teacher's Class Average for 2015-16 cohort

One of our covariates is teachers' classroom-average on APGOV or APES scores for the classes taught in 2015-16. These averages were calculated based on all students in the teacher's classroom for whom we had available data, regardless whether the student was in our analytic sample or not. The student-level APGOV and APES scores were obtained and processed using the same approach as documented in Appendices D and E. Classroom averages were computed matching each of our outcome type; for instance, percent of the class obtaining AP credit for AP credit-or-not analysis, and average fine-grained score for AP fine-grained analysis.

Standardization of fine-grained AP outcome variables

As the APES and APGOV fine-grained scores were on different metrics, to combine the scores for our pooled impact analysis, we standardized the fine-grained scores against the national norms. We standardized/rescaled only the weighted section scores and composite scores to the same metric, as norms were not available for unweighted section scores.

[6] The old PSAT's national norm was not available, so we first concorded the old PSAT to the new PSAT score scale, then standardized against new PSAT's national norm. We used this College Board concordance table: https://collegereadiness.collegeboard.org/pdf/2015-psat-nmsqt-concordance-tables.pdf

Appendix F: School- and Teacher-Level Attrition Overall, and by District and Course

Overall

Table F1: School- and teacher-level sample loss without consideration for missing outcome data

	Schools			Teachers		
	Total	Treatment	Control	Total	Treatment	Control
Year One						
Randomized	74	37	37	86	44	42
Attrited 2016-17	6 (8%)	6 (16%)	0 (0%)	12 (14%)	9 (20%)	3 (7%)
Schools and teachers in Year One sample	68	31	37	74	35	39
Year Two						
Randomized	74	37	37	86	44	42
Attrited 2016-17 and 2017-18	24 (32%)	16 (43%)	8 (22%)	33 (38%)	21 (48%)	12 (29%)
Schools and teachers in Year Two sample	50	21	29	53	23	30

By District

Table F2: Cross-course school- and teacher-level attrition by district between randomization, Year One and Year Two

District A (APGOV and APES)

	Schools			Teachers		
	Total	T	C	Total	T	C
Randomized	11	6	5	13	7	6
Losses between randomization and Year One	1	1	0	2	1	1
Year One sample	10	5	5	11	6	5
Losses between Year One and Year Two	7	4	3	8	5	3
Year Two sample	3	1	2	3	1	2

District B (APGOV and APES)

	Schools			Teachers		
	Total	T	C	Total	T	C
Randomized	6	3	3	9	4	4
Losses between randomization and Year One	0	0	0	1	0	0
Year One sample	6	3	3	8	4	4
Losses between Year One and Year Two	1	1	0	1	1	0

Year Two sample	5	2	3	7	3	4

District C (APGOV and APES)

	Schools			Teachers		
	Total	T	C	Total	T	C
Randomized	12	6	6	16	8	8
Losses between randomization and Year One	1	1	0	5	3	2
Year One sample	11	5	6	11	5	6
Losses between Year One and Year Two	3	1	2	3	1	2
Year Two sample	8	4	4	8	4	4

District D (APGOV only)

	Schools			Teachers		
	Total	T	C	Total	T	C
Randomized	12	6	6	15	8	7
Losses between randomization and Year One	1	1	0	1	1	0
Year One sample	11	5	6	14	7	7
Losses between Year One and Year Two	0	0	0	2	1	1
Year Two sample	11	5	6	12	6	6

District E (APES only)

	Schools			Teachers		
	Total	T	C	Total	T	C
Randomized	33	16	17	34	17	17
Losses between randomization and Year One	3	3	0	4	4	0
Year One sample	30	13	17	30	13	17
Losses between Year One and Year Two	7	4	3	7	4	3
Year Two sample	23	9	14	23	9	14

Historical course offerings

Districts provided administrative data describing whether each school offered either or both courses prior to randomization at the end of the 2015-16 school year.[7] Tables F3 and F4 show the counts and percentage of schools, across the districts, that always offered APES and APGOV versus those not always offering those classes over the past three years.

[7] Of our five participating districts, three (Districts A, C, and E) provided historical course offerings in the three academic years prior to KIA implementation (2013-14 to 2015-16), whereas District B provided information for two academic years prior (2014-15 and 2015-16), and District D provided only for 2015-16. Given data limitations, our measures of average and prior-year historical course offerings are identical in District D.

Table F3: Counts and percentages, by district, of schools always offering APES

	District A	District B	District C	District E
Not always	12 (67%)	20 (69%)	50 (82%)	161 (83%)
Always	6 (33%)	9 (45%)	11 (18%)	32 (17%)
Total	18 (100%)	29 (100%)	61 (100%	193 (100%)

Table F4: Counts and percentages, by district, of schools always offering APGOV

	District A	District B	District C	District D
Not always	12 (67%)	20 (69%)	50 (82%)	5 (19%)
Always	6 (33%)	9 (45%)	11 (18%)	21 (81%)
Total	18 (100%)	29 (100%)	61 (100%	26 (100%)

School-level course composition

At the time of randomization, 32 of 74 schools were participating with an APGOV teacher while 48 were participating with an APES teacher, with equal counts of treatment and control schools within course[8]. Within the Year One subsample of 68 schools, 29 had a consented APGOV teacher and 42 had a consented APES teacher. As we show in Table F5—which shows school counts and, in parentheses, attrition from the randomized sample—attrition to the Year One sample was a bit greater among APES (12.5% across treatment status) relative to APGOV (9.4% across treatment status). However, although attrition from the Year Two subsample was twice as high among treatment (43.2%) compared to control (21.6%), it was parallel between courses, 34.4% APGOV and 35.4% APES.

Table F5: Course-specific school-level attrition

	Randomization			Year One			Year Two		
	Total	T	C	Total	T	C	Total	T	C
APGOV	32	16	16	29	15	14	21	9	12
	(0.0%)	(0.0%)	(0.0%)	(9.4%)	(6.3%)	(12.5%)	(34.4%)	(43.8%)	(25.0%)
APES	48	24	24	42	18	24	31	13	18
	(0.0%)	(0.0%)	(0.0%)	(12.5%)	(25.0%)	(0.0%)	(35.4%)	(45.8%)	(25.0%)
Total	74	37	37	68	31	37	50	21	29
	(0.0%)	(0.0%)	(0.0%)	(8.1%)	(16.2%)	(0.0%)	(32.4%)	(43.2%)	(21.6%)

[8] Six schools consented teachers of both courses at randomization: three in the Year One school subsample and two in the Year Two school subsample. Thus, school-course counts do not equal total school counts.

Appendix G: School- and Teacher-Level Baseline Equivalence Across Randomized, Year One, and Year Two School Subsamples

To examine respective differences between treatment and control schools at randomization and in the subsequent Year One and Year Two school subsamples, we first calculated means on school-level covariates using a series of district fixed-effects regression models, one per covariate. We estimated our school-level regression model using school-level data, with one observation per school. Using regression-adjusted means allows us to present within-district differences between treatment conditions on each school variable of interest. For all school-level covariates, we fit fixed effects models as follows:

$$X_s = \beta_0 + \beta_1 T_s + \mu_d + \epsilon$$

where X_s is the school-level covariate, T_s is the school's treatment status, and μ_d is the district fixed effect. Additionally, all regression models are probability weighted.

We obtained adjusted means for treatment and control groups through a STATA postestimation command (margins) that predicts the estimated value on each covariate for the treatment and control conditions.

To obtain standardized mean differences—following What Works Clearinghouse (WWC) version 4.1 standards as closely as possible, though deviating given we are calculating school- rather than student-level differences—we first calculated a pooled-standard deviation, s_p, defined as the standard deviation on each covariate, as well as small sample size adjustment, ω, as follows:

$$\omega = 1 - \frac{3}{4(n_1 + n_2) - 9}$$

where n_1 and n_2 are the sample sizes of the control and treatment groups, respectively. We do not include district fixed effects because β_1, including district FE, comes from the model specified above. We calculate the standardized mean difference (SMD) as:

$$SMD = \omega \frac{\beta_1}{s_p}$$

In Table G1, we present school-level post-estimation means and SMDs.

Table G1: School-level baseline (2015-16) characteristics, means (SDs), and standardized mean differences between treatment and control schools at randomization (n=74), Year One (2016-17, n=68), and Year Two (2017-18, n=50)

	Randomized Sample			Year One Sample			Year Two Sample		
Variable	T	C	SMD	T	C	SMD	T	C	SMD
% FRPL	0.62	0.66	-0.16	0.61	0.66	-0.2	0.55	0.64	-0.31
Magnet	0.35	0.29	0.13	0.39	0.29	0.22	0.49	0.30	0.36
Enrollment	1594.25	1628.56	-0.04	1671.96	1631.41	0.04	2009.62	1830.47	0.19
Title 1	0.75	0.72	0.06	0.73	0.72	0.04	0.65	0.66	-0.01
% LEP	0.13	0.14	-0.13	0.13	0.14	-0.15	0.11	0.13	-0.29
Student-teacher ratio	19.07	18.88	0.04	18.96	18.87	0.02	19.49	19.66	-0.03
Urban	0.75	0.74	0.03	0.74	0.73	0.00	0.74	0.69	0.1
Charter	0.05	0.03	0.13	0.03	0.03	0.02	0.05	0.03	0.12
% school taking AP exam	0.21	0.21	0.01	0.20	0.20	-0.07	0.22	0.22	0
Schools (n)	37	37	74	31	37	68	21	29	50

Our teacher-level analysis had several key differences from our school-level analysis. For this, we relied on student-level data, which effectively weighted each teacher-level variable by the number of students associated with that teacher in our data. For the Years One and Two samples, we included all students from each qualifying score (full sample) analytic sample, respectively. For the randomized sample, we included data from all students of randomized teachers for whom we had student records (83 out of 86 randomized teachers). For most teachers in the randomized sample, we used 2016-17 students, corresponding to the Year One sample. There was one teacher for whom we did not have records of 2016-17 students, so here we used 2015-16 students.

For our teacher analysis, we calculated means on teacher-level covariates from the baseline year (2015-16) using the same fixed effects models outlined for the school-level analysis:

$$X_s = \beta_0 + \beta_1 T_t + \mu_d + \epsilon$$

where X_t is the teacher-level covariate, T_s is the school's treatment status, and μ_d is the district fixed effect. All regression models are probability weighted.

The procedures for calculating adjusted means and standard mean differences are the same as described for the school-level analysis.

In Table G2, we present teacher-level post-estimation means and SMDs.

Table G2: Teacher-level baseline (2015-16) characteristics, means (SDs), and standardized mean differences between treatment and control teachers at randomization (n=83), Year One (2016-17, n=74), and Year Two (2017-18, n=53)

	Randomized			Year One			Year Two		
Variable	T	C	SMD	T	C	SMD	T	C	SMD
Average class size in 2015-16	29.01	28.04	0.1	28.32	27.97	0.04	30.29	28.07	0.3

Baseline (2015-16) % earning qualifying score APES/APGOV exam (all students)	0.28	0.31	-0.08	0.28	0.31	-0.08	0.31	0.37	-0.18
Baseline (2015-16) % taking APES/APGOV exam (all students)	0.84	0.75	0.21	0.84	0.75	0.21	0.91	0.86	0.3
Baseline (2015-16) average APES/APGOV score (exam-takers)	2.01	2.10	-0.08	2.02	2.09	-0.07	2.15	2.26	-0.12
Students (n)	1845	2190	4035	1499	2146	3645	1186	1760	2946
Teachers (n)	43	40	83	35	39	74	23	30	53

Appendix H: Descriptive Statistics for Experimental Teachers in 2015-16 (baseline), Across and by Course, at Randomization, Year One, and Year Two

In Tables H1-H3, we show mean statistics describing teachers' and their students' baseline (2015-16) characteristics overall, and by treatment and control at randomization (n=86), Year One (2016-17, n=74), and Year Two (2017-18, n=53). The first table presents means across courses, the second for APES only, and the third for APGOV only. In all three tables, the first column shows means describing the randomized sample, for which we do not have imputed data due to not having data at all on four teachers who were randomized but left the study prior to Year One. The next four columns show means based on unimputed data, followed by imputed data, for the Year One and Year Two teacher subsamples, respectively. Presenting unimputed and imputed data side by side demonstrate the effectiveness of our imputation. The final three columns describe data missingness among consented teachers within randomized schools, and in the Year One and Year Two subsamples.

Table H1: Teacher-level baseline (2015-16) characteristics overall, and by treatment and control at randomization (n=86), Year One (2016-17, n=74), and Year Two (2017-18, n=53)

Variable	Randomized (2016-17)	Year One (2016-17)	Year One (2016-17) including imputed	Year Two (2017-18)	Year Two (2017-18) including imputed	Randomized missing data (not imputed)	Imputed N (2016-17)	Imputed N (2017-18)
Teacher female								
Overall	62.50	62.16	62.16	64.15	64.15	4	0	0
Treatment	57.50	57.14	57.14	52.17	52.17	2	0	0
Control	67.50	66.67	66.67	73.33	73.33	2	0	0
Teacher years APES/APGOV teaching experience								
Overall	5.75	5.83	5.76	7.62	7.80	4	4	3
Treatment	4.86	4.97	4.97	6.78	6.78	2	0	0
Control	6.69	6.69	6.47	8.33	8.58	2	4	3
Teacher average APES/APGOV class size in 2015-16								

Overall	27.87	27.57	27.39	28.40	28.32	4	8	1
Treatment	28.38	27.76	27.14	29.46	29.46	2	4	0
Control	27.37	27.40	27.63	27.57	27.45	2	4	1
Baseline year (2015-16) % earning qualifying score on APES/APGOV exam, all students								
Overall	25.88	26.53	25.87	28.84	28.82	4	8	1
Treatment	24.50	25.70	25.00	28.83	28.83	2	4	0
Control	27.27	27.28	26.66	28.85	28.81	2	4	1
Baseline year (2015-16) % taking APES/APGOV exam								
Overall	85.71	86.27	85.03	86.37	86.09	4	8	1
Treatment	86.98	88.47	87.90	88.65	88.65	2	4	0
Control	84.43	84.32	82.47	84.57	84.13	2	4	1
Average national ELA test								
Overall	-0.24	-0.20	-0.20	-0.06	-0.06	4	0	0
Treatment	-0.23	-0.18	-0.18	0.05	0.05	2	0	0
Control	-0.25	-0.23	-0.23	-0.15	-0.15	2	0	0
Average national Math test								
Overall	-0.19	-0.14	-0.14	-0.10	-0.10	4	0	0
Treatment	-0.16	-0.13	-0.13	-0.00	-0.00	2	0	0
Control	-0.22	-0.15	-0.15	-0.17	-0.17	2	0	0
Average state ELA test								
Overall	282.81	282.70	282.70	286.96	286.04	4	0	4
Treatment	284.05	283.78	283.78	287.75	287.06	2	0	1
Control	281.54	281.74	281.74	286.31	285.25	2	0	3
Average state Math test								
Overall	307.54	309.10	309.10	309.74	308.28	4	0	4
Treatment	309.45	310.41	310.41	311.33	310.64	2	0	1
Control	305.58	307.92	307.92	308.44	306.47	2	0	3

Average state Science test								
Overall	166.03	168.58	168.58	169.33	169.11	4	0	1
Treatment	165.64	168.79	168.79	172.05	171.43	2	0	1
Control	166.45	168.40	168.40	167.33	167.33	2	0	0
Average taking AP test prior year								
Overall	49.29	63.03	63.03	57.19	57.19	4	0	0
Treatment	49.41	63.38	63.38	58.21	58.21	2	0	0
Control	49.17	62.72	62.72	56.40	56.40	2	0	0
Average student economic disadvantage								
Overall	50.30	57.72	57.72	52.61	52.61	4	0	0
Treatment	49.08	56.99	56.99	50.25	50.25	2	0	0
Control	51.58	58.37	58.37	54.42	54.42	2	0	0
% female students								
Overall	55.21	59.77	59.77	56.81	56.81	4	0	0
Treatment	54.98	59.75	59.75	56.18	56.18	2	0	0
Control	55.44	59.79	59.79	57.30	57.30	2	0	0
% Asian students								
Overall	12.41	15.74	15.74	11.77	11.77	4	0	0
Treatment	12.78	15.76	15.76	11.67	11.67	2	0	0
Control	12.02	15.71	15.71	11.84	11.84	2	0	0
% Hispanic students								
Overall	44.93	52.02	52.02	42.58	42.58	4	0	0
Treatment	42.94	51.02	51.02	41.40	41.40	2	0	0
Control	47.01	52.92	52.92	43.47	43.47	2	0	0
% Black students								
Overall	13.97	10.32	10.32	10.45	10.45	4	0	0
Treatment	13.61	10.27	10.27	8.62	8.62	2	0	0
Control	14.35	10.38	10.38	11.86	11.86	2	0	0
% White students								
Overall	26.50	43.49	43.49	29.54	29.54	4	0	0

Treatment	28.48	44.41	44.41	33.46	33.46	2	0	0
Control	24.41	42.67	42.67	26.54	26.54	2	0	0
Average student grade level								
Overall	11.32	11.44	11.44	11.52	11.52	4	0	0
Treatment	11.23	11.41	11.41	11.48	11.48	2	0	0
Control	11.41	11.46	11.46	11.54	11.54	2	0	0
Teacher counts								
Overall	82	74	74	53	53	NA	NA	NA
Treatment	42	35	35	23	23	NA	NA	NA
Control	40	39	39	30	30	NA	NA	NA

Table H2: APES teacher-level baseline (2015-16) characteristics overall, and by treatment and control at randomization (n=86), Year One (2016-17, n=74), and Year Two (2017-18, n=53)

Variable	Randomized (2016-17)	Year One (2016-17)	Year One (2016-17) including imputed	Year Two (2017-18)	Year Two (2017-18) including imputed	Randomized missing data (not imputed)	Imputed N (2016-17)	Imputed N (2017-18)
Teacher female								
Overall	67.39	66.67	66.67	67.74	67.74	3	0	0
Treatment	63.64	61.11	61.11	53.85	53.85	2	0	0
Control	70.83	70.83	70.83	77.78	77.78	1	0	0
Teacher years APES/APGOV teaching experience								
Overall	5.17	5.28	5.46	6.72	6.93	3	2	2
Treatment	4.40	4.56	4.56	5.69	5.69	2	0	0
Control	5.86	5.86	6.14	7.56	7.83	1	2	2
Teacher average APES/APGOV class size in 2015-16								
Overall	28.30	28.27	28.72	29.69	29.51	3	2	1

Treatment	28.82	28.85	28.85	31.56	31.56	2	0	0
Control	27.80	27.80	28.63	28.26	28.03	1	2	1
Baseline year (2015-16) % earning qualifying score on APES/APGOV exam, all students								
Overall	12.00	12.36	13.03	13.95	14.39	3	2	1
Treatment	10.98	11.61	11.61	12.45	12.45	2	0	0
Control	12.97	12.97	14.10	15.10	15.80	1	2	1
Baseline year (2015-16) % taking APES/APGOV exam								
Overall	79.98	81.02	80.80	80.15	79.87	3	2	1
Treatment	82.22	84.92	84.92	84.01	84.01	2	0	0
Control	77.84	77.84	77.70	77.21	76.89	1	2	1
Average national ELA test								
Overall	-0.46	-0.35	-0.35	-0.28	-0.28	3	0	0
Treatment	-0.48	-0.36	-0.36	-0.28	-0.28	2	0	0
Control	-0.44	-0.35	-0.35	-0.29	-0.29	1	0	0
Average national Math test								
Overall	-0.39	-0.29	-0.29	-0.32	-0.32	3	0	0
Treatment	-0.41	-0.32	-0.32	-0.29	-0.29	2	0	0
Control	-0.36	-0.26	-0.26	-0.34	-0.34	1	0	0
Average state ELA test								
Overall	276.71	278.58	278.58	283.25	282.15	3	0	4
Treatment	275.90	278.36	278.36	282.51	281.70	2	0	1
Control	277.45	278.74	278.74	283.83	282.48	1	0	3
Average state Math test								
Overall	303.10	306.26	306.26	302.06	300.56	3	0	4
Treatment	303.51	307.22	307.22	301.59	301.12	2	0	1
Control	302.74	305.54	305.54	302.44	300.16	1	0	3
Average state Science test								

Overall	160.27	164.84	164.84	163.69	163.50	3	0	1
Treatment	158.94	164.80	164.80	163.92	163.45	2	0	1
Control	161.49	164.87	164.87	163.54	163.54	1	0	0
Average taking AP exam prior year								
Overall	43.08	61.53	61.53	56.07	56.07	3	0	0
Treatment	38.82	60.63	60.63	52.32	52.32	2	0	0
Control	47.15	62.21	62.21	58.78	58.78	1	0	0
Average student economic disadvantage								
Overall	62.40	61.32	61.32	65.71	65.71	3	0	0
Treatment	64.04	61.25	61.25	66.53	66.53	2	0	0
Control	60.83	61.37	61.37	65.12	65.12	1	0	0
% female students								
Overall	57.42	59.91	59.91	58.56	58.56	3	0	0
Treatment	58.18	60.04	60.04	57.26	57.26	2	0	0
Control	56.68	59.80	59.80	59.50	59.50	1	0	0
% Asian students								
Overall	12.80	15.78	15.78	10.44	10.44	3	0	0
Treatment	14.67	16.06	16.06	10.77	10.77	2	0	0
Control	11.01	15.57	15.57	10.20	10.20	1	0	0
% Hispanic students								
Overall	59.06	56.92	56.92	57.63	57.63	3	0	0
Treatment	55.87	56.16	56.16	56.34	56.34	2	0	0
Control	62.12	57.48	57.48	58.57	58.57	1	0	0
% Black students								
Overall	11.02	9.91	9.91	7.72	7.72	3	0	0
Treatment	10.23	9.63	9.63	8.38	8.38	2	0	0
Control	11.77	10.13	10.13	7.24	7.24	1	0	0
% White students								
Overall	15.76	40.39	40.39	17.39	17.39	3	0	0
Treatment	17.81	41.16	41.16	20.42	20.42	2	0	0

Control	13.81	39.81	39.81	15.19	15.19	1	0	0
Average student grade level								
Overall	11.19	11.38	11.38	11.41	11.41	3	0	0
Treatment	10.93	11.29	11.29	11.33	11.33	2	0	0
Control	11.43	11.44	11.44	11.47	11.47	1	0	0
Teacher counts								
Overall	47	42	42	31	31		NA	NA
Treatment	23	18	18	13	13		NA	NA
Control	24	24	24	18	18		NA	NA

Table H3: APGOV teacher-level baseline (2015-16) characteristics overall, and by treatment and control at randomization (n=86), Year One (2016-17, n=74), and Year Two (2017-18, n=53)

Variable	Randomized (2016-17)	Year One (2016-17)	Year One (2016-17) including imputed	Year Two (2017-18)	Year Two (2017-18) including imputed	Randomized missing data (not imputed)	Imputed N (2016-17)	Imputed N (2017-18)
Teacher female								
Overall	55.88	56.25	56.25	59.09	59.09	1	0	0
Treatment	50.00	52.94	52.94	50.00	50.00	0	0	0
Control	62.50	60.00	60.00	66.67	66.67	1	0	0
Teacher years APES/APGOV teaching experience								
Overall	6.57	6.57	6.16	8.86	9.02	1	2	1
Treatment	5.41	5.41	5.41	8.20	8.20	0	0	0
Control	8.08	8.08	7.00	9.45	9.71	1	2	1
Teacher average APES/APGOV class size in 2015-16								
Overall	27.25	26.50	25.65	26.65	26.65	1	6	0
Treatment	27.77	26.25	25.32	26.72	26.72	0	4	0
Control	26.69	26.74	26.02	26.58	26.58	1	2	0
Baseline year (2015-16) % earning qualifying score on APES/APGOV exam, all students								
Overall	46.47	48.34	42.72	49.15	49.15	1	6	0
Treatment	43.42	45.20	39.17	50.14	50.14	0	4	0
Control	49.73	51.49	46.74	48.32	48.32	1	2	0
Baseline year (2015-16) % taking APES/APGOV exam								

Overall	94.20	94.33	90.60	94.86	94.86	1	6	0
Treatment	93.64	93.37	91.05	94.69	94.69	0	4	0
Control	94.80	95.29	90.09	94.99	94.99	1	2	0
Average national ELA test								
Overall	0.05	-0.00	-0.00	0.25	0.25	1	0	0
Treatment	0.06	0.02	0.02	0.47	0.47	0	0	0
Control	0.03	-0.03	-0.03	0.07	0.07	1	0	0
Average national Math test								
Overall	0.07	0.05	0.05	0.22	0.22	1	0	0
Treatment	0.13	0.07	0.07	0.38	0.38	0	0	0
Control	-0.01	0.01	0.01	0.09	0.09	1	0	0
Average state ELA test								
Overall	291.32	288.12	288.12	291.51	291.51	1	0	0
Treatment	294.02	289.51	289.51	294.03	294.03	0	0	0
Control	288.09	286.54	286.54	289.42	289.42	1	0	0
Average state Math test								
Overall	313.73	312.82	312.82	319.16	319.16	1	0	0
Treatment	316.72	313.78	313.78	323.03	323.03	0	0	0
Control	310.14	311.73	311.73	315.94	315.94	1	0	0
Average state Science test								
Overall	173.83	173.50	173.50	177.01	177.01	1	0	0
Treatment	173.39	173.02	173.02	181.80	181.80	0	0	0
Control	174.38	174.05	174.05	173.01	173.01	1	0	0
Average taking AP exam prior year								
Overall	57.65	65.00	65.00	58.76	58.76	1	0	0
Treatment	62.24	66.29	66.29	65.87	65.87	0	0	0
Control	52.20	63.54	63.54	52.83	52.83	1	0	0
Average student economic disadvantage								
Overall	34.05	52.99	52.99	34.16	34.16	1	0	0
Treatment	30.97	52.48	52.48	29.09	29.09	0	0	0

Control	37.70	53.57	53.57	38.38	38.38	1	0	0
% female students								
Overall	52.24	59.60	59.60	54.35	54.35	1	0	0
Treatment	51.10	59.44	59.44	54.78	54.78	0	0	0
Control	53.59	59.77	59.77	54.00	54.00	1	0	0
% Asian students								
Overall	11.88	15.68	15.68	13.64	13.64	1	0	0
Treatment	10.48	15.46	15.46	12.84	12.84	0	0	0
Control	13.55	15.93	15.93	14.30	14.30	1	0	0
% Hispanic students								
Overall	25.95	45.60	45.60	21.36	21.36	1	0	0
Treatment	27.29	45.57	45.57	21.99	21.99	0	0	0
Control	24.35	45.63	45.63	20.83	20.83	1	0	0
% Black students								
Overall	17.94	10.86	10.86	14.31	14.31	1	0	0
Treatment	17.71	10.94	10.94	8.94	8.94	0	0	0
Control	18.21	10.77	10.77	18.78	18.78	1	0	0
% White students								
Overall	40.91	47.56	47.56	46.67	46.67	1	0	0
Treatment	41.41	47.85	47.85	50.42	50.42	0	0	0
Control	40.31	47.25	47.25	43.55	43.55	1	0	0
Average student grade level								
Overall	11.50	11.52	11.52	11.66	11.66	1	0	0
Treatment	11.60	11.53	11.53	11.67	11.67	0	0	0
Control	11.37	11.51	11.51	11.66	11.66	1	0	0
Teacher counts								
Overall	35	32	32	22	22	NA	NA	NA
Treatment	19	17	17	10	10	NA	NA	NA
Control	16	15	15	12	12	NA	NA	NA

Appendix I: Descriptive Statistics for Students of Experimental Teachers

Table I1: 74-teacher sample's students' Year One (2016-17) baseline (2015-16) characteristics overall, and by treatment and control at randomization, across and by course

Variable		Overall	APES	APGOV
Counts				
	Overall	3,645	1,952	1,693
	Treatment	1,499	766	733
	Control	2,146	1,186	960
Economic disadvantage				
	Overall	42.77%	60.19%	22.68%
	Treatment	44.90%	61.49%	27.56%
	Control	41.29%	59.36%	18.96%
Female				
	Overall	55.61%	56.61%	54.46%
	Treatment	55.90%	58.62%	53.07%
	Control	55.41%	55.31%	55.52%
Grade level				
	Overall	11.44	11.28	11.62
	Treatment	11.43	11.30	11.57
	Control	11.44	11.28	11.65
Asian				
	Overall	14.10%	13.37%	14.94%
	Treatment	15.08%	17.49%	12.55%
	Control	13.42%	10.71%	16.77%
Hispanic				
	Overall	38.16%	55.02%	18.72%
	Treatment	37.83%	51.04%	24.01%
	Control	38.40%	57.59%	14.69%
Black				
	Overall	8.72%	8.25%	9.27%
	Treatment	8.87%	7.05%	10.78%
	Control	8.62%	9.02%	8.13%
White				
	Overall	36.08%	21.57%	52.81%
	Treatment	35.36%	22.72%	48.57%
	Control	36.58%	20.83%	56.04%
Standardized national Math				
	Overall	0.01	-0.29	0.37
	Treatment	0.01	-0.27	0.32
	Control	0.01	-0.30	0.41
Standardized national ELA				
	Overall	-0.03	-0.33	0.31

	Treatment	-0.05	-0.32	0.24
	Control	-0.03	-0.34	0.37
Standardized state Math				
	Overall	312.89	306.48	320.28
	Treatment	312.91	307.48	318.59
	Control	312.88	305.84	321.57
Standardized state ELA				
	Overall	287.47	281.07	294.85
	Treatment	288.11	281.69	294.81
	Control	287.03	280.68	294.88
Standardized state Science				
	Overall	172.35	164.10	181.87
	Treatment	171.12	163.88	178.68
	Control	173.21	164.23	184.31
Took AP exam in prior year				
	Overall	48.89%	38.99%	60.31%
	Treatment	52.43%	42.17%	63.17%
	Control	46.41%	36.93%	58.13%

Table I2: 53-teacher sample students' Year One (2016-17) baseline (2015-16) characteristics overall, and by treatment and control at randomization, across and by course

Variable		Overall	APES	APGOV
Counts				
	Overall	3,100	1,627	1,473
	Treatment	1,215	631	584
	Control	1,885	996	889
Economic disadvantage				
	Overall	39.65%	57.96%	19.42%
	Treatment	42.14%	59.75%	23.12%
	Control	38.04%	56.83%	16.99%
Female				
	Overall	55.77%	56.61%	54.85%
	Treatment	55.64%	57.69%	53.42%
	Control	55.86%	55.92%	55.79%
Grade level				
	Overall	11.48	11.32	11.65
	Treatment	11.54	11.47	11.60
	Control	11.44	11.22	11.68
Asian				
	Overall	14.84%	13.58%	16.23%
	Treatment	16.13%	17.75%	14.38%
	Control	14.01%	10.94%	17.44%
Hispanic				
	Overall	34.42%	50.89%	16.23%
	Treatment	34.65%	46.12%	22.26%

	Control	34.27%	53.92%	12.26%
Black				
	Overall	8.26%	8.60%	7.88%
	Treatment	7.74%	7.61%	7.88%
	Control	8.59%	9.24%	7.87%
White				
	Overall	39.26%	24.83%	55.19%
	Treatment	38.19%	26.62%	50.68%
	Control	39.95%	23.69%	58.16%
Standardized national Math				
	Overall	0.10	-0.22	0.45
	Treatment	0.15	-0.13	0.45
	Control	0.06	-0.28	0.45
Standardized national ELA				
	Overall	0.06	-0.24	0.40
	Treatment	0.08	-0.18	0.36
	Control	0.05	-0.28	0.43
Standardized state Math				
	Overall	315.31	308.93	322.36
	Treatment	315.98	311.53	320.79
	Control	314.88	307.29	323.38
Standardized state ELA				
	Overall	289.12	282.70	296.22
	Treatment	289.49	283.68	295.77
	Control	288.88	282.08	296.51
Standardized state Science				
	Overall	174.71	166.34	183.96
	Treatment	173.56	165.96	181.78
	Control	175.45	166.57	185.40
Took AP exam in prior year				
	Overall	60.65%	49.60%	72.84%
	Treatment	63.79%	58.00%	70.03%
	Control	58.62%	44.28%	74.69%

Table I3: 53-teacher sample's students' Year Two (2017-18) baseline characteristics overall, and by treatment and control at randomization, across and by course

Variable		Overall	APES	APGOV
Counts				
	Overall	2,946	1,646	1,300
	Treatment	1,186	675	511
	Control	1,760	971	789
Economic disadvantage				
	Overall	47.05%	61.48%	28.77%
	Treatment	46.63%	61.48%	27.01%
	Control	47.33%	61.48%	29.91%

Female				
	Overall	56.85%	58.63%	54.60%
	Treatment	56.66%	57.48%	55.58%
	Control	56.98%	59.42%	53.97%
Grade level				
	Overall	11.50	11.40	11.62
	Treatment	11.45	11.36	11.57
	Control	11.53	11.43	11.65
Asian				
	Overall	13.85%	11.66%	16.62%
	Treatment	14.33%	13.48%	15.46%
	Control	13.52%	10.40%	17.36%
Hispanic				
	Overall	36.25%	51.03%	17.54%
	Treatment	39.21%	52.15%	22.11%
	Control	34.26%	50.26%	14.58%
Black				
	Overall	8.83%	8.14%	9.69%
	Treatment	7.34%	7.11%	7.63%
	Control	9.83%	8.86%	11.03%
White				
	Overall	35.51%	22.72%	51.69%
	Treatment	33.98%	22.81%	48.73%
	Control	36.53%	22.66%	53.61%
Standardized national Math				
	Overall	0.05	-0.21	0.38
	Treatment	0.08	-0.12	0.34
	Control	0.02	-0.28	0.40
Standardized national ELA				
	Overall	0.09	-0.14	0.38
	Treatment	0.09	-0.13	0.38
	Control	0.09	-0.14	0.39
Standardized state Math				
	Overall	313.58	304.12	325.55
	Treatment	311.32	303.35	321.85
	Control	315.09	304.65	327.95
Standardized state ELA				
	Overall	288.33	284.21	293.53
	Treatment	287.52	283.98	292.19
	Control	288.87	284.37	294.40
Standardized state Science				
	Overall	173.01	166.20	181.63
	Treatment	172.53	166.10	181.02
	Control	173.33	166.27	182.02
Took AP exam in prior year				

Overall	60.35%	57.53%	63.92%
Treatment	61.64%	57.33%	67.32%
Control	59.49%	57.67%	61.72%

Appendix J: Impact Analysis Methodology

Within years, for both our experimental and non-experimental samples, we followed a standard randomized controlled trial analysis protocol for our impact analyses, assessing baseline equivalence and estimating intent-to-treat (ITT) effects, conducting pre-defined exploratory subgroup analyses for course and socio-economic status (SES) subgroups, and addressing sensitivity to modeling choices. To properly account for nested data, our primary analytic method for was Hierarchical Linear Modeling (HLM). We included in our ITT analysis all students with given outcomes of complier and non-complier teachers. In this section, we first describe covariates, imputation, and baseline equivalence, followed by the details of our HLM ITT model, such as covariate selection, accounting for multiple hypothesis tests, sensitivity, and our subgroup analytic approach.

While most analytic steps described were the same for the experimental and non-experimental approaches, there were three differences for non-experimental (i.e., Research Question Three Approach 2). First, we fit two-level HLMs at the teacher-level rather than school-level. Second, we used matching weights. Third, we did not include among our covariate set a measure of how many years of experience teachers had instructing their APGOV/APES course, as it was not available for non-experimental teachers from all five districts.

Covariates

Drawing from education literature, we chose the following covariates as substantively important for consideration in our impact models for both Year One (2016-17) and Year Two (2017-18). For student-level covariates, we consider the year prior as baseline (e.g., test score data from spring 2016 and earlier for 2016-17 students). For all teacher- and school-level covariates, including the performance of teachers' prior students, we used 2015-16 cohort data as baseline. All covariates were correlated to one or more AP outcomes with rho greater or equal to 0.1.

- Student-level covariates
  - Math and ELA prior achievement, as measured by national assessments (PSAT/SAT/ACT)
  - Math, ELA, and Science prior achievement, as measured by eighth-grade state standardized tests
  - Socio-economic status, as measured by student eligibility for free or reduced-price lunch[9]

[9] In four of the five participating districts, we used eligibility for free and reduced-price lunch as a dichotomous proxy for whether students were from lower- or higher income households. For the district in which lunch-program data was unavailable, we instead used students' home ZIP codes as a proxy for household income. We referred to the U.S. Census Bureau's American Community Survey 2012-2016 Five-Year Estimates to obtain median household incomes for each student's ZIP code in 2016. We then compared the medians to income-eligibility guidelines under the Department of Agriculture's Child Nutrition Programs. For a household size of four within the contiguous United States during this time, the income-eligibility guideline was $44,955. We designated each student in this district as eligible if their ZIP code's median household income fell below this threshold.

  - Sex
  - Race/Ethnicity
  - Grade level
  - Whether the student took any AP exam in spring 2016 or 2017 (baseline depending on year in which student was enrolled in APGOV/APES)
  - All prior achievement scores interacted with course
- Teacher-level covariates
  - Course (APES or APGOV)
  - Averages of teachers' 2015-16 APGOV/APES students' exam results (each 2016-17 or 2017-18 outcome uses the corresponding covariate from 2015-16) These include, for each teachers 2015-16 students: 1) percent of all students taking the exam; 2) percent among all students earning a qualifying score; 3) percent among exam-takers earning a qualifying score; 4) average free-response subscore among exam-takers; 5) average multiple-choice subscore among exam-takers, 6) average total score among exam-takers)
  - 2015-16 class size
  - Number of years teaching APES/APGOV (experimental arm only)
- School-level covariates
  - Proportion free/reduced-price lunch
  - Student-teacher ratio
  - Proportion of Students taking AP exam
  - All student-level covariates (except interactions) averaged at the school level

To place all covariates on the same scale, we standardized numeric covariates to mean 0 and variance 1. For Year One data, we standardized across the full experimental sample; for Year Two data we standardized across our full dataset (i.e., including experimental and non-experimental data from 2015-16 through 2017-18). We standardized because with unstandardized data we received warning messages from *lmer()* and *glmer()* in R, prompting us to rescale our variables, and logistic regression models were not converging.

Following Altonji and Mansfield (2018), we created school-level averages of student-level variables to help adjust for any bias due to baseline imbalance at the school level—the level of treatment assignment. We averaged over all students in KIA classrooms at each school. To avoid outcome-dependence in covariates, averages were over all eligible students as opposed to all students with outcomes. Though potential bias stemming from using only observed values should be equally distributed across treatment and control schools, averages also included multiply-imputed covariate values to avoid potential bias from using only observed values. We used imputed values from the AP exam-taking and qualifying-score outcome (full sample) imputation model, as this was the only model imputing values for all students, as opposed to only students with outcomes.

Multiple imputation

We imputed missing covariate values through multilevel joint modelling multiple imputation, implemented with the *jomoImpute* function from the *mitml* R package, which uses Carpenter and Kenward's (2013) MCMC algorithm. We used a two-level hierarchical linear model with district as a

fixed effect, almost paralleling our impact analysis—but for the imputation, we grouped students within teachers rather than within schools because we had several teacher-level covariates to impute. We imputed all covariates at both levels jointly according to one multivariate distribution. We imputed student-level covariates based on all other covariates at both levels, and imputed teacher-level covariates jointly with other covariates needing imputation, based on a model using observed teacher-level covariates and averages of observed student-level covariates.

Satisfying WWC 4.0 requirements for the respective imputation model, we always included as fixed effects all outcomes in the specified group, all covariates, the treatment indicator, and district indicator. We included the intercept for each variable as a random effect, allowed to vary from teacher to teacher.

We imputed missing covariates separately for experimental versus non-experimental observations because these two samples differed along observed, and likely unobserved dimensions. We also imputed missing covariates separately for each year of the experimental analysis.

We imputed covariate values separately for different outcome domains because the subsample with outcome data differs substantially by outcome. We fit separate imputation models for: 1) the AP total, multiple-choice, and free-response questions outcome sample (students who took the AP exam in all districts excepting District D); 2) the qualifying-score outcome sample (exam-takers only); and 3) the AP exam-taking and qualifying-score sample (all students in our sample).

The model had a burn-in of 10,000 iterations, with each of the 20 multiple imputations then taken 500 iterations apart.

WWC 4.0 also specifies that standard errors from the analysis must reflect the imputation, and mention that multiple imputation is one way to do so. To satisfy WWC, calculations must be based on at least five imputed datasets (we used 20), and account for: (1) the within-imputation variance component; (2) the between-imputation variance component; and (3) the number of imputations. We used Rubin's Rules for analysis on multiply-imputed datasets, which satisfies these requirements.

Baseline equivalence analysis

Establishing baseline equivalence is critical to all randomized controlled trials because even though randomization balances all covariates on average, it remains possible for some baseline covariates to be imbalanced by random chance. The rerandomization procedure promotes good balance on the baseline covariates used—in this study, baseline test scores and SES composites for 2015-16 students—and ensures better-than-random balance for all baseline covariates correlated with these two. However, it still was possible random chance imbalanced the covariates measured on the 2016-17 students, 2017-18 students, or other baseline covariates. Following standard randomized controlled trial protocol, we must empirically rule out this possibility. Moreover, we must assess baseline equivalence after attrition.

Although we always expect some covariates will be unbalanced just by random chance—at a 5% significance level, we expect 5% of differences to be significant just by random chance (without rerandomization)—our results are more trustworthy if we can empirically show our experiment was not conducted under a particularly unlucky randomization.

We divided baseline covariates into three categories: student-level standardized test-score data, student demographic data, and teacher-/class-/school-level covariates. We calculated effect size (ES) at baseline for each covariate according to WWC 4.0, though noting it only addressed baseline equivalence for student-level covariates. We conducted baseline equivalence analysis separately for each outcome group, as the analytic samples differed by outcome group. Each outcome sample was restricted to students who enrolled in their APGOV or APES class in first semester 2016-17 (Year 1) or 2017-18 (Year 2) with outcome data.

Continuous Covariates

For continuous covariates, define s_p to be a pooled individual-level standard deviation of the covariate, calculated as

$$s_p \equiv \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Regardless how the difference in means is calculated, this number always is calculated at the individual level.[10]

This denominator causes a slight upward bias in small sample sizes, so we follow WWC 4.0 and apply a small sample size adjustment, ω, where

$$\omega \equiv 1 - \frac{3}{4(n_1 + n_2) - 9}$$

although in our case the student sample sizes are large enough to render this irrelevant, as ω will be very close to 1.

The effect size then is calculated as

$$ES \equiv \omega \frac{\overline{X}_1 - \overline{X}_2}{s_p}$$

According to WWC 4.0, the difference in means can be calculated using either cluster- or individual-level data, as long as the weighting is consistent with the weighting used in the analysis—in our case, our outcome HLM model. Because we estimated impact with a two-level HLM, grouping students within schools and including district fixed effects, we calculated the difference in means for the numerator of the baseline equivalence with the treatment intercept from this HLM as well (as opposed to weighting either schools or students equally).

Let i denote individual student and s denote school (the unit of randomization). We calculated baseline equivalence for covariate X by fitting a two-level HLM grouping students within schools,

[10] As a check, we fit unadjusted models respectively fitting students within districts, and school-level averages within districts. The student-level weighting results were closer to the HLM results, so we calculated baseline equivalence for all covariates at the individual level.

with treatment W_s as the sole predictor. For student-level continuous outcomes, we calculated the individual-level model as:

$$X_i \sim N(\alpha_{s[i]}, \sigma_i^2)$$

and the school-level model as

$$\alpha_s \sim N(\mu + \tau W_s + \gamma D_s, \sigma_s^2)$$

where W_s is an indicator for treatment (Wave), τ is the coefficient we care about, and D_s is an indicator for district.

The baseline effect size is then be calculated as

$$ES \equiv \omega \frac{\hat{\tau}}{s_p}$$

Categorical Covariates

WWC 4.0 only defines effect sizes for dichotomous categorical variables, indicating that for covariates with multiple categories, we must first create a binary indicator for membership in each category, then assess equivalence for each of these indicators. For example, rather than looking at race as one covariate, we created an effect size for each indicator: White, Black, Asian, Hispanic, and Other. Under this format, we analyzed all categorical covariates as binary covariates or a collection of such.

For student-level binary covariates, the WWC effect size is defined as

$$ES \equiv \omega \frac{LOR}{1.65}$$

where LOR is the log odds ratio, defined as

$$\log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)$$

where p_1 is the probability that $X_i = 1$ in the treatment group.

The actual calculation of the numerator is the same as for continuous outcomes, where the LOR is the estimated τ coefficient for Wave from a two-level logistic HLM:

$$ES = \omega \frac{\hat{\tau}}{1.65}$$

The HLM school-level models are defined equivalently for continuous outcomes, but the individual-level model can be written as

$$P(X_i = 1) = logit^{-1}(\alpha_{s[i]})$$

Unequal Assignment Probabilities

Because our randomization procedure resulted in unequal assignment probabilities for two districts (only slightly different from 0.5 in District E, but substantially different from 0.5 in District B), we

adjusted using inverse probability weighting when fitting the HLM. Thus, in the HLM and all experimental analyses, each school is weighted by

$$w_s^{raw} = \frac{1}{P(W_s = W_s^{obs})}$$

We then normalized within each district so the overall weighted sample size (i.e., number of schools) stayed the same. Some statistical packages implicitly make this step, but fitting HLMs using the *lme4* package in R requires this explicitly:

$$w_s = n_s \frac{w_s^{raw}}{\sum w_s^{raw}}$$

where n_s is the total number of schools in each district and the sum is over all schools in the district. These weights were applied in all experimental analyses because these stemmed from the initial randomization, but not in the non-experimental analyses where randomization was irrelevant.

Teacher- and School-Level Covariates

For teacher- and school-level covariates, we calculated baseline equivalence in the same way described above, except it doesn't make sense to model these as an outcome yet also include a school-level random effect—so we exclude the school-level random effect terms and simply model each covariate as an outcome with treatment as a predictor and a district fixed effect. The exclusion of the school random intercepts will have resulted in different weighting than used for our HLM impact model, but we could not mimic the HLM model here because we could not include a school-level term with school-level outcomes. To remain as consistent as possible without impact analysis, we calculated baseline equivalence at the individual level, rather than the teacher or school level.

For course, there are some districts with only one level (District D is all APGOV, as District D is all APES), making it mathematically impossible to include a district fixed effect term. Thus, by necessity, we excluded the district fixed effects from the baseline equivalence models for course.

Bounding Imputed Baseline Data Differences

When covariates contain imputed data, WWC 4.1 requires bounding the baseline difference. As WWC is only concerned with student-level covariates, we followed this procedure—specified in Appendix C of the WWC 4.1 Standards Handbook—only for prior achievement measures, the only student-level covariates with any notable missingness.[11]

In Table J1, we list the student-level covariates that were not fully observed across 2016-17 and 2017-18 experimental APES and APGOV students, and 2017-18 APGOV and APES non-experimental students, as well as their proportion of missingness. Any student-level covariates not listed were fully observed excepting for negligible missingness on grade level and sex.

[11] There was negligible missingness on student grade level and sex, which we did not impute.

Table J1: Student-level covariates with any missingness and their proportion missing in full sample outcome groups

	Proportion missing			
	Year 1 experimental sample (n=74 teachers)	Year 1 experimental sample (n=53 teachers)	Year 2 experimental sample (n=53 teachers)	Year 2 non-experimental sample (n=66 teachers)
National Math	0.15	0.14	0.15	0.16
National ELA	0.15	0.14	0.15	0.16
Eighth-grade Math	0.34	0.35	0.40	0.40
Eighth-grade English	0.30	0.30	0.39	0.39
Eighth-grade Science	0.20	0.20	0.18	0.17
Sex	0	0	0.001	0.0002
Grade level	0.0005	0.0006	0.002	0

For these five measures of students' prior achievement, we calculated the average baseline effect size across our 20 imputed datasets. We calculated baseline effect sizes under the assumption of Missing at Random (MAR), then bounded them according to WWC, making several other assumptions about the degree to which the MAR assumption holds.

We let $\overline{x}_j$ denote the full-sample (unmeasured) covariate mean for group j. For baseline equivalence, we care about $\overline{x}_t - \overline{x}_c$, but when some covariate data are missing, we cannot directly calculate this difference. Instead, we can either estimate the baseline equivalence under the MAR assumption, or bound the baseline equivalence under deviations from the same.

According to WWC 4.1, when covariate data are imputed and the outcome is observed for all subjects in the analytic sample, standards require computation of the following for each outcome and covariate combination:

(a) The means and standard deviations of the outcome for the analytic sample, separately by group: $\overline{y}_j$ and s_{jy}. To retain weighting consistency with our impact analysis, we calculated our means using a two-level HLM. We computed standard deviations within treatment groups, with weights used for all other analyses.

(b) The means of the outcome for the subjects in the analytic sample with observed covariate data, by group: $\overline{y}_{jR}$. We also calculated these means using a two-level HLM.

(c) The correlation between the covariate and the outcome: ρ. Note: This is estimated using only observed data, per WWC 4.0.

(d) An estimate of the baseline difference based on study data, g_{xI}, denotes the estimate using imputed covariate data.

We then used (a) through (d) to estimate the baseline difference under the MAR assumption (D1), then, taking the maximum of the formulas specified in WWC 4.0, bound the baseline difference under deviations (D2-D4) from this MAR assumption:

$$D1 \equiv |g_{xI}|$$

$$D2 \equiv \left| g_{xI} + \frac{\omega}{s_y} \frac{1-\rho^2}{\rho} [\overline{y}_t - \overline{y}_{tR}] \right|$$

$$D3 \equiv \left| g_{xI} + \frac{\omega}{s_y} \frac{1-\rho^2}{\rho} [\overline{y}_c - \overline{y}_{cR}] \right|$$

$$D4 \equiv \left| g_{xI} + \frac{\omega}{s_y} \frac{1-\rho^2}{\rho} ([\overline{y}_t - \overline{y}_{tR}] - [\overline{y}_c - \overline{y}_{cR}]) \right|$$

It should be noted these formulas do not exactly match WWC 4.0 because our estimates, g_{xI}, already are scaled by ω.

Hierarchical Linear Modeling (HLM) Impact Analysis

We used Hierarchical Linear Modeling (HLM) as a model-based method of accounting for the inherent multilevel structure and to estimate the causal impact of Knowledge in Action after adjusting for covariate differences between treatment groups. We modeled the KIA intervention as a multisite, cluster-randomized trial experimental design, with school districts or sites serving as randomization blocks. The units of randomization were schools (i.e., clusters). Because most schools had one participating teacher and most students only took one course (APES or APGOV), we assumed students were nested within schools, the level of treatment assignment.[12] As eight of 68 Year One schools and three of 50 Year Two schools had more than one teacher, we placed teacher-level covariates at the student level rather than aggregating up to the school level.

Initially, we specified our primary analysis as a covariate-adjusted three-level HLM, nesting students within schools within districts, with both school and district intercepts estimated as random effects. However, this model resulted in a singular fit, with estimated group level variances very close to zero—the lower bound of the parameter space—for most outcomes: Out of 25, 20 resulted in a singular fit with this three-level model. To alleviate the singularity problem for our primary model, we replaced the district random effect with a district fixed-effect term, yielding the two-level HLM as described above. For Year One, we presented three-level model results as a sensitivity analysis demonstrating robustness of results to the two-level model.

For all outcomes, let Y_i denote the outcome for individual i, $W_{s[i]}$ denote the treatment assignment for individual i (clustered at the school level), and X_i denote the covariate matrix for individual i, which includes student-, teacher-, and school-level covariates, as well as indicator variables for district.

[12] In our benchmark model, we pooled student outcomes across courses. The initial sample-size calculations for the KIA Efficacy Study assumed pooled outcomes.

All HLM models include fixed effect coefficients, β, for the covariates and district terms, and random intercepts for each school, α_s. The treatment effect coefficient is denoted by τ, so our effect sizes will be based on our estimate of τ, $\hat{\tau}$. The specific functional form for the individual level of the HLM model varies by outcome type, quantitative and binary, as elaborated on below.

Quantitative Outcomes

For quantitative outcomes, we can write the individual-level model as:

$$Y_i \sim N(\alpha + \tau W_{s[i]} + \beta X_i + \alpha_{s[i]}, \sigma_i^2)$$

where X_i includes all covariate values associated with individual *i* (student, teacher, school level, and district indicator), $\alpha_{s[i]}$ is the random intercept for school *s* associated with individual *i*, and $W_{s[i]}$ is the treatment assignment for school *s* associated with individual *i*. We then model the random intercepts as follows:

$$\alpha_s \sim N(0, \sigma_s^2)$$

Alternatively, we could have written the model by breaking up the covariates then placing student-level variables at the individual level and school-level variables in the school-level model. While the fixed-effect coefficients do not change, this does alter the interpretation of the random intercepts. Because this is what the software actually fits, with random effects centered around 0, we present it as above.

For the variances, σ_i^2 is the individual-level residual variance, and σ_s^2 represents how much the random intercepts vary across schools.

For quantitative outcomes, we fit models using the *lmer* function from the *lme4* package in R, which fits using restricted maximum likelihood (REML) to help estimate the variance of random effects.

Effect sizes

From the estimated treatment coefficient, $\hat{\tau}$, we then calculate the effect size, standardized mean difference, as Hedge's g, consistent with WWC 4.0 and following the same procedures described above under "Baseline Equivalence."

Unequal Assignment Probabilities

With the randomization procedure resulting in unequal assignment probabilities for two districts, when fitting the HLM in the experimental analyses we adjusted using inverse probability weighting, as described in the baseline equivalence section.

Variable Selection

Following WWC guidelines, we included in the corresponding impact model any covariate with an absolute baseline effect size > 0.05. We also included in all impact models the course (APES or

APGOV) covariate, because the AP examinations differed for APGOV as compared to APES students.

While this ensures we retained covariates we should be including to adjust for baseline imbalance, it does not ensure we retained covariates strongly associated with the outcome that we should be including to maximize precision of estimates. Due to the large number of covariates available, especially at the teacher- and school-level and relative to the number of schools, simply including all available covariates would result in overfitting. However, we also did not want to ignore important covariates that were balanced at baseline but could benefit the precision of the estimates. Therefore, we automated variable selection for any remaining covariates not already forced in due to baseline imbalance.

Only covariates considered potentially substantively important are on our full covariate list; hence, we used automation to supplement, rather than replace, subject-matter expertise. We automated the remaining variable selection to maintain objectivity. To select variables across multiply imputed datasets, the automated algorithm uses a two-step procedure, as developed by Brand (2003) and recommended in https://stefvanbuuren.name/fimd/sec-stepwise.html. Step 1 performs variable selection separately on each multiply-imputed dataset and retains covariates selected in most imputed datasets, then Step 2 performs backwards selection on these retained covariates by calculating p-values jointly across all imputed datasets. In step 1, we use forward selection minimizing the Akaike Information Criterion (AIC) on each imputed dataset (here all models are fit with maximum likelihood as opposed to REML so AIC makes sense), and in step 2 we perform backwards selection based on p-values calculated according to Rubin's Rules across all imputed datasets (Wood 2008). Interactions are considered in the automated variable selection, but not forced.

Inference

Once the covariates were selected, to account for any added uncertainty due to missing covariate data imputation, we fit the appropriate model to each of the 20 multiply-imputed datasets. To give overall estimates and corresponding inferences, we combined the resulting estimates according to Rubin's Rules for multiple imputation. According to Rubin's Rules, if $\hat{\tau}_m$ represents the point estimate from the m^{th} imputed dataset, the overall point estimate simply averages the estimates from the $M = 20$ different imputations:

$$\hat{\tau} = \frac{1}{M}\sum_{m=1}^{M}\hat{\tau}_m$$

For the standard error, we calculate the average within imputation variance:

$$\overline{V} = \frac{1}{M}\sum_{m=1}^{M} v\,ar(\hat{\tau}_m)$$

and the between imputation variance:

$$B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\tau}_m - \hat{\tau})^2$$

and combine these for the overall variance:

$$var(\hat{\tau}) = \overline{V} + (1 + M^{-1})B$$

The standard error reflects the clustering by school due to the two-level HLM grouped by school, except in the case of singular models that effectively eliminate the random intercept for school. In these cases, p-values are computed according to randomization-based inference, described below, reflecting the randomization at the school level.

We computed p-values according to two-sided hypothesis tests,

$$H_0: \tau = 0 \; vs \; H_a: \tau \neq 0$$

Research Question Three Approach 1 inference

Since we could not analytically calculate the true variance for the Research Question 3 Approach 1 experimental combined estimate $\hat{\tau}_2$, we instead conducted inference for $\hat{\tau}_2$using a randomization test. This was possible since the variance of the one year estimate, $\hat{\tau}_1$ and the two minus one year estimate, $\hat{\tau}_{2-1}$ both stem from the same initial randomization. We simulated 5,000 randomizations according to the original rerandomization scheme used to randomize schools within districts, and under the sharp null hypothesis that outcomes are unaffected by treatment assignment, recalculated $\hat{\tau}_1$, $\hat{\tau}_{2-1}$, and $\hat{\tau}_2$ using each simulated treatment assignment. For each simulation, we calculated $\hat{\tau}_1$, $\hat{\tau}_{2-1}$, and $\hat{\tau}_2$ following the same methodology as our primary impact analysis, with the same covariates and models—only changing the treatment assignment vector. This gives us a randomization distribution of each statistic under the null hypothesis of no KIA effect. For all $\hat{\tau}_2$ estimates, and for $\hat{\tau}_1$ or $\hat{\tau}_{2-1}$ for models resulting in a singular fit in our impact analysis, we then computed the p-value as the proportion of statistics calculated from the simulated randomizations at least as extreme as our observed statistic from the actual randomization. In addition to providing valid inferences for our combined estimate, this approach has the benefit of reflecting not just the school-level randomization, but the original rerandomization scheme used for our experimental design.

Multiple hypothesis testing

When examining impacts of a program on several outcomes within the same domain—as is the case for several measures of AP performance—we faced an inflated "Type I error rate," in which the chance of falsely finding a significant effect increases simply because we were conducting so many statistical tests. Because each test has a 5% chance of making a Type I error (rejecting the null hypothesis when it is, in fact, true), if we run many tests (one for each outcome), the overall family-wise Type I error rate compounds and increases beyond the point of acceptability. For example, with 20 different tests at a significance level of $\alpha = 0.05$, we should expect one test (0.05*20) to be significant just by random chance. When conducting multiple tests, adjustment must be made to counter the compounding Type I error rate.

There are many ways to adjust for multiple tests. A WWC-recommended approach is to lower the threshold for statistical significance for any individual test based on the number of tests conducted, making it harder to achieve statistical significance. When multiple tests are conducted, to preserve the nominal significance level of Type I Error rate of $\alpha = 0.05$, the significance threshold for each individual p-value is actually a value lower than 0.05. The exact threshold depends on the specific method; we used the Benjamini-Hochberg procedure, which is the WWC's recommended procedure for multiple testing. Regardless of method, the more tests conducted means the more difficult it is to achieve significant results for any one outcome. Besides needing to account for this in analysis, the primary implication is that we must take care to limit the number of tests performed by limiting the number of outcome variables subject to adjustments for multiple comparisons.

Again aligning with WWC standards, multiple testing adjustments should be made within outcome domains. For our primary analysis, we only used one analytic method (the Intent to Treat HLM model on all units with outcome data included in our analytic sample), with other models serving as forms of robustness assessment, exploring the sensitivity of the results to analytic modeling choices. Therefore, we did not adjust for multiple testing due to multiple analyses.

Subgroup analysis

Because the central goal of the Knowledge in Action RCT was to estimate differences between outcomes among students whose teachers were in either treatment or control, our confirmatory comparisons within domains were differences between the outcomes of students of teachers randomly assigned to treatment or control. While we also were interested in other differences (e.g., by course), we defined all subgroup analyses conducted as exploratory due to limitations of multiple hypothesis testing. We decided to conduct exploratory analyses of differences by course (APGOV and APES), and student socio-economic status. We defined our course and SES subgroups *a priori* with our rationale for looking for differences between these groups based on theory. In addition to defining socio-economic subgroups at the student level, we also analyzed socio-economic subgroups at the district level, by estimated effects separately within the two districts with lower proportions of students eligible for FRLP (Districts B and D) and the three districts with higher proportions of students eligible for FRLP (Districts A, C, and E).

For the subgroup analysis, we used our primary HLM model, adding interaction terms between treatment and the binary subgroup indicator (Wang, Lagakos, Ware, Hunter, & Drazen, 2007). As with our primary analysis, we fit this model on all multiply-imputed datasets and combined resulting inferences via Rubin's Rules. We calculated effect sizes using the overall standard deviations, as opposed to subgroup-specific standard deviations, given that we calculated all from the overall model.

We did not conduct subgroup analyses as part of our non-experimental approach. As we explain further in Appendices X and through CC, our match teacher quality, while acceptable overall, varied considerably at the district level. As districts in our study are inextricably linked with APES and APGOV courses, as well as student SES, non-experimental arm subgroup analysis on these dimensions was not appropriate.

Appendix K: Student-level Outcome Sample Attrition

Even for outcome samples meeting attrition thresholds at the student level, all study outcomes exceed thresholds at the cluster level; thus, no study outcomes can meet standards without reservations. Nonetheless, student-level attrition calculations are informative. Here we present experimental student attrition calculations for the AP qualifying score (exam-takers only) and continuous score outcomes for Years One (74- and 53-teacher sample) and Two (53-teacher sample), followed by student attrition calculations for Year One (74-teacher sample) CWRA+ and student survey outcomes. All calculations start within schools with at least one student outcome; that is, following WWC we do not double-count school and student attrition.

AP Qualifying Score (exam-takers only)

Within the subgroup of exam-takers, Year One (74-teacher) overall attrition on having a qualifying score was 19%, with a 4-percentage point differential, meeting the WWC "cautious boundary" threshold. Similarly, for the Year One 53-teacher sample, overall attrition was 17%, with a 3-percentage point differential (Table K1).

For Year Two (53-teacher), overall student-level attrition also was 17%, this time with an 8.5 percentage point differential, meeting the WWC "optimistic boundary" threshold.

Table K1: Year One (74-teacher), Year One (53-teacher), and Year Two (53-teacher) student-level attrition on AP qualifying score (exam-takers only)

	Overall	Treatment	Control	Difference
Year One (74-teacher)				
Full sample (n)	3,645	1,499	2,146	
Exam-taker subsample (n)	2,963	1,255	1,708	
Attrition	19%	16%	20%	4 percentage points
Year One (53-teacher)				
Full sample (n)	3,100	1,215	1,855	
Exam-taker subsample (n)	2,574	1,034	1,540	
Attrition	17%	15%	12%	3 percentage points
Year Two (53-teacher)				
Full sample (n)	2,946	1,186	1,760	
Exam-taker subsample (n)	2,436	1,041	1,395	
Attrition	17%	12%	21%	11 percentage points

Continuous scores

Within the continuous sample subgroup of exam-takers in schools with at least one continuous score outcome, for the 74-teacher Year One sample, overall attrition was 29% with a 6-percentage point differential, meeting the WWC optimistic threshold (Table J2).

For the 53-teacher sample, Year One overall attrition was 35% with an 11-percentage point differential. For Year 2, overall student-level attrition was 29%, this time with a 10-percentage point

differential. Thus for both 53-teacher samples, student-level attrition on the continuous AP score outcome samples exceeded the WWC cautious and optimistic thresholds.

Table K2: Year One (74-teacher), Year One (53-teacher), and Year Two (53-teacher) student-level attrition on AP continuous score outcomes

	Overall	Treatment	Control	Difference
Year One (74-teacher)				
Full sample (n)	2,249	936	1,313	
Exam-taker subsample (n)	1,599	697	902	
Attrition	29%	26%	31%	6 percentage points
Year One (53-teacher)				
Full sample (n)	2,042	764	1,278	
Exam-taker subsample (n)	1,318	544	774	
Attrition	35%	29%	39%	11 percentage points
Year Two (53-teacher)				
Full sample (n)	2,009	790	1,219	
Exam-taker subsample (n)	1,424	608	816	
Attrition	29%	23%	33%	10 percentage points

CWRA+

Attrition exceeded WWC thresholds for all three CWRA+ outcomes (Tables K3, K4, K5).

Table K3: Year One (74-teacher) student-level attrition on CWRA+ total score outcomes

	Overall	Treatment	Control	Difference
Year One (74-teacher)				
Full sample (n)	1,455	688	767	
Exam-taker subsample (n)	489	184	305	
Attrition	66%	73%	60%	13 percentage points

Table K4: Year One (74-teacher) student-level attrition on CWRA+ performance score outcomes

	Overall	Treatment	Control	Difference
Year One (74-teacher)				
Full sample (n)	1,477	702	775	
Exam-taker subsample (n)	565	229	336	
Attrition	62%	67%	57%	10 percentage points

Table K5: Year One (74-teacher) student-level attrition on CWRA+ selected response question score outcomes

	Overall	Treatment	Control	Difference
Year One (74-teacher)				
Full sample (n)	1,463	688	775	
Exam-taker subsample (n)	534	202	332	
Attrition	63%	71%	57%	13 percentage points

Student survey

For the student survey (Year One 74-teacher sample only), overall attrition was 52% with a 1.9 percentage point differential, meeting the WWC "optimistic boundary" threshold (Table K6).

Table K6: Year One (74-teacher) student-level attrition on survey outcomes

	Overall	Treatment	Control	Difference
Year One (74-teacher)				
Full sample (n)	1,560	722	838	
Exam-taker subsample (n)	744	337	407	
Attrition	52%	53.3%	51.4%	1.9 percentage points

Appendix L: Research Question One Student-Level Baseline Equivalence, Overall Sample

AP Outcomes

WWC reviewers likely will evaluate our results on AP outcomes using the Transition to College review protocol, which specifies that baseline equivalence must be established on measures of prior achievement (we have eighth-grade standardized test scores and national assessments) and SES (we have an indicator for economically disadvantaged).

In the Appendix L tables, we show standardized mean differences between treatment and control students on all student-, teacher-, and school-level covariates and associated p-values, for outcomes as specified. We include as a covariate the 2015-16 AP score average corresponding specifically to that outcome. For this reason, there are blank cells in Table L1 for covariates describing 2015-16 AP performance. Student-level prior achievement variables assume missing at randomness (MAR) in Tables L1, and bound according to potential deviations from MAR in Table L2. Figures L1, L2, and L3 visualize the content of Table L1.

Table L1: Baseline standardized mean differences and p-values between treatment and control students on all student-, teacher-, and school-level covariates, for Year One Research Question One AP outcome analytic samples

	AP QS (full sample) and exam-taking (n=3,645)		AP QS (exam-takers only) (n=2,963)		AP continuous scores (n=1,599)	
	SMD	p-value	SMD	p-value	SMD	p-value
National Assessment Math	-0.003	0.981	-0.082	0.632	-0.055	0.680
National Assessment ELA	0.051	0.694	-0.111	0.526	-0.021	0.876
Eighth-grade Math	0.084	0.514	-0.036	0.821	0.055	0.690
Eighth-grade English	0.095	0.487	-0.023	0.892	0.088	0.548
Eighth-grade Science	0.001	0.996	-0.105	0.503	-0.038	0.741
Economically disadvantaged	0.010	0.952	0.103	0.679	0.010	0.958
Took any AP exam in 2016	0.075	0.780	0.148	0.511	0.038	0.899
Female	0.019	0.675	-0.057	0.373	-0.017	0.732
Grade	-0.088	0.657	-0.002	0.989	-0.050	0.809
Asian	0.114	0.559	-0.029	0.918	-0.013	0.963
Hispanic	-0.085	0.685	-0.079	0.782	-0.080	0.716
Black	0.008	0.962	0.065	0.806	0.099	0.612

White	0.182	0.472	0.236	0.510	0.142	0.591
2016 average AP score					-0.218	0.331
2016 % earning AP QS (full sample)					-0.241	0.271
2016 % earning AP QS (exam-takers)	-0.183	0.360				
2016 % taking AP exam	0.143	0.536				
2016 Average total fine-grained score			-0.004	0.991		
2016 Average MC fine-grained score			0.028	0.929		
2016 average FR fine-grained score			-0.018	0.955		
Years teaching APES/APGOV	-0.182	0.465	0.103	0.734	-0.150	0.561
Course: APGOV	0.178	0.634	0.230	0.629	0.048	0.900
2015-16 average class size	-0.045	0.845	-0.154	0.559	-0.117	0.588
% free/reduced-price lunch	-0.012	0.944	-0.142	0.607	0.029	0.864
Student-teacher ratio	-0.106	0.357	-0.123	0.374	-0.138	0.195
% taking an AP exam	-0.018	0.949	0.151	0.651	-0.095	0.752
Average national Math	0.081	0.728	0.179	0.629	0.096	0.680
Average national ELA	0.038	0.864	0.079	0.810	0.033	0.880
Average eighth-grade Math	0.081	0.700	0.167	0.559	0.076	0.717
Average eighth-grade ELA	0.132	0.556	0.075	0.808	0.138	0.527
Average eighth-grade Science	-0.044	0.820	0.035	0.905	-0.049	0.791
Proportion school low SES	0.048	0.765	0.176	0.452	0.058	0.697
Proportion school taking 2016 AP exam	0.201	0.468	0.439	0.213	0.135	0.645
Proportion female	0.072	0.774	-0.013	0.961	-0.010	0.968
Average grade	0.005	0.985	0.046	0.897	-0.059	0.845
Proportion Asian	0.118	0.656	0.372	0.287	0.115	0.671
Proportion Hispanic	-0.065	0.753	-0.195	0.527	-0.047	0.815
Proportion Black	-0.035	0.855	-0.061	0.813	-0.031	0.875
Proportion White	0.037	0.851	0.058	0.824	0.013	0.949

Figure L1: Baseline standardized mean differences and p-values between treatment and control students on all student-, teacher-, and school-level covariates, for Year One Research Question One AP qualifying score (full) analytic sample

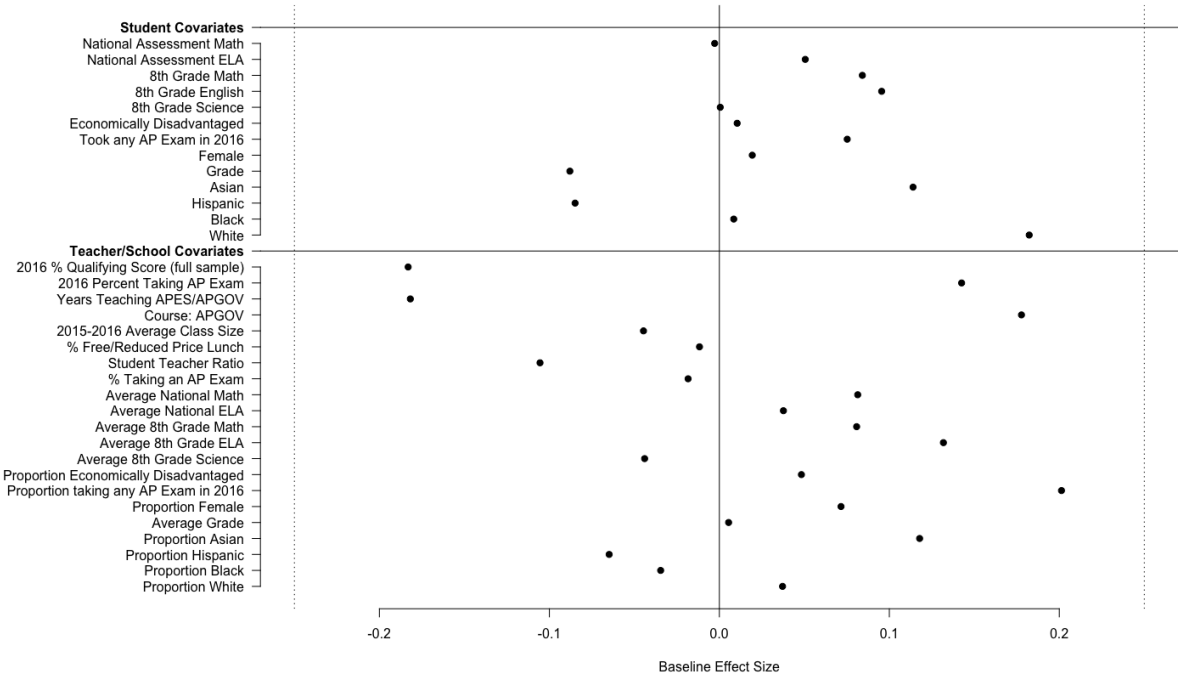


Figure L2: Baseline standardized mean differences and p-values between treatment and control students on all student-, teacher-, and school-level covariates, for Year One Research Question One AP qualifying score (exam-takers only) analytic sample

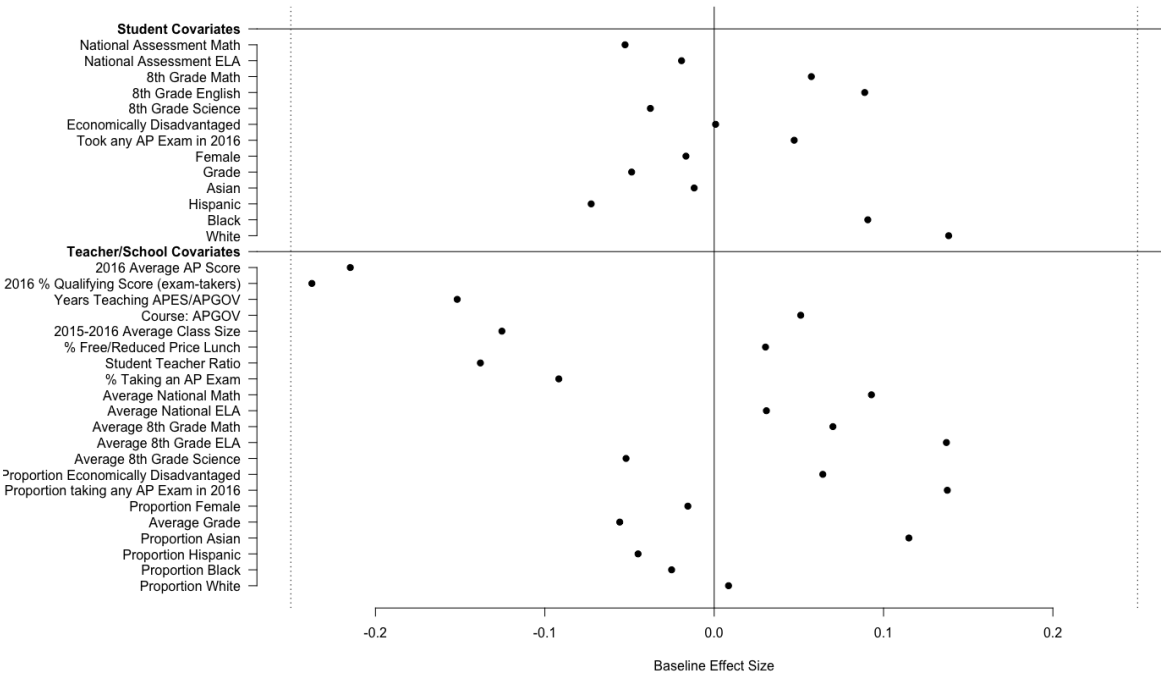


Figure L3: Baseline standardized mean differences and p-values between treatment and control students on all student-, teacher-, and school-level covariates, for Year One Research Question One AP continuous score analytic sample

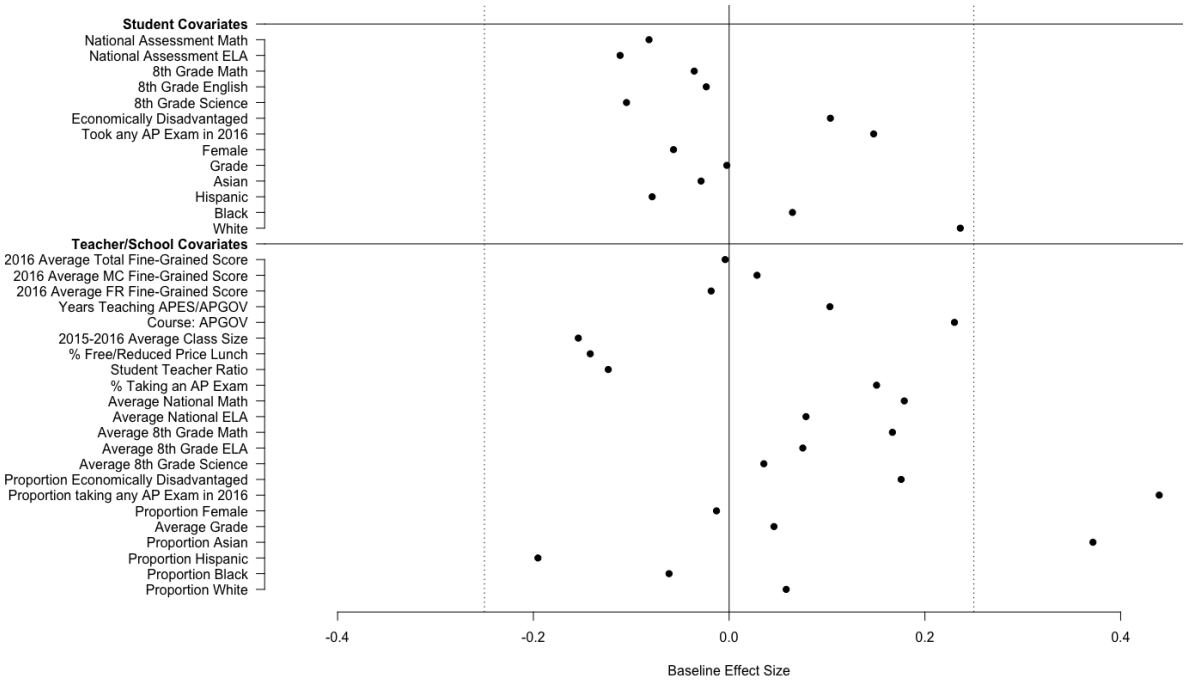


Table L2: Baseline standardized mean difference between treatment and control students on imputed student-level prior achievement covariates, for all Year One Research Question One AP analytic outcome samples

Took AP exam

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.003	0.385	0.211	0.172	0.385
National Assessment ELA	0.051	0.410	0.249	0.211	0.410
Eighth-grade Math	0.084	0.545	0.449	0.180	0.545
Eighth-grade English	0.095	0.539	0.442	0.192	0.539
Eighth-grade Science	0.001	0.113	-0.141	0.255	0.255

AP earn qualifying score (full sample)

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.003	-0.010	-0.049	0.035	-0.049
National Assessment ELA	0.051	0.044	0.014	0.081	0.081
Eighth-grade Math	0.084	0.067	0.061	0.090	0.090
Eighth-grade English	0.095	0.082	0.068	0.110	0.110
Eighth-grade Science	0.001	0.003	0.012	-0.008	0.012

AP earn qualifying score (exam-takers only)

	D1	D2	D3	D4	Most extreme

National Assessment Math	-0.055	-0.073	-0.114	-0.014	-0.114
National Assessment ELA	-0.021	-0.035	-0.067	0.011	-0.067
Eighth-grade Math	0.055	0.024	0.010	0.069	0.069
Eighth-grade English	0.088	0.056	0.042	0.103	0.103
Eighth-grade Science	-0.038	-0.044	-0.023	-0.059	-0.059

AP total score

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.082	-0.087	-0.065	-0.104	-0.104
National Assessment ELA	-0.111	-0.115	-0.099	-0.127	-0.127
Eighth-grade Math	-0.036	-0.017	0.023	-0.076	-0.076
Eighth-grade English	-0.023	-0.005	0.032	-0.060	-0.060
Eighth-grade Science	-0.105	-0.112	-0.089	-0.128	-0.128

AP multiple-choice score

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.082	-0.095	-0.072	-0.104	-0.104
National Assessment ELA	-0.111	-0.120	-0.105	-0.127	-0.127
Eighth-grade Math	-0.036	-0.009	0.004	-0.048	-0.048
Eighth-grade English	-0.023	0.003	0.013	-0.033	-0.033
Eighth-grade Science	-0.105	-0.110	-0.090	-0.124	-0.124

AP free-response score

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.082	-0.086	-0.064	-0.104	-0.104
National Assessment ELA	-0.111	-0.115	-0.097	-0.129	-0.129
Eighth-grade Math	-0.036	-0.033	0.051	-0.120	-0.120
Eighth-grade English	-0.023	-0.021	0.059	-0.104	-0.104
Eighth-grade Science	-0.105	-0.116	-0.083	-0.138	-0.138

CWRA Outcomes

WWC reviewers likely will evaluate our results on CWRA outcomes using the Transition to College review protocol, which specifies that baseline equivalence be established on measures of prior achievement students' families' household income.

Table L3: Baseline standardized mean differences and p-values between treatment and control students on all student-, teacher-, and school-level covariates, for all CWRA+ analytic outcome samples in Year One

	CWRA+ overall score		CWRA+ performance task subscore		CWRA+ selected response subscore	
	SMD	p-value	SMD	p-value	SMD	p-value
National Assessment Math	-0.065	0.754	-0.085	0.644	-0.023	0.910

National Assessment ELA	0.071	0.726	-0.001	0.998	0.113	0.563
Eighth-grade Math	0.174	0.396	0.088	0.629	0.198	0.312
Eighth-grade English	0.086	0.682	0.027	0.888	0.142	0.476
Eighth-grade Science	-0.102	0.575	-0.097	0.547	-0.083	0.636
Economically disadvantaged	-0.053	0.867	0.272	0.269	-0.151	0.620
Took any AP exam in 2016	-0.324	0.426	-0.356	0.344	-0.174	0.647
Female	0.180	0.145	0.231	0.037*	0.211	0.073
Grade	-0.093	0.696	-0.065	0.776	0.034	0.887
Asian	0.239	0.157	0.245	0.195	0.221	0.239
Hispanic	-0.244	0.536	-0.061	0.858	-0.119	0.748
Black	-0.037	0.924	-0.178	0.623	-0.099	0.775
White	-0.190	0.583	-0.191	0.536	-0.123	0.707
2016 average AP score	-0.204	0.297	-0.185	0.361	-0.193	0.340
2016 % taking AP exam	0.553	0.032*	0.390	0.160	0.499	0.049*
Years teaching APES/APGOV	-0.557	0.082	-0.548	0.057	-0.552	0.069
Course: APGOV	0.029	0.953	0.095	0.848	0.097	0.845
2015-16 average class size	-0.258	0.479	-0.122	0.694	-0.234	0.497
% free/reduced-price lunch	0.119	0.462	0.106	0.505	0.113	0.485
Student-teacher ratio	-0.093	0.507	-0.121	0.405	-0.102	0.457
% taking an AP exam	-0.004	0.991	-0.036	0.906	-0.006	0.984
Average national Math	-0.050	0.852	-0.036	0.887	-0.036	0.891
Average national ELA	0.128	0.608	0.102	0.658	0.134	0.580
Average eighth-grade Math	0.072	0.776	0.071	0.762	0.073	0.768
Average eighth-grade ELA	0.130	0.659	0.110	0.685	0.128	0.652
Average eighth-grade Science	-0.189	0.459	-0.168	0.475	-0.178	0.468
Proportion econ disadvantage	0.011	0.962	0.072	0.748	0.012	0.958
Proportion school taking 2016 AP exam	-0.039	0.899	0.014	0.960	-0.013	0.965
Proportion female	0.402	0.205	0.391	0.181	0.358	0.239
Average grade	-0.081	0.867	-0.006	0.989	-0.048	0.917
Proportion Asian	0.318	0.311	0.418	0.180	0.310	0.324

Proportion Hispanic	0.117	0.601	0.027	0.902	0.095	0.663
Proportion Black	-0.131	0.675	-0.153	0.553	-0.167	0.544
Proportion White	-0.203	0.352	-0.127	0.549	-0.157	0.472

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Figure L4: Baseline standardized mean difference between treatment and control students on all student-, teacher-, and school-level covariates, for CWRA+ overall score analytic outcome samples in Year One

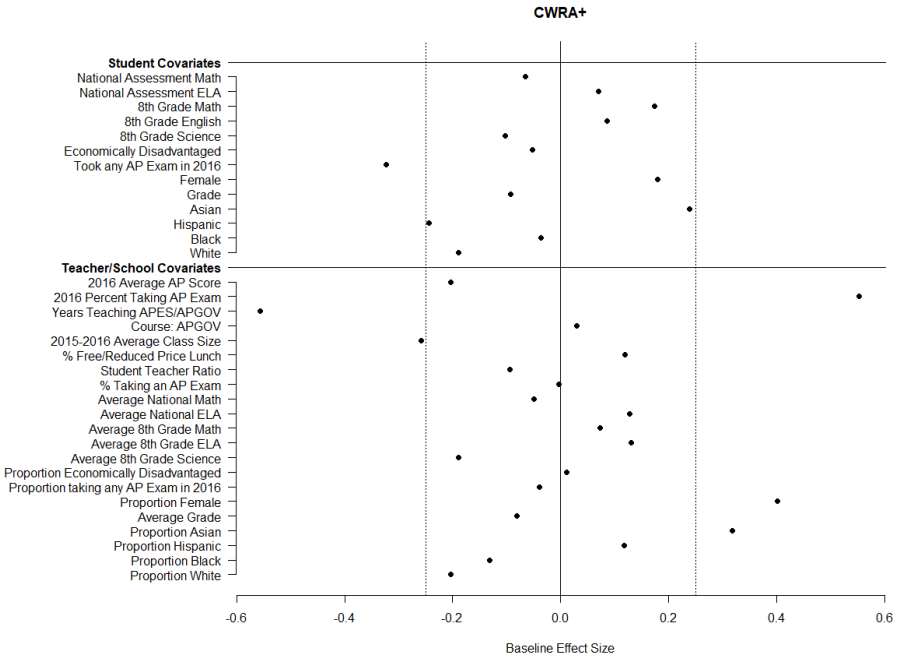


Table L4: Baseline standardized mean difference between treatment and control students on imputed student-level prior achievement covariates, for CWRA+ analytic outcome samples in Year One

CWRA+ overall score

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.065	0.022	-0.074	0.031	-0.074
National Assessment ELA	0.071	0.126	0.065	0.132	0.132
Eighth-grade Math	0.174	0.020	0.363	-0.168	0.363
Eighth-grade English	0.086	-0.005	0.210	-0.129	0.210
Eighth-grade Science	-0.102	-0.026	0.013	-0.142	-0.142

CWRA+ performance task subscore

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.085	0.493	0.013	0.395	0.493
National Assessment ELA	-0.001	0.297	0.050	0.246	0.297
Eighth-grade Math	0.088	0.225	0.550	-0.237	0.550
Eighth-grade English	0.027	0.104	0.380	-0.249	0.380
Eighth-grade Science	-0.097	0.034	0.093	-0.156	-0.156

CWRA+ selected response subscore

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.023	0.050	-0.030	0.058	0.058
National Assessment ELA	0.113	0.163	0.108	0.168	0.168
Eighth-grade Math	0.198	0.059	0.252	0.004	0.252
Eighth-grade English	0.142	0.053	0.179	0.016	0.179
Eighth-grade Science	-0.083	-0.091	-0.036	-0.139	-0.139

Survey Outcomes

WWC reviewers likely will evaluate our results on survey outcomes using the Character Education review protocol, which specifics baseline equivalence be established on the following[13]:

1) Pretest scores for at least one outcome measure
2) Grade level or age
3) Gender
4) Any special status such as special education, ELL, etc.
5) Location of the schools involved (urban, suburban, or rural; geographical region)

As we do not have pretest scores, most likely we would not be able meet WWC standards for the survey outcomes. We calculated baseline equivalence on survey outcomes to inform impact model variable selection.

[13] This protocol is from 2006 and may have been updated.

Table L5: Baseline standardized mean difference between treatment and control students on all student-, teacher-, and school-level covariates, for all analytic survey outcome samples in Year One

	Collaboration	Opportunities for leadership	Self-efficacy	Grit	Growth mindset	Appreciation for diversity	Civic/political efficacy	Whether expects to vote regularly at 18	Participatory citizenship	Interest in politics	Political voice	Concern for the environment
National Assessment Math	-0.107	-0.111	-0.103	-0.084	-0.096	-0.101	-0.095	-0.097	-0.099	-0.095	-0.100	-0.098
National Assessment ELA	-0.038	-0.074	-0.042	-0.019	-0.032	-0.017	-0.015	-0.015	-0.017	-0.013	-0.019	-0.015
Eighth-grade Math	0.023	0.007	-0.024	-0.002	-0.017	0.002	0.006	0.003	0.002	0.005	-0.001	0.004
Eighth-grade English	0.045	0.016	0.027	0.037	0.035	0.017	0.020	0.019	0.018	0.022	0.015	0.020
Eighth-grade Science	-0.132	-0.147	-0.131	-0.123	-0.123	-0.145	-0.141	-0.144	-0.146	-0.140	-0.144	-0.141
Economically disadvantaged	0.243	0.219	0.267	0.253	0.274	0.255	0.259	0.256	0.271	0.265	0.264	0.256
Took any AP exam in 2016	-0.349	-0.448	-0.454	-0.438	-0.450	-0.438	-0.440	-0.438	-0.438	-0.440	-0.447	-0.442
Female	0.185	0.161	0.151	0.154	0.147	0.161	0.153	0.160	0.160	0.160	0.157	0.161
Grade	-0.141	-0.117	-0.169	-0.129	-0.168	-0.133	-0.136	-0.135	-0.136	-0.135	-0.137	-0.134
Asian	0.300	0.254	0.238	0.266	0.257	0.254	0.251	0.255	0.264	0.255	0.252	0.254
Hispanic	-0.042	-0.018	-0.015	-0.007	-0.017	-0.004	-0.015	-0.004	-0.015	-0.015	-0.008	-0.016
Black	-0.332	-0.319	-0.317	-0.345	-0.312	-0.350	-0.350	-0.347	-0.350	-0.348	-0.346	-0.350
White	-0.126	-0.143	-0.136	-0.139	-0.150	-0.131	-0.114	-0.133	-0.122	-0.116	-0.119	-0.113
2016 average AP score	-0.205	-0.193	-0.199	-0.194	-0.201	-0.185	-0.187	-0.186	-0.186	-0.189	-0.186	-0.188
2016 % taking AP exam	0.687	0.708	0.704	0.713	0.705	0.714	0.711	0.714	0.711	0.713	0.713	0.713
Years teaching APES/APGOV	-0.530	-0.496	-0.512	-0.513	-0.509	-0.490	-0.494	-0.491	-0.493	-0.495	-0.489	-0.494

	Collaboration	Opportunities for leadership	Self-efficacy	Grit	Growth mindset	Appreciation for diversity	Civic/political efficacy	Whether expects to vote regularly at 18	Participatory citizenship	Interest in politics	Political voice	Concern for the environment
Course: APGOV	0.312	0.256	0.278	0.268	0.266	0.261	0.254	0.254	0.263	0.258	0.257	0.261
2015-16 average class size	-0.152	-0.183	-0.189	-0.176	-0.183	-0.179	-0.172	-0.178	-0.175	-0.174	-0.177	-0.175
% free/reduced-price lunch	0.150	0.145	0.147	0.139	0.144	0.140	0.141	0.141	0.141	0.142	0.143	0.141
Student-teacher ratio	-0.063	-0.060	-0.060	-0.056	-0.061	-0.055	-0.056	-0.056	-0.057	-0.057	-0.058	-0.056
% taking an AP exam	-0.195	-0.209	-0.214	-0.198	-0.204	-0.201	-0.200	-0.202	-0.201	-0.202	-0.196	-0.200
Average national Math	-0.107	-0.111	-0.120	-0.116	-0.117	-0.109	-0.106	-0.110	-0.111	-0.112	-0.110	-0.111
Average national ELA	-0.007	-0.019	-0.026	-0.021	-0.024	-0.016	-0.013	-0.017	-0.019	-0.018	-0.017	-0.017
Average eighth-grade Math	-0.086	-0.083	-0.086	-0.081	-0.085	-0.078	-0.076	-0.079	-0.080	-0.081	-0.076	-0.080
Average eighth-grade ELA	0.128	0.116	0.108	0.107	0.108	0.120	0.122	0.119	0.116	0.115	0.116	0.116
Average eighth-grade Science	-0.123	-0.126	-0.133	-0.127	-0.131	-0.117	-0.116	-0.118	-0.121	-0.124	-0.121	-0.123
Proportion school low SES	0.164	0.144	0.144	0.144	0.145	0.143	0.141	0.144	0.143	0.140	0.142	0.139
Proportion school taking 2016 AP exam	-0.173	-0.193	-0.203	-0.187	-0.194	-0.189	-0.184	-0.189	-0.187	-0.185	-0.184	-0.185
Proportion female	0.335	0.318	0.322	0.312	0.316	0.311	0.306	0.312	0.308	0.308	0.308	0.307
Average grade	-0.271	-0.281	-0.293	-0.279	-0.285	-0.285	-0.280	-0.284	-0.282	-0.275	-0.278	-0.276
Proportion Asian	0.514	0.502	0.504	0.502	0.502	0.504	0.508	0.503	0.507	0.502	0.501	0.503
Proportion Hispanic	-0.007	-0.005	-0.002	0.001	-0.003	-0.003	-0.005	-0.003	-0.002	0.000	0.000	0.000
Proportion Black	-0.137	-0.118	-0.119	-0.154	-0.127	-0.146	-0.147	-0.143	-0.144	-0.144	-0.150	-0.146
Proportion White	-0.160	-0.160	-0.163	-0.153	-0.159	-0.152	-0.150	-0.152	-0.155	-0.156	-0.153	-0.155

Table L6: P-values on baseline standardized mean differences between treatment and control students on all student-, teacher-, and school-level covariates, for all student survey analytic outcome samples in Year One

	Collaboration	Opportunities for leadership	Self-efficacy	Grit	Growth mindset	Appreciation for diversity	Civic/political efficacy	Whether expects to vote regularly at 18	Participatory citizenship	Interest in politics	Political voice	Concern for the environment
National Assessment Math	0.553	0.547	0.562	0.634	0.591	0.578	0.600	0.590	0.584	0.601	0.581	0.588
National Assessment ELA	0.826	0.679	0.812	0.915	0.857	0.926	0.931	0.934	0.923	0.940	0.914	0.931
Eighth-grade Math	0.893	0.968	0.892	0.990	0.923	0.992	0.973	0.987	0.991	0.976	0.996	0.981
Eighth-grade English	0.813	0.933	0.881	0.836	0.845	0.925	0.912	0.917	0.921	0.903	0.934	0.911
Eighth-grade Science	0.375	0.320	0.374	0.399	0.404	0.338	0.350	0.341	0.333	0.352	0.343	0.350
Economic disadvantage	0.281	0.330	0.233	0.252	0.219	0.251	0.247	0.250	0.231	0.257	0.244	0.259
Took any AP exam in 2016	0.353	0.218	0.211	0.230	0.217	0.229	0.227	0.228	0.229	0.227	0.220	0.226
Female	0.056	0.088	0.108	0.101	0.118	0.087	0.104	0.090	0.090	0.089	0.097	0.087
Grade	0.532	0.611	0.456	0.572	0.460	0.561	0.553	0.556	0.553	0.553	0.550	0.557
Asian	0.170	0.239	0.270	0.212	0.234	0.240	0.245	0.236	0.215	0.236	0.237	0.239
Hispanic	0.894	0.955	0.962	0.982	0.956	0.990	0.962	0.991	0.961	0.962	0.979	0.959
Black	0.348	0.367	0.366	0.331	0.372	0.331	0.330	0.334	0.329	0.333	0.336	0.330
White	0.646	0.616	0.637	0.626	0.600	0.646	0.690	0.641	0.672	0.686	0.679	0.693
2016 average AP score	0.330	0.354	0.344	0.353	0.337	0.374	0.368	0.370	0.371	0.363	0.370	0.366
2016 % taking any AP exam	0.003**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**

	Collaboration	Opportunities for leadership	Self-efficacy	Grit	Growth mindset	Appreciation for diversity	Civic/political efficacy	Whether expects to vote regularly at 18	Participatory citizenship	Interest in politics	Political voice	Concern for the environment
Years teaching APES/APGOV	0.061	0.079	0.069	0.067	0.070	0.080	0.078	0.079	0.079	0.078	0.080	0.079
Course: APGOV	0.484	0.562	0.531	0.545	0.548	0.555	0.566	0.566	0.552	0.559	0.561	0.556
2015-16 Avg Class Size	0.540	0.462	0.448	0.474	0.462	0.485	0.500	0.486	0.494	0.495	0.490	0.494
% free/reduced-price lunch	0.283	0.301	0.296	0.321	0.305	0.316	0.313	0.312	0.311	0.309	0.307	0.313
Student-teacher ratio	0.654	0.664	0.666	0.683	0.661	0.689	0.688	0.687	0.684	0.682	0.673	0.684
% taking an AP exam	0.486	0.448	0.436	0.471	0.458	0.463	0.464	0.460	0.463	0.461	0.472	0.464
Average national Math	0.624	0.602	0.576	0.588	0.586	0.606	0.614	0.603	0.601	0.598	0.605	0.602
Average national ELA	0.972	0.928	0.900	0.920	0.908	0.939	0.949	0.935	0.928	0.929	0.934	0.933
Average eighth-grade Math	0.685	0.692	0.686	0.699	0.686	0.708	0.712	0.702	0.701	0.695	0.715	0.701
Average eighth-grade ELA	0.623	0.656	0.678	0.679	0.677	0.641	0.634	0.643	0.652	0.657	0.652	0.655
Average eighth-grade Science	0.588	0.579	0.558	0.572	0.563	0.601	0.603	0.599	0.590	0.582	0.589	0.585
Proportion school low SES	0.408	0.476	0.471	0.473	0.471	0.476	0.483	0.474	0.478	0.487	0.482	0.488
% school taking 2016 AP exam	0.517	0.460	0.436	0.474	0.460	0.467	0.478	0.467	0.473	0.477	0.480	0.476
Proportion female	0.252	0.281	0.274	0.288	0.283	0.291	0.296	0.288	0.295	0.294	0.294	0.296
Average grade	0.479	0.463	0.443	0.463	0.457	0.454	0.461	0.454	0.459	0.469	0.464	0.469
Proportion Asian	0.128	0.132	0.133	0.131	0.133	0.129	0.125	0.130	0.127	0.130	0.129	0.129

	Collaboration	Opportunities for leadership	Self-efficacy	Grit	Growth mindset	Appreciation for diversity	Civic/political efficacy	Whether expects to vote regularly at 18	Participatory citizenship	Interest in politics	Political voice	Concern for the environment
Proportion Hispanic	0.972	0.982	0.991	0.998	0.988	0.989	0.979	0.989	0.994	0.999	0.998	0.999
Proportion Black	0.572	0.609	0.607	0.512	0.585	0.530	0.528	0.537	0.537	0.536	0.518	0.531
Proportion White	0.414	0.411	0.400	0.429	0.412	0.430	0.434	0.428	0.421	0.421	0.427	0.422

Figure L5: Baseline standardized mean difference between treatment and control students on all student-, teacher-, and school-level covariates, for all survey outcome analytic samples in Year One

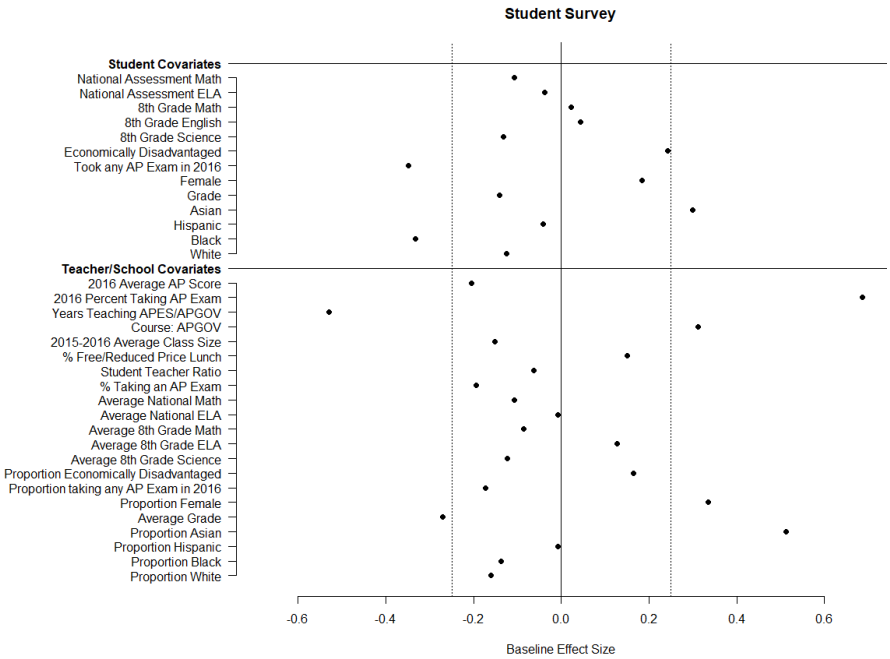


Table L7: Baseline standardized mean difference between treatment and control students on imputed student-level prior achievement covariates, for all (Year One) student survey outcome analytic samples

Collaboration

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.107	-0.455	-0.529	-0.033	-0.529
National Assessment ELA	-0.038	-0.204	-0.240	-0.003	-0.240
Eighth-grade Math	0.023	1.748	0.679	1.092	1.748
Eighth-grade English	0.045	3.014	1.141	1.917	3.014
Eighth-grade Science	-0.132	0.577	0.263	0.181	0.577

Opportunities for leadership

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.111	0.180	0.852	-0.782	0.852
National Assessment ELA	-0.074	0.484	1.776	-1.365	1.776
Eighth-grade Math	0.007	-2.854	-1.470	-1.377	-2.854
Eighth-grade English	0.016	-26.926	-10.530	-16.381	-26.926
Eighth-grade Science	-0.147	-0.629	-1.768	0.991	-1.768

Self-efficacy

	D1	D2	D3	D4	Most extreme

National Assessment Math	-0.103	0.185	-0.012	0.094	0.185
National Assessment ELA	-0.042	0.206	0.036	0.127	0.206
Eighth-grade Math	-0.024	0.608	0.166	0.417	0.608
Eighth-grade English	0.027	0.614	0.265	0.376	0.614
Eighth-grade Science	-0.131	0.071	0.048	-0.108	-0.131

Grit

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.084	-0.487	0.056	-0.628	-0.628
National Assessment ELA	-0.019	0.140	-0.074	0.195	0.195
Eighth-grade Math	-0.002	1.847	0.309	1.536	1.847
Eighth-grade English	0.037	1.980	0.526	1.491	1.980
Eighth-grade Science	-0.123	-0.111	-0.250	0.015	-0.250

Growth mindset

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.096	14.463	12.543	1.824	14.463
National Assessment ELA	-0.032	-2.483	-2.160	-0.355	-2.483
Eighth-grade Math	-0.017	2.812	1.375	1.420	2.812
Eighth-grade English	0.035	1.394	0.761	0.668	1.394
Eighth-grade Science	-0.123	1.011	0.567	0.321	1.011

Appreciation for diversity

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.101	-1.238	-1.570	0.231	-1.570
National Assessment ELA	-0.017	-0.205	-0.260	0.038	-0.260
Eighth-grade Math	0.002	0.054	0.144	-0.088	0.144
Eighth-grade English	0.017	-0.147	0.171	-0.301	-0.301
Eighth-grade Science	-0.145	-0.437	-0.282	-0.299	-0.437

Civic/political efficacy

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.095	0.808	0.425	0.289	0.808
National Assessment ELA	-0.015	-1.031	-0.599	-0.447	-1.031
Eighth-grade Math	0.006	-0.023	0.448	-0.465	-0.465
Eighth-grade English	0.020	-0.499	1.620	-2.099	-2.099
Eighth-grade Science	-0.141	1.967	1.654	0.172	1.967

Whether expects to vote regularly at 18

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.097	-0.194	-0.306	0.015	-0.306
National Assessment ELA	-0.015	-0.083	-0.161	0.064	-0.161
Eighth-grade Math	0.003	0.878	0.159	0.722	0.878
Eighth-grade English	0.019	0.647	0.225	0.441	0.647
Eighth-grade Science	-0.144	0.121	0.070	-0.093	-0.144

Participatory citizenship

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.099	5.420	4.084	1.237	5.420
National Assessment ELA	-0.017	-1.135	-0.864	-0.288	-1.135
Eighth-grade Math	0.002	-0.059	1.617	-1.673	-1.673
Eighth-grade English	0.018	-1.167	-3.364	2.214	-3.364
Eighth-grade Science	-0.146	-0.633	-0.381	-0.399	-0.633

Interest in politics

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.095	-0.272	0.445	-0.812	-0.812
National Assessment ELA	-0.013	-0.095	0.235	-0.344	-0.344
Eighth-grade Math	0.005	-26.175	1.997	-28.167	-28.167
Eighth-grade English	0.022	1.068	-0.011	1.101	1.101
Eighth-grade Science	-0.140	0.730	0.262	0.328	0.730

Political voice

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.100	-3.953	13.406	-17.459	-17.459
National Assessment ELA	-0.019	-0.089	0.225	-0.333	-0.333
Eighth-grade Math	-0.001	-3.328	-0.156	-3.173	-3.328
Eighth-grade English	0.015	-12.467	-0.930	-11.522	-12.467
Eighth-grade Science	-0.144	2.453	1.327	0.983	2.453

Concern for the environment

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.098	-18.453	-11.532	-7.019	-18.453
National Assessment ELA	-0.015	-0.635	-0.401	-0.249	-0.635
Eighth-grade Math	0.004	4.073	-0.090	4.166	4.166
Eighth-grade English	0.020	1.001	0.103	0.918	1.001
Eighth-grade Science	-0.141	0.158	0.030	-0.012	0.158

Course relevance for the future

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.105	0.064	-0.136	0.096	-0.136
National Assessment ELA	-0.025	0.160	-0.060	0.195	0.195
Eighth-grade Math	-0.012	-0.384	-0.043	-0.353	-0.384
Eighth-grade English	0.011	-0.747	0.009	-0.744	-0.747
Eighth-grade Science	-0.133	0.011	-0.117	-0.005	-0.133

Course satisfaction

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.113	-0.100	0.019	-0.233	-0.233
National Assessment ELA	-0.027	-0.014	0.111	-0.152	-0.152
Eighth-grade Math	-0.013	-0.503	0.905	-1.421	-1.421

Eighth-grade English	0.011	-0.095	0.766	-0.850	-0.850
Eighth-grade Science	-0.133	0.651	0.908	-0.390	0.908

Student engagement

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.110	0.116	0.153	-0.148	0.153
National Assessment ELA	-0.047	0.188	0.227	-0.086	0.227
Eighth-grade Math	-0.017	-0.987	-0.170	-0.834	-0.987
Eighth-grade English	-0.002	-1.146	-0.146	-1.002	-1.146
Eighth-grade Science	-0.142	-0.188	-0.201	-0.128	-0.201

Appendix M: Research Question One Student-Level Baseline Equivalence for AP Analytic Outcome Samples, Subgroups

In the Appendix M tables, we show standardized mean differences between treatment and control students on all student-, teacher-, and school-level covariates for outcomes as specified within subgroups: APGOV, APES, students from lower-income households, and students from higher-income households. We include as a covariate the 2015-16 AP score average corresponding specifically to that outcome, as opposed to the average AP score and exam-taking rate for all outcomes (as we had done previously). For this reason, there are blank cells for covariates describing 2015-16 AP performance. Student-level prior achievement variables assume missing at randomness (MAR) in Tables M1, M3, M5, and M7, and bound according to potential deviations from MAR in Tables M2, M4, M6, and M8.

Table M1: Baseline standardized mean differences between treatment and control students on all student-, teacher-, and school-level covariates for Year One AP outcome analytic samples within the APGOV subgroup

	AP QS (full sample) and exam-taking (n=1,693)	AP QS (exam-takers only scores (n=1,587)	AP continuous scores (n=505)
National Assessment Math	0.162	0.183	0.187
National Assessment ELA	0.180	0.186	0.166
Eighth-grade Math	0.188	0.244	0.268
Eighth-grade English	0.271	0.328	0.275
Eighth-grade Science	0.027	0.046	-0.065
Economically disadvantaged	-0.138	-0.232	-0.129
Took any AP exam in 2016	0.658	0.683	0.850
Female	-0.057	-0.057	-0.058
Grade	0.179	0.161	0.335
Asian	-0.125	-0.128	-0.042
Hispanic	0.039	-0.059	0.250
Black	0.197	0.223	0.265
White	0.116	0.120	0.175
2016 average AP score		-0.438	
2016 % earning AP QS (exam-takers)		-0.481	
2016 % earning AP QS (full sample)	-0.434		
2016 % taking AP exam	0.107		

2016 Average total fine-grained score			-0.291
2016 Average MC fine-grained score			-0.179
2016 average FR fine-grained score			-0.336
Years teaching APES/APGOV	-0.266	-0.265	-0.397
Course: APGOV	NA	NA	NA
2015-16 average class size	-0.091	-0.057	0.007
% free/reduced-price lunch	0.209	0.247	0.279
Student-teacher ratio	-0.128	-0.166	0.030
% taking an AP exam	-0.290	-0.341	-0.125
Average national Math	0.169	0.170	-0.025
Average national ELA	0.050	0.039	-0.031
Average eighth-grade Math	0.159	0.179	0.178
Average eighth-grade ELA	0.386	0.405	0.132
Average eighth-grade Science	-0.056	-0.061	-0.097
Proportion school low SES	-0.076	-0.075	-0.026
Proportion school taking any 2016 AP exam	0.061	0.034	0.271
Proportion female	-0.437	-0.464	-0.531
Average grade	-0.038	-0.063	-0.354
Proportion Asian	-0.308	-0.306	0.093
Proportion Hispanic	-0.022	-0.031	0.064
Proportion Black	0.247	0.284	0.583
Proportion White	0.040	0.028	-0.552

Table M2: Baseline standardized mean differences between treatment and control students on imputed student-level covariates for Year One AP outcome analytic samples within the APGOV subgroup

Took AP exam

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.162	1.427	1.790	-0.202	1.790
National Assessment ELA	0.180	1.397	1.747	-0.170	1.747
Eighth-grade Math	0.188	0.979	1.274	-0.107	1.274
Eighth-grade English	0.271	1.178	1.374	0.075	1.374

Eighth-grade Science	0.027	0.142	-0.017	0.186	0.186

AP qualifying score (exam-takers only)

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.183	0.088	-0.014	0.285	0.285
National Assessment ELA	0.186	0.115	0.039	0.263	0.263
Eighth-grade Math	0.244	0.104	0.058	0.290	0.290
Eighth-grade English	0.328	0.181	0.125	0.384	0.384
Eighth-grade Science	0.046	0.035	0.057	0.024	0.057

AP qualifying score (full sample)

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.162	0.111	0.039	0.233	0.233
National Assessment ELA	0.180	0.141	0.086	0.234	0.234
Eighth-grade Math	0.188	0.103	0.101	0.191	0.191
Eighth-grade English	0.271	0.184	0.153	0.302	0.302
Eighth-grade Science	0.027	0.024	0.043	0.008	0.043

AP total score

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.187	0.212	0.300	0.099	0.300
National Assessment ELA	0.166	0.183	0.240	0.109	0.240
Eighth-grade Math	0.268	0.313	0.432	0.149	0.432
Eighth-grade English	0.275	0.328	0.473	0.130	0.473
Eighth-grade Science	-0.065	-0.086	-0.044	-0.108	-0.108

AP multiple-choice score

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.187	0.191	0.262	0.116	0.262
National Assessment ELA	0.166	0.169	0.214	0.121	0.214
Eighth-grade Math	0.268	0.300	0.410	0.159	0.410
Eighth-grade English	0.275	0.314	0.447	0.142	0.447
Eighth-grade Science	-0.065	-0.083	-0.041	-0.108	-0.108

AP free-response score

	D1	D2	D3	D4	Most extreme

National Assessment Math	0.187	0.232	0.342	0.077	0.342
National Assessment ELA	0.166	0.198	0.277	0.087	0.277
Eighth-grade Math	0.268	0.324	0.457	0.135	0.457
Eighth-grade English	0.275	0.339	0.498	0.117	0.498
Eighth-grade Science	-0.065	-0.092	-0.046	-0.111	-0.111

Table M3: Baseline standardized mean differences between treatment and control students on all student-, teacher-, and school-level covariates for Year One AP outcome analytic samples within the APES subgroup

	AP QS (full sample) and exam-taking (n=1,952)	AP QS (exam-takers only) (n=1,376)	AP continuous scores (n=1,094)
National Assessment Math	-0.077	-0.183	-0.162
National Assessment ELA	-0.001	-0.138	-0.213
Eighth-grade Math	0.062	-0.032	-0.172
Eighth-grade English	0.010	-0.037	-0.134
Eighth-grade Science	-0.005	-0.095	-0.132
Economically disadvantaged	0.049	0.112	0.188
Took any AP exam in 2016	-0.237	-0.361	-0.166
Female	0.084	0.072	0.022
Grade	-0.213	-0.145	-0.071
Asian	0.304	0.097	0.038
Hispanic	-0.149	-0.072	-0.176
Black	-0.205	-0.114	-0.119
White	0.255	0.203	0.360
2016 average AP score		-0.031	
2016 % earning AP QS (exam-takers)		-0.037	
2016 % earning AP QS (full sample)	-0.031		
2016 % taking AP exam	0.251		
2016 Average total fine-grained score			0.130
2016 Average MC fine-grained score			0.128
2016 average FR fine-grained score			0.133
Years teaching APES/APGOV	-0.294	-0.215	0.050
Course: APGOV	NA	NA	NA

2015-16 average class size	0.001	-0.151	-0.292
% free/reduced-price lunch	-0.302	-0.285	-0.477
Student-teacher ratio	-0.053	-0.098	-0.074
% taking an AP exam	0.283	0.260	0.244
Average national Math	0.070	0.099	0.310
Average national ELA	0.109	0.121	0.177
Average eighth-grade Math	0.081	0.047	0.229
Average eighth-grade ELA	0.069	0.034	0.074
Average eighth-grade Science	-0.053	-0.059	0.101
Proportion school low SES	0.266	0.319	0.448
Proportion school taking any 2016 AP exam	0.402	0.327	0.500
Proportion female	0.519	0.508	0.346
Average grade	0.132	0.047	0.412
Proportion Asian	0.513	0.601	0.580
Proportion Hispanic	-0.143	-0.117	-0.377
Proportion Black	-0.356	-0.512	-0.507
Proportion White	0.024	-0.008	0.360

Table M4: Baseline standardized mean differences between treatment and control students on imputed student-level covariates, for Year One AP outcome analytic samples within the APES subgroup

Took AP exam

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.077	0.251	-0.192	0.366	0.366
National Assessment ELA	-0.001	0.302	-0.108	0.408	0.408
Eighth-grade Math	0.062	0.289	-0.036	0.388	0.388
Eighth-grade English	0.010	0.191	-0.069	0.270	0.270
Eighth-grade Science	-0.005	0.248	-0.183	0.425	0.425

AP qualifying score (full sample)

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.077	-0.051	-0.103	-0.025	-0.103

National Assessment ELA	-0.001	0.022	-0.025	0.045	0.045
Eighth-grade Math	0.062	0.104	0.085	0.081	0.104
Eighth-grade English	0.010	0.051	0.032	0.029	0.051
Eighth-grade Science	-0.005	0.004	0.011	-0.013	-0.013

AP qualifying score (exam-takers only)

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.183	-0.160	-0.196	-0.147	-0.196
National Assessment ELA	-0.138	-0.119	-0.149	-0.108	-0.149
Eighth-grade Math	-0.032	0.004	-0.011	-0.017	-0.032
Eighth-grade English	-0.037	0.003	-0.017	-0.017	-0.037
Eighth-grade Science	-0.095	-0.105	-0.071	-0.130	-0.130

AP total score

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.162	-0.161	-0.166	-0.157	-0.166
National Assessment ELA	-0.213	-0.213	-0.216	-0.210	-0.216
Eighth-grade Math	-0.172	-0.144	-0.139	-0.177	-0.177
Eighth-grade English	-0.134	-0.109	-0.108	-0.135	-0.135
Eighth-grade Science	-0.132	-0.126	-0.111	-0.147	-0.147

AP multiple-choice score

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.162	-0.156	-0.160	-0.158	-0.162
National Assessment ELA	-0.213	-0.209	-0.212	-0.211	-0.213
Eighth-grade Math	-0.172	-0.115	-0.158	-0.129	-0.172
Eighth-grade English	-0.134	-0.085	-0.125	-0.095	-0.134
Eighth-grade Science	-0.132	-0.121	-0.114	-0.139	-0.139

AP free-response score

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.162	-0.167	-0.175	-0.155	-0.175
National Assessment ELA	-0.213	-0.218	-0.224	-0.208	-0.224

Eighth-grade Math	-0.172	-0.185	-0.108	-0.249	-0.249
Eighth-grade English	-0.134	-0.144	-0.079	-0.199	-0.199
Eighth-grade Science	-0.132	-0.132	-0.102	-0.161	-0.161

Table M5: Baseline standardized mean differences between treatment and control students on all student-, teacher-, and school-level covariates for Year One AP outcome analytic samples within the lower-income household student subgroup

	AP QS (full sample) and exam-taking (n=1,159)	AP QS (exam-takers only) (n=1,125)	AP continuous scores (n=805)
National Assessment Math	0.069	0.000	0.023
National Assessment ELA	0.144	0.007	-0.128
Eighth-grade Math	0.161	0.124	-0.011
Eighth-grade English	0.177	0.160	-0.020
Eighth-grade Science	0.065	-0.015	-0.068
Economically disadvantaged	NA	NA	NA
Took any AP exam in 2016	0.235	0.133	0.075
Female	0.047	0.006	-0.083
Grade	-0.073	-0.048	0.011
Asian	0.210	0.013	0.156
Hispanic	-0.144	-0.128	-0.236
Black	-0.116	-0.030	-0.064
White	0.279	0.192	0.321
2016 average AP score		-0.025	
2016 % earning AP QS (exam-takers)		0.002	
2016 % earning AP QS (full sample)	0.028		
2016 % taking AP exam	0.274		
2016 Average total fine-grained score			0.056
2016 Average MC fine-grained score			0.085
2016 average FR fine-grained score			0.044
Years teaching APES/APGOV	-0.297	-0.230	-0.059
Course: APGOV	0.315	0.121	0.458
2015-16 average class size	-0.005	-0.094	-0.230
% free/reduced-price lunch	-0.141	-0.112	-0.266

Student-teacher ratio	-0.067	-0.089	-0.136
% taking an AP exam	0.506	0.429	0.412
Average national Math	0.088	0.120	0.184
Average national ELA	0.057	0.071	0.017
Average eighth-grade Math	0.019	-0.007	0.083
Average eighth-grade ELA	0.012	-0.007	-0.061
Average eighth-grade Science	-0.146	-0.159	-0.102
Proportion school low SES	0.166	0.177	0.365
Proportion school taking any 2016 AP exam	0.577	0.503	0.552
Proportion female	0.068	-0.026	-0.109
Average grade	0.236	0.206	0.134
Proportion Asian	0.319	0.336	0.413
Proportion Hispanic	-0.021	0.019	-0.107
Proportion Black	-0.248	-0.310	-0.309
Proportion White	0.002	-0.025	0.128

Table M6: Baseline standardized mean differences between treatment and control students on imputed student-level covariates for Year One AP outcome analytic samples within the lower-income household student subgroup

Took AP exam

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.069	0.711	0.311	0.470	0.711
National Assessment ELA	0.144	0.769	0.380	0.533	0.769
Eighth-grade Math	0.161	0.811	0.383	0.589	0.811
Eighth-grade English	0.177	0.765	0.423	0.519	0.765
Eighth-grade Science	0.065	0.318	-0.097	0.481	0.481

AP (full sample)

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.069	0.068	-0.015	0.152	0.152
National Assessment ELA	0.144	0.143	0.074	0.214	0.214

Eighth-grade Math	0.161	0.242	0.142	0.261	0.261
Eighth-grade English	0.177	0.270	0.138	0.310	0.310
Eighth-grade Science	0.065	0.082	0.083	0.065	0.083

AP qualifying score (exam-takers only)

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.000	-0.030	-0.103	0.072	-0.103
National Assessment ELA	0.007	-0.016	-0.075	0.065	-0.075
Eighth-grade Math	0.124	0.234	0.108	0.250	0.250
Eighth-grade English	0.160	0.266	0.123	0.303	0.303
Eighth-grade Science	-0.015	0.005	0.013	-0.022	-0.022

AP total score

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.023	-0.037	-0.018	0.003	-0.037
National Assessment ELA	-0.128	-0.169	-0.155	-0.142	-0.169
Eighth-grade Math	-0.011	0.018	-0.012	0.019	0.019
Eighth-grade English	-0.020	0.014	-0.030	0.024	-0.030
Eighth-grade Science	-0.068	-0.066	-0.052	-0.082	-0.082

AP multiple-choice score

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.023	-0.057	-0.024	-0.010	-0.057
National Assessment ELA	-0.128	-0.180	-0.158	-0.150	-0.180
Eighth-grade Math	-0.011	0.049	-0.016	0.054	0.054
Eighth-grade English	-0.020	0.037	-0.033	0.050	0.050
Eighth-grade Science	-0.068	-0.054	-0.049	-0.073	-0.073

AP free-response score

	D1	D2	D3	D4	Most extreme
National Assessment Math	0.023	-0.028	-0.019	0.013	-0.028

National Assessment ELA	-0.128	-0.168	-0.160	-0.136	-0.168
Eighth-grade Math	-0.011	-0.015	0.005	-0.031	-0.031
Eighth-grade English	-0.020	-0.012	-0.016	-0.016	-0.020
Eighth-grade Science	-0.068	-0.078	-0.050	-0.096	-0.096

Table M7: Baseline standardized mean differences between treatment and control students on all student-, teacher-, and school-level covariates for Year One AP outcome analytic samples within the higher-income household student subgroup

	AP QS (full sample) and exam-taking (n=2,086)	AP QS (exam-takers only) (n=1,838)	AP continuous scores (n=794)
National Assessment Math	-0.018	-0.060	-0.072
National Assessment ELA	0.105	0.056	0.050
Eighth-grade Math	0.096	0.039	-0.015
Eighth-grade English	0.122	0.121	0.047
Eighth-grade Science	-0.002	-0.039	-0.053
Economically disadvantaged	NA	NA	NA
Took any AP exam in 2016	0.041	0.030	0.210
Female	-0.008	-0.039	-0.054
Grade	-0.112	-0.087	0.012
Asian	0.073	0.027	-0.034
Hispanic	-0.175	-0.183	-0.230
Black	0.100	0.159	-0.023
White	0.152	0.147	0.420
2016 average AP score		-0.421	
2016 % earning AP QS (exam-takers)		-0.475	
2016 % earning AP QS (full sample)	-0.391		
2016 % taking AP exam	0.019		
2016 Average total fine-grained score		-0.128	-0.148
2016 Average MC fine-grained score		0.113	-0.106
2016 average FR fine-grained score		-0.154	-0.168
Years teaching APES/APGOV	-0.128	0.168	0.251
Course: APGOV	0.188	-0.182	0.181

2015-16 average class size	-0.090	-0.459	-0.110
% free/reduced-price lunch	0.111	0.059	0.066
Student-teacher ratio	-0.142	-0.021	-0.103
% taking an AP exam	-0.421	0.107	-0.219
Average national Math	0.062	0.237	0.099
Average national ELA	0.003	-0.040	0.054
Average eighth-grade Math	0.114	0.008	0.177
Average eighth-grade ELA	0.239	-0.084	0.134
Average eighth-grade Science	-0.015	-0.011	0.035
Proportion school low SES	-0.013	-0.271	0.133
Proportion school taking any 2016 AP exam	-0.040	-0.030	0.243
Proportion female	0.059	-0.073	0.081
Average grade	-0.202	0.238	-0.141
Proportion Asian	-0.036	-0.011	0.344
Proportion Hispanic	-0.090		-0.248
Proportion Black	0.219		0.290
Proportion White	0.027		-0.080

Table M8: Baseline standardized mean differences between treatment and control students on imputed student-level covariates for Year One AP outcome analytic samples within the higher-income household student subgroup

Took AP exam

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.018	0.099	0.046	0.034	0.099
National Assessment ELA	0.105	0.214	0.165	0.154	0.214
Eighth-grade Math	0.096	0.293	0.558	-0.168	0.558
Eighth-grade English	0.122	0.404	0.489	0.037	0.489
Eighth-grade Science	-0.002	0.200	-0.053	0.251	0.251

AP qualifying score (full sample)

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.018	-0.051	-0.079	0.010	-0.079
National Assessment ELA	0.105	0.078	0.054	0.128	0.128
Eighth-grade Math	0.096	-0.015	-0.005	0.086	0.096
Eighth-grade English	0.122	0.003	0.022	0.103	0.122
Eighth-grade Science	-0.002	-0.021	0.003	-0.026	-0.026

AP qualifying score (exam-takers only)

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.060	-0.101	-0.125	-0.035	-0.125
National Assessment ELA	0.056	0.023	0.003	0.076	0.076
Eighth-grade Math	0.039	-0.101	-0.101	0.039	-0.101
Eighth-grade English	0.121	-0.031	-0.010	0.100	0.121
Eighth-grade Science	-0.039	-0.071	-0.036	-0.074	-0.074

AP total score

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.072	-0.070	-0.043	-0.099	-0.099
National Assessment ELA	0.050	0.051	0.073	0.028	0.073
Eighth-grade Math	-0.015	-0.055	0.058	-0.128	-0.128
Eighth-grade English	0.047	0.001	0.132	-0.084	0.132
Eighth-grade Science	-0.053	-0.077	-0.053	-0.078	-0.078

AP multiple-choice score

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.072	-0.075	-0.051	-0.095	-0.095
National Assessment ELA	0.050	0.047	0.067	0.031	0.067
Eighth-grade Math	-0.015	-0.048	0.033	-0.095	-0.095
Eighth-grade English	0.047	0.008	0.104	-0.049	0.104
Eighth-grade Science	-0.053	-0.073	-0.057	-0.069	-0.073

AP free-response score

	D1	D2	D3	D4	Most extreme
National Assessment Math	-0.072	-0.068	-0.037	-0.102	-0.102
National Assessment ELA	0.050	0.053	0.079	0.024	0.079
Eighth-grade Math	-0.015	-0.078	0.090	-0.183	-0.183
Eighth-grade English	0.047	-0.023	0.164	-0.140	0.164
Eighth-grade Science	-0.053	-0.090	-0.050	-0.093	-0.093

Appendix N: Year One Impact Sensitivity Results

This section explores the sensitivity of our model to various modeling choices.

Sensitivity to Two-Level HLM

We compared the sensitivity to the choice of our primary two-level HLM by also providing results fit from a three-level HLM (nesting students within schools within districts), and an ordinary least squares linear model, ignoring the grouping by school and only including district as a fixed effect. We compared the following models:

1) Three-level HLM (students nested within schools nested within districts: school and district random effects)
2) Two-level HLM (students nested within schools: school random, district fixed effect)
3) One-level LM (no random effects, district as fixed effect)

To isolate changes due to random versus fixed effects, we fit all models with the same covariates as selected above for our primary model.

We provide these results as a sensitivity check, so neither the three-level results nor the one-level results should be interpreted as fully accurate. Most three-level models suffer from a singular fit, indicating a possible underestimation of the true standard error. The one-level linear models ignore clustering by school; hence, the resulting standard errors are underestimates of the truth, which artificially inflates statistical significance.

Table N1: Sensitivity of two-level HLM Year One covariate-adjusted impact estimates to modeling choice

	3-level HLM: school and district random	2-level HLM: school random; district fixed	1-level LM: no school; district fixed
Took AP exam	-0.031 (0.16)	-0.009 (0.15)	0 (0.08)
AP qualifying score (full sample)	0.207 (0.11)	0.264 (0.11)*	0.267 (0.09)**
AP qualifying score (exam-takers)	0.327 (0.15)*	0.457 (0.14)**[S]	0.468 (0.13)***
AP total score	0.185 (0.07)**	0.192 (0.07)**	0.156 (0.04)***
AP multiple-choice score	0.179 (0.07)*	0.188 (0.07)**	0.147 (0.04)***
AP free-response score	0.174 (0.07)*	0.181 (0.07)*	0.154 (0.04)***
CWRA+ overall score	0.182 (0.17)[S]	0.202 (0.2)	0.171 (0.13)
CWRA+ performance task subscore	0.284 (0.19)[S]	0.268 (0.22)	0.169 (0.1)
CWRA+ selected response subscore	0.025 (0.14)[S]	0.048 (0.16)	0.016 (0.1)
Collaboration	-0.014 (0.12)[S]	0.015 (0.12)[S]	0.015 (0.12)
Opportunities for leadership	0.14 (0.12)[S]	0.118 (0.13)[S]	0.11 (0.12)
Self-efficacy	-0.209 (0.14)[S]	-0.208 (0.15)	-0.215 (0.11)
Grit	-0.122 (0.14)[S]	-0.099 (0.16)	-0.108 (0.12)

Growth mindset	0.036 (0.11)[S]	0.038 (0.11)[S]	0.038 (0.11)
Appreciation for diversity	0.201 (0.13)[S]	0.181 (0.13)	0.18 (0.12)
Civic/political efficacy	-0.081 (0.13)[S]	-0.131 (0.15)	-0.161 (0.12)
Whether expects to vote at 18	-0.134 (0.11)[S]	-0.126 (0.12)[S]	-0.126 (0.12)
Participatory citizenship	0.019 (0.13)[S]	0.025 (0.15)	-0.01 (0.13)
Interest in politics	-0.204 (0.14)[S]	-0.251 (0.15)	-0.281 (0.12)*
Political voice	-0.07 (0.14)[S]	-0.077 (0.14)	-0.109 (0.12)
Concern for the environment	0.078 (0.11)[S]	0.128 (0.11)[S]	0.127 (0.11)
Course relevance for the future	0.11 (0.14)[S]	0.144 (0.15)	0.104 (0.11)
Course satisfaction	0.071 (0.16)	0.102 (0.16)	0.015 (0.11)
Student engagement	-0.006 (0.15)[S]	-0.005 (0.16)	-0.076 (0.11)

Notes: Table columns show standardized effect sizes, standard errors, confidence intervals, p-values, and analytic sample sizes. Asterisks denote statistical significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

Sensitivity to Covariates

We also investigated the sensitivity of our primary results to modeling choices around which covariates are included. Here we compare models with the following sets of covariates:

1) No covariates
2) Covariates with baseline imbalance
3) Covariates with baseline imbalance or selected by the automated variable selection [PRIMARY]
4) All covariates

For models 2 and 3, per WWC we define "baseline imbalance" as an absolute baseline effect size of greater than 0.05.

Table N2: Sensitivity of Year One overall impact estimates to covariates.

	No covariates	Covariates with ABE > 0.05	Primary	All covariates
Took AP exam	0.125 (0.18)	-0.007 (0.15)	-0.009 (0.15)	0.056 (0.16)
AP QS (full sample)	-0.005 (0.21)	0.16 (0.13)	0.264 (0.11)*	0.351 (0.13)**[S]
AP QS (exam-takers)	-0.043 (0.21)	0.145 (0.13)	0.457 (0.14)**[S]	0.353 (0.14)*[S]
AP total score	0.035 (0.17)	0.196 (0.07)**	0.192 (0.07)**	0.188 (0.08)*
AP multiple-choice score	0.049 (0.17)	0.196 (0.07)**	0.188 (0.07)**	0.17 (0.08)*
AP free-response score	0.016 (0.16)	0.185 (0.07)**	0.181 (0.07)*	0.191 (0.07)*
CWRA+ overall score	-0.07 (0.18)	0.202 (0.2)	0.202 (0.2)	0.149 (0.22)
CWRA+ PT subscore	0.005 (0.19)	0.28 (0.22)	0.268 (0.22)	0.323 (0.24)
CWRA+ SR subscore	-0.086 (0.16)	0.048 (0.16)	0.048 (0.16)	-0.025 (0.17)
Collaboration	-0.123 (0.09)	0.015 (0.12)[S]	0.015 (0.12)[S]	0.013 (0.12)[S]
Opportunities for leadership	0.043 (0.09)	0.118 (0.13)[S]	0.118 (0.13)[S]	0.121 (0.14)[S]
Self-efficacy	-0.279 (0.11)*	-0.208 (0.15)	-0.208 (0.15)	-0.211 (0.15)
Grit	-0.096 (0.1)	-0.099 (0.16)	-0.099 (0.16)	-0.107 (0.17)
Growth mindset	-0.147 (0.08)	0.038 (0.11)[S]	0.038 (0.11)[S]	0.04 (0.12)[S]
Appreciation for diversity	0.009 (0.09)	0.181 (0.13)	0.181 (0.13)	0.168 (0.15)
Civic/political efficacy	-0.154 (0.08)	-0.131 (0.15)	-0.131 (0.15)	-0.148 (0.15)
Whether expects to vote at 18	-0.176 (0.09)	-0.068 (0.14)	-0.126 (0.12)[S]	-0.123 (0.12)[S]
Participatory citizenship	-0.09 (0.09)	0.086 (0.16)	0.025 (0.15)	0.037 (0.15)
Interest in politics	-0.253 (0.09)**	-0.251 (0.15)	-0.251 (0.15)	-0.261 (0.15)
Political voice	-0.207 (0.08)*	-0.077 (0.14)	-0.077 (0.14)	-0.09 (0.14)
Concern for the environment	0.022 (0.09)	0.128 (0.11)[S]	0.128 (0.11)[S]	0.114 (0.12)[S]
Course relevance for the future	0.058 (0.1)	0.144 (0.15)	0.144 (0.15)	0.135 (0.15)
Course satisfaction	-0.13 (0.11)	0.102 (0.16)	0.102 (0.16)	0.104 (0.16)
Student engagement	-0.119 (0.1)	-0.005 (0.16)	-0.005 (0.16)	-0.003 (0.16)

Notes: Table columns show standardized effect sizes and standard errors. Asterisks denote statistical significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

Appendix O: Student Sorting in Experimental Schools with Both Experimental and Non-Participating Teachers of the Same Course

In schools with both consented KIA teachers and non-participating teachers of the same course (e.g., a treatment APGOV teacher and a non-participating APGOV teacher), we can examine average characteristics of students enrolled in experimental and non-experimental conditions. Analysis of these average characteristics in 2015-16 (baseline year), 2016-17 (Year One school sample), and 2017-18 (Year Two school sample) can shed light on the extent to which the average characteristics of students enrolled in KIA versus non-participating teachers' classrooms may have changed over time.

In the baseline year, 10 of 62 experimental schools (16%) fit this description, with both experimental and non-participating teachers offering the same target course in the same year. In the Year One sample, 13 of 68 (19%) did, and in the Year Two sample 9 of 50 (18%) did. Though a minority of schools in the overall study, concern of potential non-random sorting of students into KIA versus non-KIA classrooms motivated this investigation. (See applicable student counts in the "Sample" section below.) Systematic changes in measured covariates for students enrolled in treatment classrooms, compared to non-participating classrooms, might suggest observed and unobserved bias. For example, if over time, KIA APGOV classrooms had more higher-performing students than non-participating APGOV classrooms, this pattern might suggest that students with certain characteristics were sorted into KIA classrooms. While we can statistically control for observed bias attributable to this behavior, we are concerned about the possibility of unobserved bias, particularly in a direction favoring students of experimental treatment teachers.

Sample

Within some experimental schools, there were non-participating teachers who did not enroll in the KIA RCT but did teach APGOV or APES in the same year. In 2015-16, immediately prior to the first year of the KIA offer for treatment teachers, 10 schools that ultimately enrolled in the KIA study had both experimental and non-participating teachers. In Year One, five schools had both an implementing (treatment) teacher and a non-participant, while eight schools had both a not-yet-implementing (control) teacher and a non-participating teacher of the same course. In Year Two, four treatment and five control schools had both a teacher implementing KIA and a non-participating teacher. We show counts in Table O1.

Table O1: School and teacher counts of baseline, Year One, and Year Two school samples with experimental and non-participating teachers of the same course

	Baseline 2015-16	Year One 2016-17 (treatment teachers had KIA offer)	Year Two 2017-18 (treatment and control teachers had KIA offer)
Schools with at least one experimental and one non-experimental teacher of the same course	10	13	9
Teachers within schools			
Experimental treatment	4	5	4
Experimental control	7	9	5
Non-participating	15	25	14

Within these schools, we show counts of students in Table O2.

Table O2: Counts of students within baseline, Year One, and Year Two school samples with experimental and non-participating teachers of the same course

	Baseline	Year One	Year Two
Whole School	2,088	2,709	1,814
Treatment students	326	311	255
Control students	616	674	424
Non-participating students	1,146	1,724	1,135

Below, we show all participating schools in each district in each year, shaded to show the combination of treatment and non-participating, or control and non-participating, teachers in the baseline (Figure O1), Year One (Figure O2), and Year Two (Figure O3) school samples. These figures highlight that virtually all schools (9 of 10) with both experimental and non-participating teachers of the same course were clustered in District D, which offered only APGOV. This concentration of schools with a combination of experimental and non-participating teachers in District D persisted in each year of the study (Figures O2 and O3).

Figure O1: Schools with experimental and non-participating teachers of the same course, 2015-16

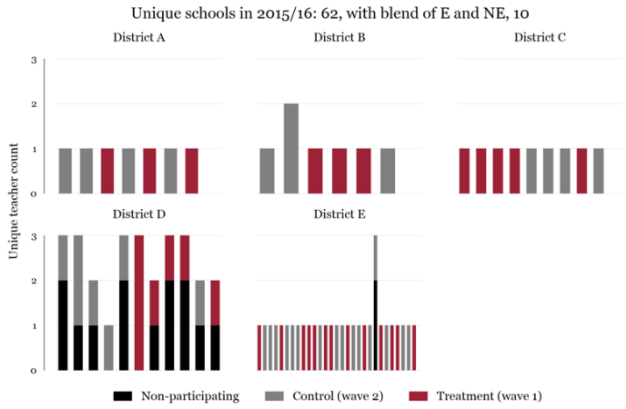


Figure O2: Schools with experimental and non-participating teachers of the same course, 2016-17

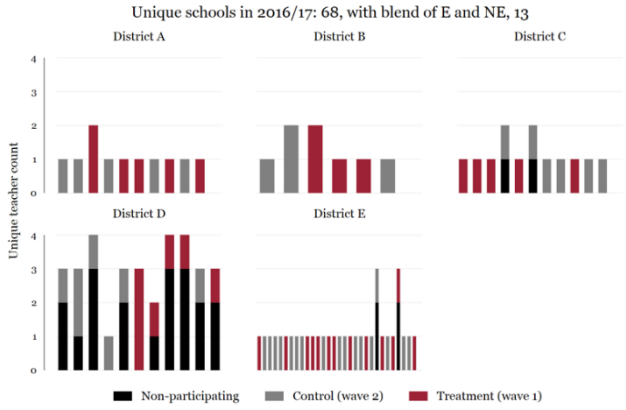
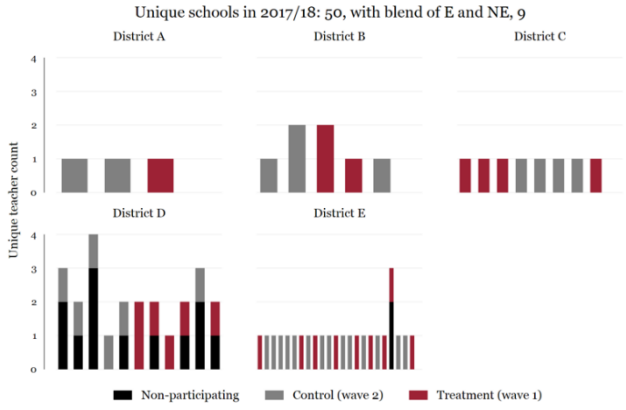


Figure O3: Schools with experimental and non-participating teachers of the same course, 2017-18



Analytic Approach

We expect to see fluctuation in the composition of students enrolled in experimental and non-participating teachers' classrooms from year to year. We also expect principals might non-randomly

sort students to teachers from year to year for reasons unassociated with teachers' participation (or not) in KIA. Relevant to our Year One impact estimates, we would be concerned if we observed systematic sorting between 2015-16 and 2016-17 of higher-performing and/or more advantaged (e.g., higher SES) students to treatment experimental teachers and did not see the same among control. If principals in treatment schools were sorting higher-performing students to treatment classrooms while control principals were not doing the same during the experimental year (2016-17), our estimated differences in AP performance between treatment and control students could be due, at least in part, to the principals' sorting rather than due to the KIA intervention. On the other hand, evidence of this type of sorting from 2015-16 to 2016-17 for both treatment teachers (with the KIA offer in 2016-17) and control teachers (business-as-usual in 2016-17), could suggest positive sorting associated with experimental teachers, but not with the KIA intervention (as control teachers were not implementing in 2016-17). If the same positive sorting took place in both treatment and control classrooms, the potential concern about overestimated treatment effects due to positive sorting would be lessened.

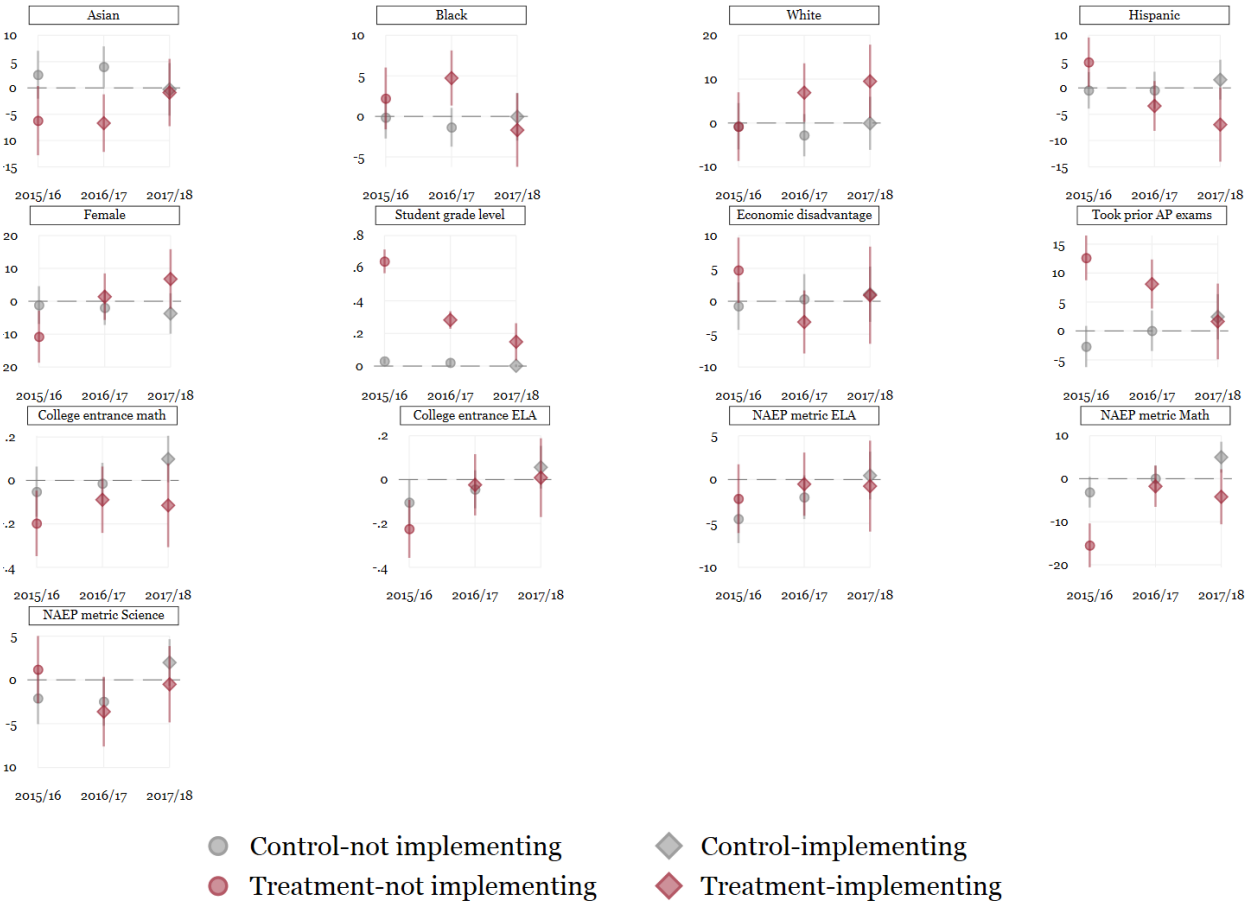
For Year Two impact estimates, we should be concerned about positive sorting from 2016-17 to 2017-18 among treatment but not control teachers. In this year, both groups had the KIA offer.

To find evidence of sorting from baseline to 2016-17 and/or 2017-18, first we limited our respective baseline, Year One, and Year Two samples to include only experimental schools with both experimental and non-experimental teachers of APGOV and APES within the same school and year (which were shared above). We then fit a series of saturated linear probability models (LPM), conditioning on treatment status for each pre-treatment covariates, with school-level fixed effects. We included school fixed effects to partial out mean differences across schools (and districts), thus estimating within-school mean differences on each covariate between treatment and non-participating teachers. Unlike for our baseline equivalence and impact analyses, we used unimputed data as we did not impute data for students of non-participating teachers within experimental schools.

Results

Figure O4 shows the average within-school differences between experimental and non-experimental teachers' students across all schools at three points in time, separately for each covariate of interest. The Y-axis shows the three points in time (i.e., 2015-16, 2016-17, and 2017-18) and the X-axis shows the magnitude of the differences, the parameter estimates. Point estimates show the standardized mean differences between experimental and non-experimental teachers from the LPM model described above. Whiskers represent 95% confidence intervals. Red represents differences between treatment and non-participating teachers' students' covariate averages while gray represents differences between control and non-participating teachers' students. We use circles to represent differences prior to initiation of the first year of the KIA offer (i.e., 2016-17 for treatment and 2017-18 for control) and diamonds to represent differences during and after the initiation of the KIA offer.

Figure O4: Standardized differences between experimental and non-participating teachers' students' average covariate values in the subsamples of baseline, Year One, and Year Two schools with experimental and non-participating teachers of the same course



Year One

As noted above, our primary concern about sorting from the baseline to Year One is whether there was non-random sorting of students to treatment teachers that did not occur among control teachers (i.e., unsynchronized). Based on College Board data describing qualifying score rates by race and socio-economic status (e.g., College Board, 2018), we see sorting favoring treatment classrooms on covariates describing whether students, on average, were White (positive direction), Hispanic (negative direction), and economically disadvantaged (negative direction). We do not see the same pattern for control. Related to students' prior achievement, patterns are synchronized. Between 2015-16 and 2016-17, treatment and control students had higher scores relative to non-participating on four of five prior achievement tests excepting eighth-grade Science.[14] Thus, the preliminary exploration provides mixed evidence supporting asynchronous positive sorting of students to Year One treatment

[14] Keeping in mind that whether student took at least one AP exam the prior year is correlated with grade level, these covariates also show evidence of negative systemic sorting favoring control.

classrooms. Assuming prior achievement scores supersede direct racial influences, we do not see evidence of asynchronous positive sorting to treatment.

Year Two

Our concern for Year Two impact analyses is positive sorting among treatment teachers' students', but not control, from 2016-17 to 2017-18. Among treatment teachers, we see sorting favoring positive treatment bias from 2016-17 and 2017-18 on all covariates describing race, though negative on economic disadvantage. College-entry Math and ELA, and eighth-grade ELA were quite constant year to year, while eighth-grade treatment Math was lower and eighth-grade Science higher for both treatment and control. Like for the Efficacy year, results are thus mixed. If the importance of prior achievement scores superseded direct racial influences, for Year Two as well, we do not see evidence of asynchronous positive sorting to treatment.

Limitations

There are limitations to this exploration. First, the samples of schools with both experimental and non-experimental teachers of the same course changed over time, both in overall counts (i.e., 10 to 13 to 9 across the three years) and in terms of which schools were included within those counts. Therefore, year-to-year differences may simply reflect changes in school composition rather than demonstrate systematic patterns. In addition, the samples we focus on represent less than one-fifth of our baseline, Year One, and Year Two school samples. Within the school sample, there were only five of 35 treatment teachers in Year One and four of 23 in Year Two, so any potential bias affects only a fraction of the full sample. Another point of consideration is how the majority of the schools with both experimental and non-experimental teachers were concentrated in District D: This district only participated in the KIA RCT with APGOV teachers, their students were considerably higher SES compared to other districts' students, and we do not have access to continuous APGOV/APES score outcomes for this district. It is different in these observed ways from the other four districts and likely differs in unobserved ways as well. We further address the differences between District D and the others in our discussion of subgroup results. A final limitation is, unlike our impact analyses, our sorting analysis results reflect the sample for which we have non-missing data.

Appendix P: Student Enrollment Sorting in KIA and Matched Non-Volunteering Teachers' Classrooms Over Time

In Appendix O, we described our examination of the extent to which the average characteristics of students enrolled in KIA may have changed over time compared with those in non-participating teachers' classrooms of the same course within the same school. Across the baseline (2015-16), Year One (2016-17), and Year Two (2017-18), a total of 32 schools included both experimental and non-experimental teachers of the same course in the same year. Changes in measured covariates for students enrolled in treatment classrooms compared to non-participating classrooms, not synchronously observed when comparing experimental control to non-participating classrooms, might have suggested bias that could have affected the internal validity of experimental estimates. We were particularly concerned about the possibility of unobserved bias in a direction favoring students of experimental treatment teachers. We did not find evidence of this type of sorting posing a threat to the internal validity experimental analyses conducted to address our first three research questions.

Relevant to those questions, we conducted a similar analysis to investigate the possibility of systematic sorting into KIA RCT teachers' APGOV or APES classrooms over time, potentially driven by students, teachers, and/or administrators' growing understanding of the KIA program. We were again concerned about the possibility of systematic sorting of higher-performing students into experimental treatment but not control classrooms. For this analysis, we included in the sample RCT teachers and matched non-volunteering teachers (n=118 teachers across 102 schools).

Sample

The sample includes all (non-joining) students in each randomized teacher's APGOV or APES classrooms and matched non-experimental teacher's APGOV or APES classrooms. The counts of students, teachers, and schools vary because not all teachers taught eligible courses in each year.

Table P1: Counts of students, teachers, and schools in Baseline, Year One, and Year Two base samples

		Baseline 2015-16	Year One 2016-17 (treatment had KIA offer)	Year Two 2017-18 (treatment and control had KIA offer)
Students	Non-Experimental	1,596	2,303	2,213
	Control	1,698	2,144	1,754
	Treatment	1,457	1,496	1,180
Teachers	Non-Experimental	29	44	44
	Control	35	39	30
	Treatment	31	35	23
Schools	Non-Experimental	27	38	38
	Control	33	37	29
	Treatment	29	31	21

Analytic Approach

We expect to see year-to-year fluctuation in the composition of students enrolled in experimental and non-participating teachers' classrooms. We also expect principals might non-randomly sort students to teachers from year to year for reasons unassociated with teachers' participation (or not) in KIA. Relevant to our Year One impact estimates, we would be concerned if we observed systematic sorting between 2015-16 and 2016-17 of higher-performing and/or more advantaged (e.g., higher household income) students to treatment experimental teachers yet did not see the same among control. If principals in treatment schools were systematically sorting higher-performing students to treatment classrooms while control principals were not doing the same during the experimental year (2016-17), our estimated differences in AP performance between treatment and control students could be due, at least in part, to this rather than due to the KIA intervention. On the other hand, evidence of this type of sorting from 2015-16 to 2016-17 for both treatment teachers (with the KIA offer in 2016-17) and control teachers (business-as-usual in 2016-17) could suggest positive sorting associated with experimental teachers but not with the KIA intervention (since control teachers were not implementing in 2016-17). If the same positive sorting took place in both treatment and control classrooms, the potential concern about overestimated treatment effects due to positive sorting would be lessened.

For Year Two, we should be concerned about positive sorting from 2016-17 to 2017-18 among treatment but not control teachers. In this year, both groups had the KIA offer.

To see whether there is evidence of such sorting, we included all randomized teachers and non-experimental teachers who met the matching criteria for the non-experimental analysis.[15] We next fit a series of linear probability models (LPM), conditioning on treatment status for each pre-treatment covariates, with district-level fixed effects. We included district fixed effects to partial out mean differences across districts, thus estimating within-district mean differences on each covariate between randomized and non-participating teachers. Unlike for our baseline equivalence and impact analyses, we used unimputed data as we did not impute data for students of non-participating teachers within experimental schools.

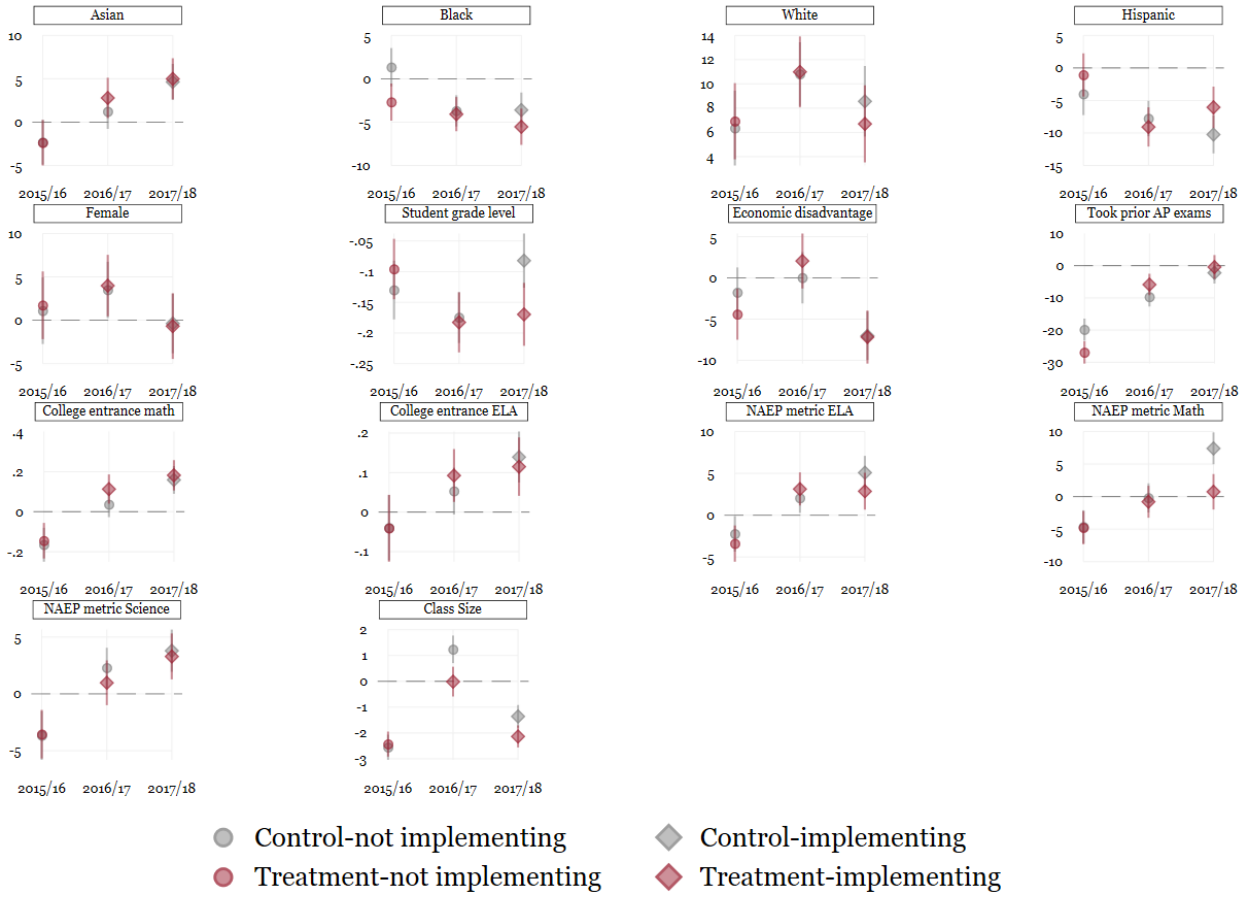
Results

Figure P1 shows the average within-school differences between experimental and non-experimental teachers' students across all schools at three points in time, separately for each covariate of interest. The Y-axis shows the three points in time (i.e., 2015-16, 2016-17, and 2017-18) and the X-axis shows the magnitude of the differences, the parameter estimates. Point estimates show the standardized mean differences between experimental and non-experimental teachers from the LPM model described above. Whiskers represent 95% confidence intervals. Red represents differences between treatment and non-participating teachers' students' covariate averages while gray represents differences between control and non-participating teachers' students. We use circles to represent

[15] The one unmatched treatment teacher is included in the present analysis. Treatment and control teachers each had a weight of "1."

differences prior to initiation of the first year of the KIA offer (i.e., 2016-17 for treatment and 2017-18 for control) and diamonds to represent differences during and after the initiation of the KIA offer.

Figure P1: Standardized differences between experimental and non-participating teachers' students' average covariate values for students in randomized and (Round Two) matched non-experimental classrooms



Year One

As noted above, our primary concern about sorting from the baseline to Year One is whether there was non-random sorting of students to treatment teachers that did not occur among control and non-experimental teachers (i.e., unsynchronized). Based on College Board data describing AP exam qualifying score rates by race and socio-economic status (e.g., College Board, 2018), we see synchronous sorting between treatment and control groups on covariates describing whether students, on average, were White (positive direction), Asian (positive direction), Hispanic (negative direction), Black (negative direction), and economically disadvantaged (positive direction; i.e., increase proportion). Related to students' prior achievement, patterns are also synchronized. Between 2015-16 and 2016-17, relative to non-participating, treatment and control students had higher scores relative to non-participating on all five prior achievement measures as well as whether students took any AP exam in May 2016. Thus, our preliminary exploration provides no evidence of positive sorting to treatment classrooms unobserved in control.

Year Two

Our concern for the Year Two impact analyses is positive sorting among treatment but not control teachers' students' from 2016-17 to 2017-18. Among treatment teachers, we see sorting favoring positive treatment bias from 2016-17 and 2017-18 on all covariates describing race, though negative on economic disadvantage. Like with Year One, we see treatment and control classrooms following the same synchronous pattern and no evidence of positive sorting to treatment classrooms unobserved in control.

Limitations

As before, there are limits to this exploration. First, the samples of teachers changed over time, both in overall counts (i.e., 95 to 118 to 103 across the three years) and in terms of which schools were included within those counts. Therefore year-to-year differences may simply reflect changes in school composition rather than demonstrate systematic patterns. In addition, unlike our impact analyses, our sorting analysis results reflect the sample for which we have non-missing data.

Appendix Q: National Prior Achievement Scores Among Experimental Exam-Taker and Non-Exam-Taker Students

The top row of Table Q1 shows students' national Math and ELA scores across courses, by treatment and control, for the full student samples of Research Question One 2016-17, and Research Question Two 2016-17 and 2017-18. The second and third rows, respectively, show the same breakdowns but by course, for APES followed by APGOV. Whereas across courses, scores are slightly above the national average, APES scores (both treatment and control and across both years) are below average and APGOV are well above the national average.

Table Q1: Baseline national Math and ELA scores for the full samples of Research Question One 2016-17, and Research Question Two 2016-17 and 2017-18 students who did and did not take the APGOV/APES examination in the relevant outcome year, by treatment status, across courses, and by course

	Year One (n=3,645 students of 74 teachers)		Year One (n=3,100 students of 53 teachers)		Year Two (n=2,946 students of 53 teachers)	
All students	Treatment 2016-17	Control 2016-17	Treatment 2016-17	Control 2016-17	Treatment 2017-18	Control 2017-18
Across courses	Math: 0.109 ELA: 0.039	Math: 0.060 ELA: 0.034	Math: 0.224 ELA: 0.151	Math: 0.118 ELA: 0.115	Math: 0.084 ELA: 0.108	Math: 0.080 ELA: 0.151
APES	Math: -0.177 ELA: -0.275	Math: -0.285 ELA: -0.321	Math: -0.079 ELA: -0.177	Math: -0.252 ELA: -0.246	Math: -0.123 ELA: -0.108	Math: -0.249 ELA: -0.097
APGOV	Math: 0.416 ELA: 0.376	Math: 0.493 ELA: 0.478	Math: 0.600 ELA: 0.558	Math: 0.521 ELA: 0.508	Math: 0.370 ELA: 0.405	Math: 0.539 ELA: 0.497
Took AP exam	Treatment 2016-17	Control 2016-17	Treatment 2016-17	Control 2016-17	Treatment 2017-18	Control 2017-18
Across courses	Math: 0.217 ELA: 0.133	Math: 0.210 ELA: 0.194	Math: 0.345 ELA: 0.262	Math: 0.261 ELA: 0.261	Math: 0.166 ELA: 0.179	Math: 0.209 ELA: 0.279
APES	Math: -0.066 ELA: -0.194	Math: -0.150 ELA: -0.170	Math: 0.040 ELA: -0.080	Math: -0.114 ELA: -0.098	Math: -0.024 ELA: -0.056	Math: -0.132 ELA: 0.029
APGOV	Math: 0.468 ELA: 0.422	Math: 0.522 ELA: 0.509	Math: 0.657 ELA: 0.613	Math: 0.546 ELA: 0.534	Math: 0.409 ELA: 0.479	Math: 0.568 ELA: 0.542
Did not take AP exam	Treatment 2016-17	Control 2016-17	Treatment 2016-17	Control 2016-17	Treatment 2017-18	Control 2017-18

Across courses	Math: -0.414	Math: -0.556	Math: -0.354	Math: -0.539	Math: -0.314	Math: -0.314
	ELA: -0.415	ELA: -0.627	ELA: -0.382	ELA: -0.555	ELA: -0.238	ELA: -0.239
APES	Math: -0.513	Math: -0.558	Math: -0.447	Math: -0.545	Math: -0.531	Math: -0.479
	ELA: -0.520	ELA: -0.628	ELA: -0.474	ELA: -0.559	ELA: -0.324	ELA: -0.345
APGOV	Math: -0.115	Math: -0.524	Math: -0.017	Math: -0.456	Math: 0.120	Math: 0.326
	ELA: -0.098	ELA: -0.606	ELA: -0.047	ELA: -0.500	ELA: -0.067	ELA: 0.172

Appendix R: Research Question One Subgroup Impact Results

Within Course Subgroups

Table R1: Covariate-adjusted standardized effect sizes for all Year One outcomes, comparing APGOV treatment and control students

	Effect Size (SE)	95% Confidence Interval	p-value	n
Took AP exam	-0.4 (0.23)	(-0.86, 0.06)	0.088	1693
AP qualifying score (full sample)	0.227 (0.15)	(-0.064, 0.518)	0.13	1693
AP qualifying score (exam-takers only)	0.506 (0.17)**[S]	(0.163, 0.849)	0.004	1587
AP total score	0.403 (0.12)***	(0.174, 0.633)	0.00058	505
AP multiple-choice score	0.403 (0.12)***	(0.171, 0.635)	0.00067	505
AP free-response score	0.37 (0.12)**	(0.131, 0.61)	0.0025	505
CWRA+ overall score	0.193 (0.3)	(-0.399, 0.786)	0.52	184
CWRA+ performance task subscore	0.357 (0.34)	(-0.314, 1.027)	0.3	201
CWRA+ selected response subscore	-0.156 (0.24)	(-0.625, 0.313)	0.52	208
Collaboration	-0.059 (0.17)[S]	(-0.394, 0.277)	0.73	286
Opportunities for leadership	0.242 (0.18)[S]	(-0.106, 0.59)	0.17	299
Self-efficacy	-0.343 (0.2)	(-0.739, 0.053)	0.089	298
Grit	-0.157 (0.22)	(-0.587, 0.273)	0.47	299
Growth mindset	0.017 (0.16)[S]	(-0.301, 0.335)	0.92	300
Appreciation for diversity	0.029 (0.19)	(-0.337, 0.394)	0.88	298
Civic/political efficacy	-0.109 (0.2)	(-0.5, 0.282)	0.58	299
Whether expects to vote regularly at 18	-0.001 (0.16)[S]	(-0.323, 0.321)	0.99	298
Participatory citizenship	0.044 (0.2)	(-0.353, 0.442)	0.83	298
Interest in politics	-0.067 (0.21)	(-0.479, 0.344)	0.75	299
Political voice	-0.079 (0.2)	(-0.478, 0.319)	0.7	297
Concern for the environment	0.099 (0.16)[S]	(-0.213, 0.41)	0.53	298
Course relevance for the future	-0.04 (0.2)	(-0.425, 0.345)	0.84	298
Course satisfaction	-0.039 (0.22)	(-0.47, 0.392)	0.86	300
Student engagement	-0.188 (0.21)	(-0.595, 0.22)	0.37	299

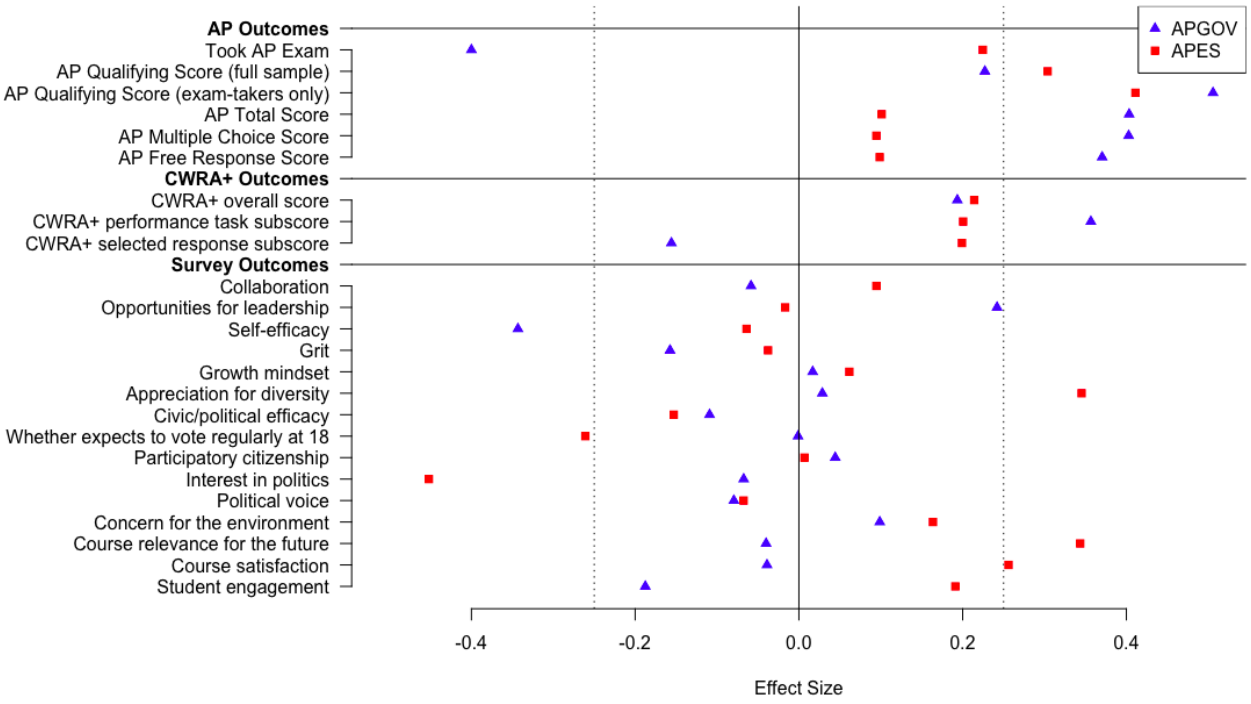
Notes: Table columns show standardized effect sizes, standard errors, confidence intervals, p-values, and analytic sample sizes. Asterisks denote statistical significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

Table R2: Covariate-adjusted standardized effect sizes for all Year One outcomes, comparing APES treatment and control students

	Effect Size (SE)	95% Confidence Interval	p-value	n
Took AP exam	0.225 (0.18)	(-0.132, 0.581)	0.22	1952
AP qualifying score (full sample)	0.304 (0.15)*	(0.003, 0.604)	0.048	1952
AP qualifying score (exam-takers only)	0.411 (0.18)*[S]	(0.056, 0.766)	0.024	1376
AP total score	0.101 (0.08)	(-0.058, 0.261)	0.21	1094
AP multiple-choice score	0.095 (0.08)	(-0.067, 0.256)	0.25	1094
AP free-response score	0.099 (0.08)	(-0.066, 0.263)	0.24	1094
CWRA+ overall score	0.214 (0.27)	(-0.319, 0.747)	0.43	305
CWRA+ performance task subscore	0.201 (0.31)	(-0.405, 0.806)	0.52	364
CWRA+ selected response subscore	0.199 (0.21)	(-0.208, 0.606)	0.34	326
Collaboration	0.095 (0.17)[S]	(-0.246, 0.435)	0.59	458
Opportunities for leadership	-0.017 (0.19)[S]	(-0.387, 0.353)	0.93	475
Self-efficacy	-0.064 (0.21)	(-0.484, 0.356)	0.77	479
Grit	-0.038 (0.24)	(-0.514, 0.438)	0.88	480
Growth mindset	0.062 (0.16)[S]	(-0.26, 0.383)	0.71	480
Appreciation for diversity	0.345 (0.2)	(-0.039, 0.73)	0.078	476
Civic/political efficacy	-0.153 (0.23)	(-0.596, 0.29)	0.5	474
Whether expects to vote regularly at 18	-0.261 (0.18)[S]	(-0.614, 0.092)	0.15	476
Participatory citizenship	0.007 (0.22)	(-0.433, 0.447)	0.98	474
Interest in politics	-0.452 (0.23)	(-0.904, 0)	0.05	476
Political voice	-0.068 (0.21)	(-0.487, 0.351)	0.75	474
Concern for the environment	0.164 (0.17)[S]	(-0.168, 0.495)	0.33	476
Course relevance for the future	0.344 (0.21)	(-0.06, 0.747)	0.095	473
Course satisfaction	0.256 (0.23)	(-0.203, 0.715)	0.27	480
Student engagement	0.191 (0.22)	(-0.246, 0.628)	0.39	476

Notes: Table columns show standardized effect sizes, standard errors, confidence intervals, p-values, and analytic sample sizes. Asterisks denote statistical significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

Figure R1: Covariate-adjusted standardized effect sizes for all Year One outcomes, comparing APGOV treatment and control students, and APES treatment and control students



Within student household income groups

Table R3: Covariate-adjusted standardized effect sizes for all Year One outcomes, comparing higher-income household treatment and control students

	Effect Size (SE)	95% Confidence Interval	p-value	n
Took AP exam	-0.002 (0.17)	(-0.326, 0.322)	0.99	1559
AP qualifying score (full sample)	0.319 (0.12)**	(0.087, 0.552)	0.0072	1559
AP qualifying score (exam-takers only)	0.496 (0.16)**[S]	(0.186, 0.806)	0.0019	1125
AP total score	0.173 (0.08)*	(0.017, 0.33)	0.03	805
AP multiple-choice score	0.183 (0.08)*	(0.024, 0.342)	0.024	805
AP free-response score	0.145 (0.08)	(-0.018, 0.309)	0.082	805
CWRA+ overall score	0.243 (0.21)	(-0.169, 0.655)	0.25	193
CWRA+ performance task subscore	0.321 (0.23)	(-0.123, 0.765)	0.16	233
CWRA+ selected response subscore	0.094 (0.17)	(-0.241, 0.43)	0.58	219
Collaboration	0.01 (0.14)[S]	(-0.257, 0.277)	0.94	324
Opportunities for leadership	0.1 (0.14)[S]	(-0.183, 0.383)	0.49	337
Self-efficacy	-0.199 (0.16)	(-0.508, 0.11)	0.21	338
Grit	-0.09 (0.17)	(-0.423, 0.243)	0.6	340
Growth mindset	0.044 (0.13)[S]	(-0.207, 0.296)	0.73	339

Appreciation for diversity	0.251 (0.15)	(-0.04, 0.542)	0.091	337
Civic/political efficacy	-0.088 (0.16)	(-0.397, 0.222)	0.58	336
Whether expects to vote regularly at 18	-0.078 (0.13)[S]	(-0.341, 0.185)	0.56	336
Participatory citizenship	0.075 (0.16)	(-0.241, 0.391)	0.64	335
Interest in politics	-0.249 (0.17)	(-0.575, 0.078)	0.14	338
Political voice	-0.068 (0.16)	(-0.376, 0.239)	0.66	337
Concern for the environment	0.135 (0.13)[S]	(-0.112, 0.382)	0.28	338
Course relevance for the future	0.121 (0.16)	(-0.188, 0.429)	0.44	337
Course satisfaction	0.16 (0.18)	(-0.184, 0.504)	0.36	340
Student engagement	-0.01 (0.17)	(-0.34, 0.319)	0.95	338

Notes: Table columns show standardized effect sizes, standard errors, confidence intervals, p-values, and analytic sample sizes. Asterisks denote statistical significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

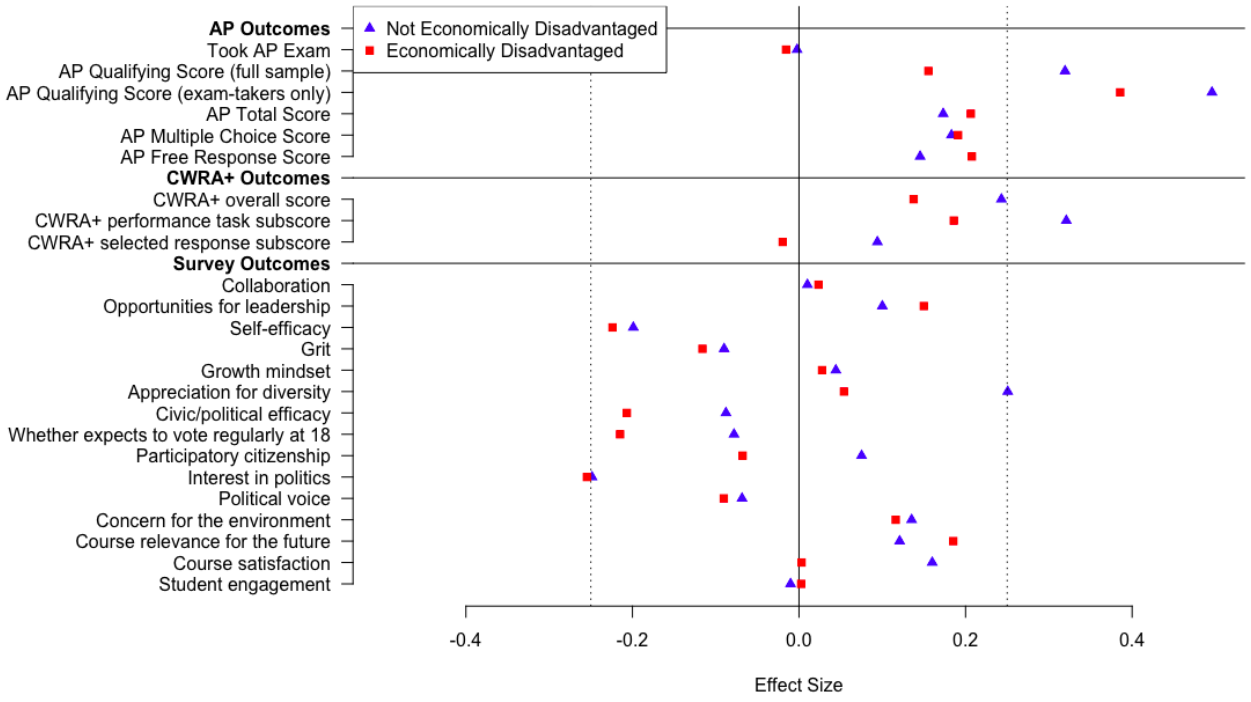
Table R4: Covariate-adjusted standardized effect sizes for all Year One outcomes, comparing lower-income household treatment and control students

	Effect Size (SE)	95% Confidence Interval	p-value	n
Took AP exam	-0.016 (0.16)	(-0.334, 0.303)	0.92	2086
AP qualifying score (full sample)	0.155 (0.14)	(-0.123, 0.434)	0.27	2086
AP qualifying score (exam-takers only)	0.386 (0.18)*[S]	(0.041, 0.73)	0.028	1838
AP total score	0.206 (0.08)**	(0.058, 0.355)	0.0064	794
AP multiple-choice score	0.191 (0.08)*	(0.041, 0.341)	0.013	794
AP free-response score	0.208 (0.08)**	(0.053, 0.362)	0.0085	794
CWRA+ overall score	0.138 (0.23)	(-0.315, 0.59)	0.55	296
CWRA+ performance task subscore	0.186 (0.24)	(-0.279, 0.651)	0.43	332
CWRA+ selected response subscore	-0.02 (0.19)	(-0.386, 0.347)	0.92	315
Collaboration	0.023 (0.16)[S]	(-0.291, 0.338)	0.88	420
Opportunities for leadership	0.15 (0.17)[S]	(-0.181, 0.481)	0.37	437
Self-efficacy	-0.224 (0.18)	(-0.578, 0.13)	0.21	439
Grit	-0.116 (0.2)	(-0.499, 0.267)	0.55	439
Growth mindset	0.028 (0.16)[S]	(-0.277, 0.333)	0.86	441
Appreciation for diversity	0.054 (0.18)	(-0.29, 0.398)	0.76	437
Civic/political efficacy	-0.207 (0.19)	(-0.573, 0.159)	0.27	437
Whether expects to vote regularly at 18	-0.215 (0.16)[S]	(-0.532, 0.103)	0.18	438
Participatory citizenship	-0.068 (0.19)	(-0.442, 0.306)	0.72	437
Interest in politics	-0.255 (0.19)	(-0.627, 0.118)	0.18	437
Political voice	-0.09 (0.18)	(-0.447, 0.266)	0.62	434

Concern for the environment	0.116 (0.16)[S]	(-0.188, 0.421)	0.45	436
Course relevance for the future	0.185 (0.18)	(-0.166, 0.537)	0.3	434
Course satisfaction	0.003 (0.2)	(-0.381, 0.387)	0.99	440
Student engagement	0.003 (0.19)	(-0.368, 0.373)	0.99	437

Notes: Table columns show standardized effect sizes, standard errors, confidence intervals, p-values, and analytic sample sizes. Asterisks denote statistical significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

Figure R2: Covariate-adjusted standardized effect sizes for all Year One outcomes, comparing higher-income household treatment and control students, and between lower-income household treatment and control students



Within higher- and lower-income household districts

Districts serving mostly students from higher-income households

Table R5: Covariate-adjusted standardized effect sizes for all Year One outcomes, comparing treatment and control students within districts serving mostly higher-income households

	Effect Size (SE)	95% Confidence Interval	p-value	n
Took AP exam	-0.299 (0.31)	(-0.91, 0.313)	0.34	1437
AP qualifying score (full sample)	0.214 (0.15)	(-0.078, 0.507)	0.15	1437
AP qualifying score (exam-takers only)	0.473 (0.18)**[S]	(0.129, 0.818)	0.0072	1391
AP total score	0.183 (0.24)	(-0.281, 0.647)	0.44	331
AP multiple-choice score	0.241 (0.24)	(-0.229, 0.712)	0.31	331

AP free-response score	0.14 (0.24)	(-0.329, 0.608)	0.56	331
CWRA+ overall score	0.186 (0.32)	(-0.44, 0.812)	0.56	180
CWRA+ performance task subscore	0.085 (0.37)	(-0.633, 0.802)	0.82	195
CWRA+ selected response subscore	-0.129 (0.26)	(-0.635, 0.378)	0.62	200
Collaboration	-0.041 (0.19)[S]	(-0.409, 0.327)	0.83	258
Opportunities for leadership	0.264 (0.19)[S]	(-0.116, 0.644)	0.17	267
Self-efficacy	-0.393 (0.23)	(-0.836, 0.049)	0.082	266
Grit	-0.186 (0.25)	(-0.668, 0.295)	0.45	267
Growth mindset	-0.016 (0.18)[S]	(-0.363, 0.331)	0.93	268
Appreciation for diversity	0.053 (0.21)	(-0.353, 0.46)	0.8	266
Civic/political efficacy	-0.121 (0.22)	(-0.561, 0.32)	0.59	267
Whether expects to vote regularly at 18	-0.005 (0.18)[S]	(-0.355, 0.346)	0.98	267
Participatory citizenship	0.045 (0.22)	(-0.394, 0.483)	0.84	266
Interest in politics	-0.124 (0.24)	(-0.593, 0.346)	0.61	267
Political voice	-0.025 (0.23)	(-0.468, 0.418)	0.91	265
Concern for the environment	0.163 (0.18)[S]	(-0.181, 0.508)	0.35	266
Course relevance for the future	-0.176 (0.21)	(-0.596, 0.244)	0.41	266
Course satisfaction	-0.291 (0.24)	(-0.752, 0.171)	0.22	268
Student engagement	-0.293 (0.23)	(-0.746, 0.161)	0.21	267

Notes: Table columns show standardized effect sizes, standard errors, confidence intervals, p-values, and analytic sample sizes. Asterisks denote statistical significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

Districts serving mostly students from lower-income households

Table R6: Covariate-adjusted standardized effect sizes for all Year One outcomes, comparing treatment and control students within districts serving mostly lower-income households

	Effect Size (SE)	95% Confidence Interval	p-value	n
Took AP exam	0.077 (0.17)	(-0.265, 0.42)	0.66	2208
AP qualifying score (full sample)	0.317 (0.15)*	(0.014, 0.619)	0.041	2208
AP qualifying score (exam-takers only)	0.445 (0.19)*[S]	(0.075, 0.815)	0.019	1572
AP total score	0.194 (0.08)*	(0.041, 0.347)	0.013	1268
AP multiple-choice score	0.181 (0.08)*	(0.026, 0.336)	0.022	1268
AP free-response score	0.188 (0.08)*	(0.03, 0.345)	0.02	1268
CWRA+ overall score	0.22 (0.29)	(-0.343, 0.783)	0.44	309
CWRA+ performance task subscore	0.416 (0.33)	(-0.228, 1.06)	0.21	370
CWRA+ selected response subscore	0.18 (0.22)	(-0.251, 0.611)	0.41	334
Collaboration	0.072 (0.18)[S]	(-0.288, 0.432)	0.7	486
Opportunities for leadership	-0.028 (0.2)[S]	(-0.419, 0.363)	0.89	507

Self-efficacy	-0.03 (0.22)	(-0.47, 0.41)	0.89	511
Grit	-0.015 (0.25)	(-0.509, 0.48)	0.95	512
Growth mindset	0.092 (0.17)[S]	(-0.244, 0.428)	0.59	512
Appreciation for diversity	0.305 (0.21)	(-0.1, 0.71)	0.14	508
Civic/political efficacy	-0.139 (0.23)	(-0.599, 0.322)	0.55	506
Whether expects to vote regularly at 18	-0.248 (0.19)[S]	(-0.617, 0.121)	0.19	507
Participatory citizenship	0.008 (0.24)	(-0.46, 0.476)	0.97	506
Interest in politics	-0.375 (0.25)	(-0.864, 0.113)	0.13	508
Political voice	-0.121 (0.23)	(-0.565, 0.322)	0.59	506
Concern for the environment	0.097 (0.18)[S]	(-0.25, 0.444)	0.58	508
Course relevance for the future	0.449 (0.21)*	(0.037, 0.861)	0.033	505
Course satisfaction	0.465 (0.23)*	(0.009, 0.92)	0.046	512
Student engagement	0.267 (0.23)	(-0.185, 0.719)	0.25	508

Notes: Table columns show standardized effect sizes, standard errors, confidence intervals, p-values, and analytic sample sizes. Asterisks denote statistical significance: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

Appendix S: Summary of College and Work Readiness Data Collection, Sample, and Results

One of the hypothesized strengths of PBL, compared to transmission instruction, is that students should learn more deeply and can transfer skills and knowledge to other contexts. We sought to measure ability to transfer skills in critical thinking, problem-solving, and writing mechanics to other contexts.

Outcome measures

In Year One, we administered the College and Work Readiness Assessment (CWRA+), designed to measure students' critical thinking and written communication skills. The CWRA+ is composed of three measures we used as outcomes. While the first and second are independent of each other, the third is a composite of the two:

- The performance task measures students' skills in analysis, problem-solving, and writing mechanics, as well as their writing effectiveness.
- Selected response questions measure students' skills in analysis and problem-solving, including their ability to reason scientifically, read and evaluate critically, and critique an argument.
- The CWRA+ total score combines the performance task and selected response scores.

Sample

To estimate the impact of the KIA offer on students' skills in critical thinking, we selected one class section for each teacher in which they would administer the CWRA+. Selecting one class section was intended to help participating teachers who had multiple KIA classrooms by reducing the burden in matters such as obtaining parental consent, reserving computer space, and administering the assessment.

More than half the analytic sample of teachers taught one section of APGOV or APES, and for these teachers that section was automatically defined as the "participating" section. Approximately one-quarter taught two sections, while another quarter taught more than two sections (Tables S2-4).[16]

For teachers who taught two or more sections, we excluded sections with fewer than six students, then randomly selected one section from the remaining, with larger class sections having greater odds of selection. This approach maximized the probability of selecting larger classes to participate while still maintaining the integrity and benefits of random selection. For example, randomly-selected sections were not always a teachers' largest; the teacher could not choose their "favorite" and/or highest-performing group of students; and class periods throughout the day and week were represented. We defined the selected section as their "participating" section yet assumed they used the KIA approach to teaching APGOV or APES in all of their sections. The "participating" section was

[16] In some districts, data describing the number of sections per teacher was available through administrative records in January 2017 when we randomly selected participating sections. In other districts, we relied on teacher self-reports. In rare cases, we did not have this information and attempted to reach the teacher. In those cases, the teacher never responded but also did not participate in data collection.

the one participating in research activities, including CWRA+, the student survey, and student interviews.

Data collection

Across all five districts, in accordance with district research review board regulations, we facilitated administration of the CWRA+ after the AP examination period and before the end of the school year. In the two districts requiring parental consent, we administered the CWRA+ only to students in the "participating" section, previously described, with the necessary affirmative consent. In the other three districts, we administered the CWRA+ to all students in the "participating" section.

Teachers had a window of 2-4 weeks during which they could administer the CWRA+. Often taking place simultaneously was other district testing (statewide assessments, end-of-year exams), which affected computer and district technical support availability, and might have affected participation and completion rates as well as teachers' and students' engagement with the CWRA+.

Student-level attrition and baseline equivalence

The sample of students with valid CWRA+ outcomes suffered from high levels of student-level attrition (as well as school-level attrition), as seen in Appendix K. Attrition stemmed from several different sources, including lack of teacher participation in the process (in which case the entire classroom, and sometimes school, attrited), lack of parental consent (either due to active denial or simply not completing and returning the permission form), student absences on day of testing, student lack of assent, and technology problems at the classroom or individual levels. Baseline equivalence on all three CWRA+ outcomes exceeded the WWC threshold on six covariates, five of which were teacher- or school-level, plus the student-level covariate describing whether a student took an AP exam in 2016. Several imputed prior achievement covariates did not meet WWC thresholds for baseline equivalence of imputed data (Appendix L). With this level of attrition and baseline imbalance, results cannot meet WWC standards, with or without reservations. Any differences observed must not be interpreted as causal and should be interpreted with caution.

CWRA+ impact results

There were no significant differences between KIA and non-KIA student performance on the CWRA+ test of critical thinking, either on the subgroup sections (the performance task and selected responses sections individually) or on the composite overall score. As shown in Table S1, the magnitude of the estimated performance task effect size, 0.268, was meaningful, though the confidence interval included zero. Given high attrition, lack of baseline imbalance, and lack of significance, these effects, though in the positive direction, are inconclusive.

Table S1: Covariate-adjusted standardized effect sizes for (Year One) CWRA+ outcomes, comparisons between treatment and controls students

	Effect Size (SE)	95% Confidence Interval	p-value	N
CWRA+ overall score	0.202 (0.2)	(-0.193, 0.597)	0.32	489
CWRA+ performance task subscore	0.268 (0.22)	(-0.158, 0.693)	0.22	565
CWRA+ selected response subscore	0.048 (0.16)	(-0.259, 0.355)	0.76	534

Notes: Table shows standardized effect sizes, standard errors, confidence intervals, p-values, and analytic sample sizes. Asterisks denote statistical significance: *p<0.05, **p<0.01, ***p<0.001.

Limitations to Research Question One CWRA+ Outcome Results

The sample of students with valid CWRA+ outcomes suffered from high levels of attrition and lacked equivalence on students' prior achievement measures. Observed results may be due to random chance, and/or differences in the composition of the treatment compared to control analytic samples with CWRA+ outcomes. As such, we urge caution in interpreting CWRA+ impact results.

Table S2: Year One teachers' (n=74) 2016-17 students per teacher-section, sections per teacher, and teachers per school, overall and by treatment and control, across and within course

Count		Overall	Treatment	Control
Students per section				
	Overall	28.98 (10.62)	28.96 (9.34)	29.00 (11.48)
	APES	31.48 (12.58)	31.92 (10.30)	31.21 (13.96)
	APGOV	26.52 (7.58)	26.33 (7.66)	26.67 (7.63)
Sections per teacher				
	Overall	1.69 (1.05)	1.46 (0.85)	1.90 (1.17)
	APES	1.48 (0.83)	1.33 (0.77)	1.58 (0.88)
	APGOV	1.97 (1.23)	1.59 (0.94)	2.40 (1.40)
Teachers per school				
	Overall	1.09 (0.33)	1.13 (0.43)	1.05 (0.23)
	APES	1.07 (0.26)	1.11 (0.32)	1.04 (0.20)
	APGOV	1.21 (0.49)	1.27 (0.59)	1.14 (0.36)

Notes: Section was missing for 22 District E students (one teacher at one school). Standard deviations in parentheses.

Table S3: Year Two teachers' (n=53) 2016-17 students per teacher-section, sections per teacher, and teachers per school, overall and by treatment and control, across and within course.

Count		Overall	Treatment	Control
Students per section				
	Overall	30.18	30.59	29.92
		(10.38)	(7.65)	(11.81)
	APES	33.20	33.21	33.20
		(12.45)	(8.72)	(14.47)
	APGOV	27.38	28.10	26.94
		(7.04)	(5.63)	(7.83)
Sections per teacher				
	Overall	1.92	1.70	2.10
		(1.12)	(0.97)	(1.21)
	APES	1.58	1.46	1.67
		(0.89)	(0.88)	(0.91)
	APGOV	2.41	2.00	2.75
		(1.26)	(1.05)	(1.36)
Teachers per school				
	Overall	1.06	1.10	1.03
		(0.24)	(0.30)	(0.19)
	APES	1.06	1.08	1.06
		(0.25)	(0.28)	(0.24)
	APGOV	1.14	1.22	1.08
		(0.36)	(0.44)	(0.29)

Notes: Section was missing for 22 District E students (one teacher at one school). Standard deviations in parentheses.

Table S4: Year Two teachers' (n=53) 2017-18 students per teacher-section, sections per teacher, and teachers per school, overall and by treatment and control, across and within course

Count		Overall	Treatment	Control
Students per section				
	Overall	28.08	28.40	27.86
		(7.45)	(6.16)	(8.27)
	APES	27.92	29.76	26.68
		(8.16)	(6.71)	(8.90)
	APGOV	28.26	26.89	29.22
		(6.64)	(5.25)	(7.41)
Sections per teacher				
	Overall	1.85	1.74	1.93
		(1.01)	(0.86)	(1.11)
	APES	1.68	1.62	1.72
		(0.83)	(0.87)	(0.83)
	APGOV	2.09	1.90	2.25
		(1.19)	(0.88)	(1.42)
Teachers per school				
	Overall	1.06	1.10	1.03
		(0.24)	(0.30)	(0.19)
	APES	1.06	1.08	1.06
		(0.25)	(0.28)	(0.24)
	APGOV	1.14	1.22	1.08
		(0.36)	(0.44)	(0.29)

Notes: Section was missing for 194 District E students (21 teachers at 21 schools). Standard deviations in parentheses.

Appendix T: Student Survey Summary

Another hypothesized strength of PBL, compared to transmission instruction, is students improving their intrapersonal skills (e.g., grit, self-efficacy) and interpersonal skills (e.g., collaboration and leadership). In addition, developers designed the APGOV and APES KIA courses with the goal of building students' awareness and engagement in civics, politics, and/or the environment.

Outcome measures

We developed an end-of-year student survey to measure such outcomes in Year One. The survey included items measuring students' attitudes towards learning adapted from the Consortium on Chicago School Research's "Becoming Effective Learner's" student survey (Farrington et al., 2014); on "collaboration" from the AIR Deeper Learning survey (American Institutes for Research, 2016 cite); and on "leadership" from the "Yes 2.0" survey (Hansen & Larson, 2005). We adapted items on civic engagement from the California Civic Index and other instruments compiled in Flanagan, Syvertsen & Stout (2007).

Sample

To estimate the impact of the KIA offer on students' inter- and intra-personal skills, we selected one class section for each teacher in which they would administer student surveys. Selecting one class section was intended to help participating teachers who had multiple KIA classrooms by reducing the burden in matters such as obtaining parental consent, reserving computer space, and administering the surveys.

More than half of the analytic sample of teachers taught one section of APGOV or APES, and for these teachers that section was automatically defined as the "participating" section. Approximately one-quarter taught two sections, while another quarter taught more than two sections (Tables S2-4).[17]

For teachers who taught two or more sections, we excluded sections with fewer than six students, then randomly selected one section from the remaining, with larger class sections having greater odds of selection. This approach maximized the probability of selecting larger classes while maintaining the integrity and benefits of random selection. For example, randomly-selected sections were not always a teachers' largest; the teacher could not choose their "favorite" and/or highest-performing group of students; and class periods throughout the day and week were represented. We defined the selected section as their "participating" section, yet assumed they used the KIA approach to teaching APGOV or APES in all of their sections. The "participating" section was the one participating in research activities including the survey, the CWRA+, and interviews.

[17] In some districts, data describing the number of sections per teacher was available through administrative records in January 2017 when we randomly selected participating sections. In other districts, we relied on teacher self-reports. In rare cases, we did not have this information and attempted to reach the teacher. In those cases, the teacher never responded but also did not participate in data collection.

Data collection

Across all five districts, in accordance with district research review board regulations, we facilitated administration of the student survey after the AP examination period and before the end of the school year. We administered the student survey only in teachers' "participating" section, previously described, providing a paper copy to all students with consent, across the analytic sample classrooms with at least one student with affirmative consent.[18] On average, the survey took students around 30 minutes to complete.

Student-level attrition and baseline equivalence

The sample of students with valid survey outcomes met WWC attrition thresholds at the student level but not the school level (Appendix K). However, baseline equivalence on student survey outcomes exceeded the WWC thresholds on six teacher- or school-level covariates, and three student-level covariates, including measures of prior achievement (Appendix L). In addition, we did not have a WWC-required baseline measure of students interpersonal, intrapersonal, or civic engagement constructs. As such, like for the CWRA+, audiences should interpret student survey outcomes with caution.

Student survey impact results

There were no significant differences between KIA and non-KIA students' survey outcomes (Table T1). Almost all survey outcome estimates are less than 0.20 in magnitude and none are statistically significant. The estimate on the "self-efficacy" construct is -0.21 and the estimate on the "interest in politics" construct is -0.25. Given high school-level attrition, lack of baseline balance on key covariates, and lack of statistical significance, all student survey estimates could be due to random chance and/or to differences between the composition of students in the treatment compared to control group.

Table T1: Covariate-adjusted standardized effect sizes for student survey outcomes in Year One, comparing treatment and control students

	Effect Size	Confidence Interval	p-value	N
Self-efficacy	-0.208 (0.15)	(-0.496, 0.08)	0.16	777
Grit	-0.099 (0.16)	(-0.412, 0.213)	0.53	779
Growth mindset	0.038 (0.11)[S]	(-0.186, 0.263)	0.74	780
Appreciation for diversity	0.181 (0.13)	(-0.083, 0.445)	0.18	774
Civic/political efficacy	-0.131 (0.15)	(-0.417, 0.156)	0.37	773
Whether expects to vote regularly at 18	-0.126 (0.12)[S]	(-0.363, 0.111)	0.3	774
Participatory citizenship	0.025 (0.15)	(-0.27, 0.319)	0.87	772
Interest in politics	-0.251 (0.15)	(-0.554, 0.053)	0.11	775
Political voice	-0.077 (0.14)	(-0.359, 0.206)	0.59	771
Concern for the environment	0.128 (0.11)[S]	(-0.092, 0.349)	0.25	774

[18] In one district, the district research review board required that members of the research team administer the student survey and CWRA+.

Course relevance for the future	0.144 (0.15)	(-0.143, 0.431)	0.32	771
Course satisfaction	0.102 (0.16)	(-0.218, 0.421)	0.53	780
Student engagement	-0.005 (0.16)	(-0.314, 0.303)	0.97	775

Notes: Table shows standardized effect sizes, standard errors, confidence intervals, p-values, and analytic sample sizes. Asterisks denote statistical significance: *p<0.05, **p<0.01, ***p<0.001. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

Limitations to Research Question One student survey outcome results

The sample of students with valid survey outcomes suffered from high levels of school-level attrition and lacked equivalence on students' prior achievement measures. Observed results may be due to random chance, and/or differences in the composition of the treatment compared to control analytic samples with survey outcomes. As such, we urge caution in interpreting results on the various survey outcome constructs.

Appendix U: Student-level Baseline Equivalence for Research Question Two and Research Question Three Approach 1

Given that overall and differential school-level attrition exceed WWC thresholds, baseline equivalence analysis is necessary to investigate the extent to which groups differed after attrition. To meet WWC baseline equivalence standards, baseline differences between treatment and control groups on respective relevant student-level covariates must be less than an absolute value of 0.25 standard deviations. In addition, we must include in the impact model any relevant student-level covariates with effect sizes greater than 0.05 in absolute value. Though baseline equivalence analysis of teacher- and school-level covariates is not required per WWC, our most substantial forms of attrition were at the school and teacher levels, so we include teacher- and school-level covariates as part of our baseline equivalence analysis. Thus, baseline equivalence analysis is a necessary step to a) selecting which covariates to include in outcomes models, and b) informing interpretation of impact estimates.

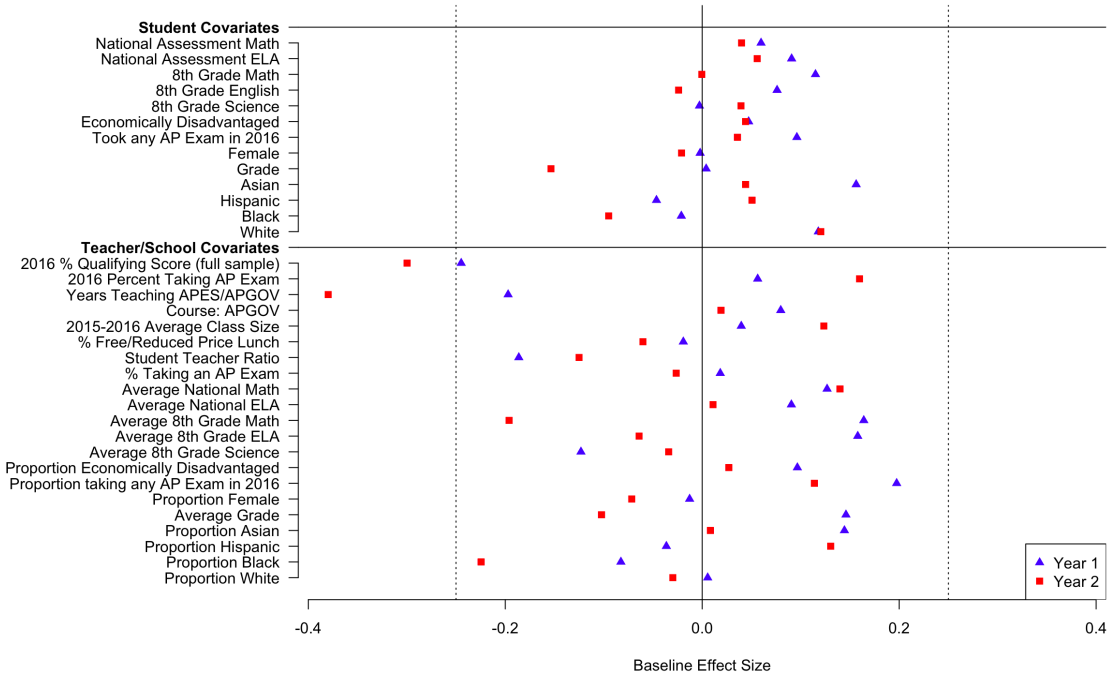
In this Appendix, we use figures to present these differences, using the same format for every figure. We list each of the student-, and teacher-/school-level covariates on the y-axis, demarcate the threshold for baseline equivalence of ±0.25 standard deviations with vertical lines, and use dots to show the point estimate of baseline differences for each covariate. Blue dots represent Year One and red represent Year Two. At the end of this Appendix, we show full tables of standardized mean difference effect sizes (ES) and associated p-values between treatment and control groups on all relevant covariates for all outcome samples.

Below, we first describe baseline equivalence for our primary analytic outcome sample of interest: qualifying score (full sample). This outcome is not contingent upon exam-taking, so samples include all 2016-17 and 2017-18 students of the 53 Research Question Two sample teachers. There was no student-level attrition on the qualifying score (full sample) outcome.

Qualifying score (full sample)

We first present baseline equivalence on the full samples of 2016-17 (n=3,100) and 2017-18 (n=2,946) students composing our samples of exam-taking and qualifying score (full sample). Figure U1 provides an overview of the magnitude of differences between groups across all student-, teacher- and school-level variables.

Figure U1: Baseline standardized mean differences between treatment and control students on all student-, teacher-, and school-level covariates, for Research Questions 2 and 3 Year One (n=3,100 students) and Year Two (n=2,946 students) qualifying score (full sample) analytic outcome samples



Student-level covariates

Figure U1 highlights that effect sizes for all student-level covariates (the top half of each figure) fall within the WWC ±0.25 SD threshold.[19] For both student cohorts, the difference between treatment and control students on prior achievement variables was small and positive, indicating treatment students, on average, performed slightly higher than control. None exceeded an absolute value of 0.134 (see details in tables below). In the opposite direction to the potential positive bias of the prior achievement variables, Year One (ES=0.047) and Year Two (ES=0.044) control students were marginally more economically disadvantaged than treatment.

Though not required by WWC, we also examined baseline equivalence on all other measured student-level characteristics. Though with small effect sizes, treatment groups were composed of a greater proportion of White and Asian students, and smaller proportion of Black students, though effect sizes never exceed 0.16. While Year One control students were a higher proportion Hispanic (ES=-0.046), in Year Two treatment students were a higher proportion Hispanic (ES=0.051).

There were virtually no differences between treatment and control in either year on proportion of female students (Year One ES=-0.002, Year 2 ES=-0.021). Whereas the grade-level difference in Year One was negligible (ES=0.004), in Year Two treatment students were in lower grade levels relative to control with SMD=-0.154.

[19] We anticipate that WWC reviewers will apply the Transition to College protocol when reviewing our evidence of KIA impacts on academic outcomes (i.e., AP scores).

Teacher- and school-level covariates

Teacher- and school-level covariates, though also not required by WWC, provide context to interpreting impact results and informed our covariate selection. Referring back to Figure V1, two variables in Year Two exceeded the WWC threshold: the percent of teachers' 2015-16 APGOV or APES students who earned qualifying scores (full sample) on their respective May 2016 exam, and teacher years of experience teaching APGOV or APES. For both, control teachers were higher than treatment. Though not exceeding thresholds, in Year One the differences followed the same pattern. These large baseline differences indicate covariate adjustment is necessary in our impact models. With baseline differences this large, estimated impacts will be model-dependent, depending on correctly modeling the relationship between this baseline covariate and the outcome.

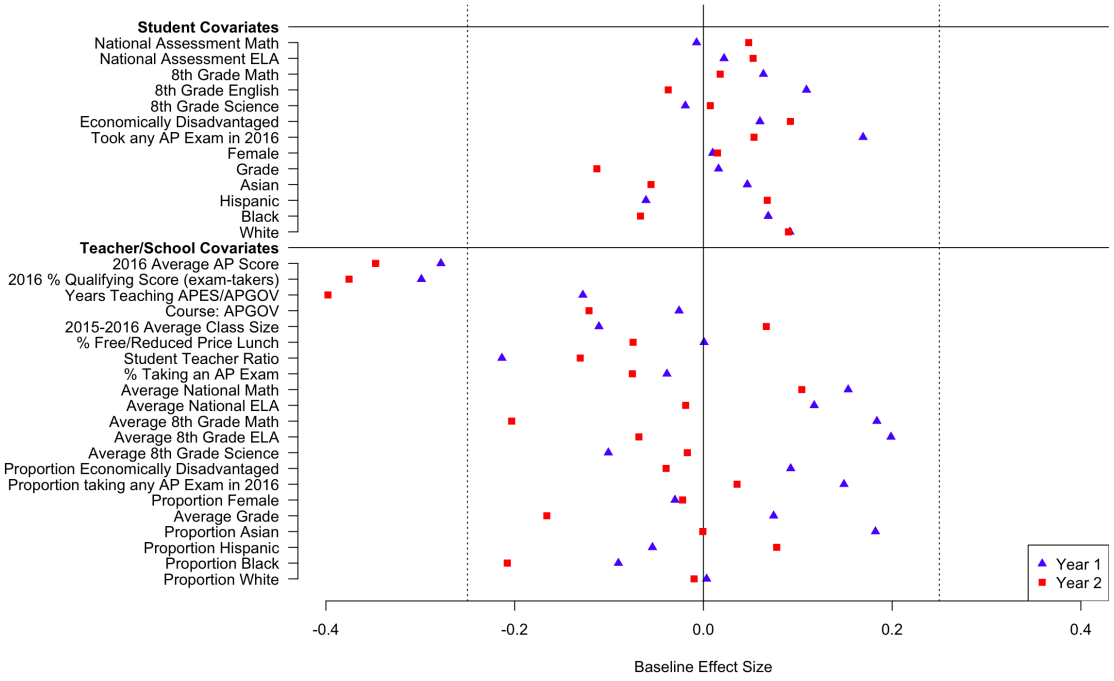
Notably, treatment teachers' students in 2015-16 (baseline) had a lower propensity to earn qualifying scores on the APES/APGOV examination, implying that, in the absence of any treatment, we would expect fewer treatment teachers' students to earn qualifying scores on the 2016-17 and 2017-18 APES/APGOV examinations. In addition, treatment teachers' average 2015-16 class sizes were larger than control in both years, but more so in Year Two. Larger class sizes can present more challenging teaching conditions.

No school-level covariate differences exceeded the WWC ±0.25 standard deviation threshold, and differences were low in magnitude for both percent of the school (in 2015-16) eligible for free or reduced-price meals and percent of the school taking an AP exam in 2015-16, though the former slightly favored treatment schools. The magnitude of the school-level student-to-teacher ratio difference was larger, with smaller ratios in treatment schools, which could also suggest a marginal treatment school advantage.

Qualifying score, exam-taking samples

Those with a qualifying score (exam-takers only) took the exam, and as such, these students are a subset of those in the analyses presented above. Student-level attrition on this outcome meets WWC thresholds (see Appendix K), though due to cluster-level attrition, this study cannot meet WWC standards without reservations. Figure U2 provides the summary view of baseline equivalence across all student-, and teacher-/school-level covariates for the sample of students who took the APES/APGOV exam. We see the same general baseline equivalence patterns as when looking at the full sample of students—that is, among those who took the exam there were no student-level covariates exceeding the WWC threshold of 0.25. Teacher-level covariates describing baseline students' May 2016 APGOV/APES performance exceeded in the negative direction, this time for both Years One and Two, suggesting that treatment teachers' students would also have worse 2017 and 2018 APGOV/APES performance in the absence of any positive KIA impact. Again, including this covariate in impact models will be critical to adjust for this baseline imbalance, and results will be model-dependent. In Year Two, years of teaching experience also exceeded the 0.25 threshold in the negative direction; i.e., treatment teachers were less experienced. Student-level group differences were similar in magnitude as those described above, with treatment students demonstrating higher prior achievement and AP exam taking.

Figure U2: Baseline standardized mean differences between treatment and control students on all student-, teacher-, and school-level covariates for Research Questions 2 and 3 Year One (n=2,537 students) and Year Two (n=2,311 students) qualifying score (exam-takers only) analytic outcome samples



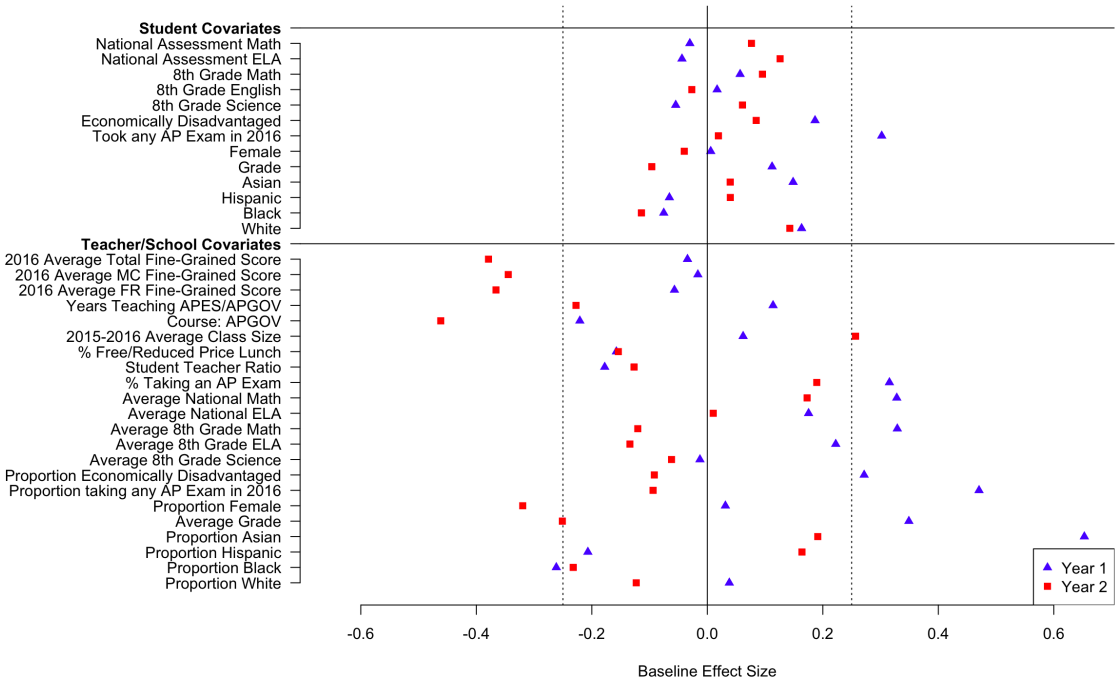
Continuous score sample

Student-level attrition on the continuous score outcome does not meet WWC thresholds (Appendix K). The Year 1 and Year 2 samples of students with continuous total scores, and their multiple-choice and free-response subsection scores, differed from the samples of students with qualifying score (full sample) and qualifying score (exam-takers only) outcomes in several important ways. This was because these samples did not include District D, as it would not permit the research team to access continuous scores. Because the missing district contributed only the APGOV course, the four-district sample consequently is composed of a greater proportion of APES students. In addition, the sample is composed of a greater proportion of lower-income students, as students in the missing district are more advantaged that the other four. For 2017-18 students, whereas 47% of the five-district sample were eligible for free or reduced-price lunch, 61% of the four-district sample were eligible.

Baseline equivalence patterns for the students with continuous scores were similar to those observed for the qualifying score (full sample) and qualifying score (exam-takers only) samples, albeit with more teacher- and school-level differences exceeding thresholds. (Figure U3 provides the summary view.) Year One baseline equivalence shows nine covariates exceeding WWC thresholds (albeit only one at the student-level), including those describing treatment prior achievement classroom averages as higher than control. In Year Two, we see treatment teachers' students performed substantially worse on the APES/APGOV examination, with a large negative effect size of -0.379 for 2016 average AP total score. This difference again suggests treatment teachers' students' 2018 AP continuous outcomes

would likely have been substantially lower than control absent a positive two- to one-year KIA impact, and again necessitates covariate adjustment and model-dependence in our impact estimates.

Figure U3: Baseline standardized mean differences between treatment and control students on all student-, teacher-, and school-level covariates for Research Questions 2 and 3 Year One (n=1,318 students) and Year Two (n=1,424 students) continuous score analytic outcome samples, four districts only



In the following tables of Appendix U, we show standardized mean differences between treatment and control students, and associated p-values, on all student-, teacher-, and school-level covariates, for outcomes as specified. For our AP outcome models, we include as a covariate the 2015-16 AP score average corresponding specifically to that outcome, as opposed to the average AP score and exam-taking rate for all outcomes. For this reason, there are blank cells in Tables U1 and U3 for covariates describing 2015-16 AP performance. Student-level prior achievement variables assume missing at randomness (MAR) in Tables U1 and U3, and bound according to potential deviations from MAR in Tables U2 and U4.

Table U1: Baseline standardized mean differences and associated p-values between treatment and control students on all student-, teacher-, and school-level covariates for Research Questions 2 and 3 Year One (2016-17 students of 53 Year Two teachers) analytic AP outcome samples

	AP QS (full sample) and exam-taking (n=3,100)		AP QS (exam-takers only) (n=2,537)		AP continuous scores (n=1,318)	
	SMD	p-value	SMD	p-value	SMD	p-value
National Assessment Math	0.060	0.689	-0.009	0.954	-0.030	0.892
National Assessment ELA	0.091	0.567	0.020	0.906	-0.044	0.846
Eighth-grade Math	0.115	0.451	0.063	0.712	0.057	0.819
Eighth-grade English	0.076	0.624	0.110	0.523	0.017	0.939
Eighth-grade Science	-0.003	0.984	-0.018	0.897	-0.055	0.781
Economically disadvantaged	0.047	0.835	0.073	0.771	0.187	0.592
Took any AP exam in 2016	0.096	0.759	0.163	0.632	0.302	0.232
Female	-0.002	0.963	0.009	0.858	0.006	0.934
Grade	0.004	0.986	0.011	0.965	0.112	0.571
Asian	0.156	0.425	0.052	0.805	0.149	0.605
Hispanic	-0.046	0.842	-0.071	0.772	-0.066	0.840
Black	-0.021	0.914	0.080	0.704	-0.075	0.808
White	0.118	0.689	0.097	0.730	0.173	0.673
2016 average AP score			-0.283	0.258		
2016 % earned QS (exam-takers)			-0.305	0.213		
2016 % earned QS (full sample)	-0.245	0.270				
2016 % taking AP exam	0.056	0.833				
2016 average total score					-0.034	0.927
2016 average MC score					-0.016	0.965
2016 average FRQ score					-0.057	0.880
Years teaching APES/APGOV	-0.197	0.482	-0.127	0.666	0.114	0.743
Course: APGOV	0.080	0.848	-0.030	0.944	-0.221	0.699
2015-16 average class size	0.040	0.877	-0.099	0.706	0.062	0.839
% free/reduced-price lunch	-0.019	0.916	0.002	0.993	-0.158	0.601
Student-teacher ratio	-0.186	0.136	-0.213	0.052	-0.178	0.261
% taking an AP exam	0.018	0.953	-0.046	0.891	0.315	0.405
Average national Math	0.127	0.649	0.157	0.575	0.328	0.426
Average national ELA	0.091	0.725	0.118	0.643	0.175	0.638
Average eighth-grade Math	0.164	0.482	0.191	0.423	0.329	0.304

Average eighth-grade ELA	0.158	0.531	0.199	0.426	0.222	0.523
Average eighth-grade Science	-0.123	0.587	-0.098	0.658	-0.013	0.971
Proportion school low SES	0.097	0.599	0.084	0.625	0.272	0.311
Proportion school taking 2016 AP exam	0.197	0.564	0.141	0.702	0.471	0.288
Proportion female	-0.013	0.962	-0.022	0.937	0.031	0.916
Average grade	0.146	0.654	0.069	0.842	0.349	0.416
Proportion Asian	0.144	0.632	0.182	0.556	0.653	0.102
Proportion Hispanic	-0.036	0.876	-0.056	0.806	-0.207	0.561
Proportion Black	-0.083	0.672	-0.094	0.622	-0.261	0.283
Proportion White	0.006	0.979	0.007	0.976	0.038	0.887

Table U2: Baseline standardized mean differences between treatment and control students on imputed student-level covariates for Research Questions 2 and 3 Year One (2016-17 students of 53 Year Two teachers) analytic AP outcome samples

AP qualifying score (exam-takers only)

	D1	D2	D3	D4	Most extreme
National Math	-0.009	0.001	-0.029	0.021	-0.029
National ELA	0.020	0.028	0.005	0.043	0.043
Eighth-grade Math	0.063	0.088	0.064	0.087	0.088
Eighth-grade ELA	0.110	0.138	0.109	0.139	0.139
Eighth-grade Science	-0.018	-0.027	-0.013	-0.032	-0.032

AP qualifying score (full sample)

	D1	D2	D3	D4	Most extreme
National Math	0.060	0.067	0.041	0.086	0.086
National ELA	0.091	0.097	0.076	0.112	0.112
Eighth-grade Math	0.115	0.134	0.121	0.129	0.134
Eighth-grade ELA	0.076	0.102	0.075	0.103	0.103
Eighth-grade Science	-0.003	-0.003	-0.003	-0.002	-0.003

Took AP exam

	D1	D2	D3	D4	Most extreme
National Math	0.060	0.164	0.015	0.208	0.208
National ELA	0.091	0.187	0.050	0.228	0.228
Eighth-grade Math	0.115	0.153	-0.107	0.374	0.374
Eighth-grade ELA	0.076	0.124	-0.144	0.344	0.344
Eighth-grade Science	-0.003	0.060	-0.256	0.313	0.313

AP total score

	D1	D2	D3	D4	Most extreme
National Math	-0.030	-0.054	-0.048	-0.037	-0.054
National ELA	-0.044	-0.061	-0.056	-0.049	-0.061
Eighth-grade Math	0.057	0.091	0.062	0.085	0.091
Eighth-grade ELA	0.017	0.053	0.020	0.050	0.053
Eighth-grade Science	-0.055	-0.048	-0.044	-0.059	-0.059

AP multiple-choice score

	D1	D2	D3	D4	Most extreme
National Math	-0.030	-0.060	-0.050	-0.040	-0.060
National ELA	-0.044	-0.065	-0.058	-0.051	-0.065
Eighth-grade Math	0.057	0.095	0.048	0.104	0.104
Eighth-grade ELA	0.017	0.058	0.006	0.069	0.069
Eighth-grade Science	-0.055	-0.047	-0.046	-0.055	-0.055

AP free-response score

	D1	D2	D3	D4	Most extreme
National Math	-0.030	-0.055	-0.050	-0.036	-0.055
National ELA	-0.044	-0.063	-0.059	-0.048	-0.063
Eighth-grade Math	0.057	0.084	0.085	0.057	0.085
Eighth-grade ELA	0.017	0.047	0.043	0.022	0.047
Eighth-grade Science	-0.055	-0.048	-0.037	-0.066	-0.066

Table U3: Baseline standardized mean differences and associated p-values between treatment and control students on all student-, teacher-, and school-level covariates for Research Questions 2 and 3 Year Two (2017-18 students of 53 Year Two teachers) analytic AP outcome samples

	AP QS (full sample) and exam-taking (n=2,946)		AP QS (exam-takers only) (n=2,311)		AP continuous scores (n=1,424)	
	SMD	p-value	SMD	p-value	SMD	p-value
National Assessment Math	0.040	0.804	0.049	0.774	0.076	0.712
National Assessment ELA	0.056	0.750	0.054	0.777	0.126	0.589
Eighth-grade Math	0.000	0.999	0.019	0.909	0.095	0.677
Eighth-grade English	-0.024	0.882	-0.039	0.825	-0.027	0.905
Eighth-grade Science	0.039	0.798	0.008	0.961	0.061	0.774
Economically disadvantaged	0.044	0.836	0.093	0.693	0.085	0.782
Took any AP exam in 2016	0.036	0.921	0.057	0.882	0.019	0.952

Female	-0.021	0.735	0.012	0.863	-0.040	0.628
Grade	-0.154	0.534	-0.110	0.656	-0.096	0.730
Asian	0.044	0.815	-0.063	0.727	0.040	0.880
Hispanic	0.051	0.813	0.072	0.743	0.040	0.894
Black	-0.095	0.631	-0.062	0.782	-0.114	0.687
White	0.120	0.651	0.098	0.751	0.143	0.731
2016 average AP score			-0.345	0.164		
2016 % earned QS (exam-takers only)			-0.373	0.120		
2016 % earned QS (full sample)	-0.300	0.192				
2016 % taking AP exam	0.160	0.483				
2016 average total score					-0.379	0.311
2016 average MC score					-0.345	0.346
2016 average FRQ score					-0.366	0.333
Years teaching APES/APGOV	-0.380	0.195	-0.397	0.197	-0.227	0.508
Course: APGOV	0.019	0.964	-0.122	0.776	-0.462	0.434
2015-16 average class size	0.124	0.616	0.068	0.784	0.257	0.415
% free/reduced-price lunch	-0.060	0.740	-0.077	0.680	-0.154	0.586
Student-teacher ratio	-0.125	0.368	-0.130	0.306	-0.127	0.518
% taking an AP exam	-0.026	0.934	-0.075	0.823	0.190	0.602
Average national Math	0.140	0.625	0.106	0.717	0.173	0.645
Average national ELA	0.011	0.968	-0.016	0.953	0.010	0.977
Average eighth-grade Math	-0.196	0.349	-0.200	0.349	-0.120	0.694
Average eighth-grade ELA	-0.064	0.802	-0.065	0.800	-0.134	0.675
Average eighth-grade Science	-0.034	0.872	-0.014	0.947	-0.062	0.828
Proportion school low SES	0.027	0.892	-0.039	0.840	-0.092	0.798
Proportion school taking any 2016 AP exam	0.114	0.748	0.035	0.924	-0.094	0.816
Proportion female	-0.072	0.815	-0.019	0.953	-0.320	0.389
Average grade	-0.102	0.758	-0.165	0.638	-0.251	0.502
Proportion Asian	0.008	0.979	-0.003	0.993	0.191	0.663
Proportion Hispanic	0.130	0.560	0.076	0.728	0.164	0.619
Proportion Black	-0.225	0.313	-0.208	0.341	-0.232	0.353
Proportion White	-0.030	0.888	-0.006	0.976	-0.123	0.637

Table U4: Baseline standardized mean differences and associated p-values between treatment and control students on imputed student-level covariates for Research Questions 2 and 3 Year Two (2017-18 students of 53 Year Two teachers) analytic AP outcome samples

Took AP exam

	D1	D2	D3	D4	Most extreme
National Math	0.040	0.133	-0.093	0.266	0.266
National ELA	0.056	0.149	-0.077	0.281	0.281
Eighth-grade Math	0.000	0.119	-0.231	0.349	0.349
Eighth-grade ELA	-0.024	0.072	-0.213	0.261	0.261
Eighth-grade Science	0.039	0.158	-0.039	0.237	0.237

AP Qualifying score outcome (full sample)

	D1	D2	D3	D4	Most extreme
National Math	0.040	0.062	0.015	0.087	0.087
National ELA	0.056	0.075	0.034	0.097	0.097
Eighth-grade Math	0.000	0.106	0.069	0.036	0.106
Eighth-grade ELA	-0.024	0.081	0.074	-0.017	0.081
Eighth-grade Science	0.039	0.124	0.102	0.061	0.124

AP Qualifying Score outcome (exam-takers only)

	D1	D2	D3	D4	Most extreme
National Math	0.049	0.062	0.035	0.076	0.076
National ELA	0.054	0.064	0.041	0.077	0.077
Eighth-grade Math	0.019	0.116	0.080	0.055	0.116
Eighth-grade ELA	-0.039	0.069	0.056	-0.027	0.069
Eighth-grade Science	0.008	0.084	0.072	0.020	0.084

AP total score

	D1	D2	D3	D4	Most extreme
National Math	0.076	0.069	0.083	0.063	0.083
National ELA	0.126	0.121	0.131	0.116	0.131
Eighth-grade Math	0.095	0.153	0.065	0.184	0.184
Eighth-grade ELA	-0.027	0.027	-0.059	0.058	-0.059
Eighth-grade Science	0.061	0.087	0.058	0.090	0.090

AP multiple-choice score

	D1	D2	D3	D4	Most extreme
National Math	0.076	0.073	0.080	0.070	0.080
National ELA	0.126	0.124	0.128	0.122	0.128
Eighth-grade Math	0.095	0.146	0.068	0.173	0.173
Eighth-grade ELA	-0.027	0.020	-0.055	0.048	-0.055
Eighth-grade Science	0.061	0.081	0.059	0.083	0.083

AP free-response score

	D1	D2	D3	D4	Most extreme
National Math	0.076	0.062	0.089	0.050	0.089
National ELA	0.126	0.115	0.135	0.105	0.135
Eighth-grade Math	0.095	0.178	0.056	0.218	0.218
Eighth-grade ELA	-0.027	0.047	-0.069	0.089	0.089
Eighth-grade Science	0.061	0.101	0.058	0.104	0.104

Appendix V: Research Question Three Approach 1 Sensitivity Analyses

In Table V1, for the separate Years One and Two impact estimates informing the combined two-year estimate, we show estimated effect sizes a) without covariates (i.e., unadjusted), b) using our primary approach, c) including all covariates, and d) including all covariates with baseline equivalence differences of greater than 0.05 (i.e., those required for inclusion in impact models per the WWC). Cells show effect sizes with standard errors in parentheses.

Table V1: Sensitivity to covariate selection of estimates of the differences in student outcomes between one and zero years of the KIA offer to teachers, and differences in student outcomes between two and one years of the KIA offer to teachers

	1 vs. 0 years of KIA offer				2 vs. 1 years of KIA offer			
	No covs.	Primary	All covs.	Covs. ABE > 0.05	No covs.	Primary	All covs.	Covs. ABE > 0.05
Took AP exam	-0.014 (0.2)	0.025 (0.13)	0.051 (0.14)	-0.038 (0.14)	0.444* (0.19)	0.222 (0.18)	0.121 (0.2)	0.223 (0.18)
QS (full sample)	-0.043 (0.24)	0.374 (0.11)***[S]	0.454 (0.16)**[S]	0.129 (0.13)	0.062 (0.31)	0.212 (0.13)[S]	0.171 (0.17)[S]	0.151 (0.12)[S]
QS (exam-takers only)	-0.059 (0.24)	0.44 (0.14)**[S]	0.495 (0.19)*[S]	0.207 (0.16)	-0.048 (0.32)	0.087 (0.14)[S]	0.129 (0.19)[S]	0.043 (0.16)[S]
Total score	0.044 (0.21)	0.186 (0.12)	0.170 (0.12)	0.185 (0.11)	0.136 (0.22)	0.163 (0.11)	0.141 (0.11)	0.16 (0.11)
Multiple-choice	0.086 (0.21)	0.204 (0.12)	0.178 (0.12)	0.203 (0.12)	0.122 (0.22)	0.151 (0.11)	0.138 (0.11)	0.148 (0.11)
Free-response	-0.002 (0.2)	0.154 (0.11)	0.147 (0.12)	0.153 (0.11)	0.143 (0.21)	0.126 (0.11)	0.085 (0.1)	0.124 (0.11)

*=p<0.05, **=p<0.01, **=p<0.001. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

The substance of most estimated effect sizes is consistent between our primary model and models including all covariates. AP qualifying score (full sample) and qualifying score (exam-takers only) estimates are lower when we do not include in adjusted models those covariates with baseline differences of less than 0.05 ES (selected through automation to improve the precision of estimates), indicating the importance of those covariates.[20]

[20] For Year 1, baseline equivalence differences were less than 0.05 on measures of student eighth-grade Science scores (ES=-0.003), student economic disadvantage (ES=0.047), proportion of sample students who were Black (ES=-0.021) or Hispanic (ES=-0.046), average teacher 2015-16 class size (ES=0.040), percentage of the school as of 2015-16 eligible for free or reduced-price lunch (ES=-0.019), and percentage of the school taking

The difference in magnitude between the unadjusted and adjusted estimates indicates the necessity of statistical controls for estimating a positive KIA effect. This means our results are model-dependent; i.e., they depend on correctly modeling the relationship between the covariates and the outcomes.

Notably, the covariate describing teachers' baseline students' AP performance on the analog outcome measure accounts for the differences between "2 versus 1 years of the KIA offer" estimated effect sizes in models with no covariates compared to our primary model estimates (Table V2, with cells showing effect sizes and standard errors). We chose to investigate this one covariate because 1) it is a pre-treatment version of each outcome, averaged at the teacher level, and 2) differences between treatment and control teachers' baseline students' APGOV/APES exam performance exceeded WWC thresholds, with students of teachers with one year of the KIA offer outperforming those with two years of the offer, while baseline students of two-year KIA teachers had higher exam-taking rates.

Table V2: Sensitivity of impact estimates describing one versus zero and two versus one years of the KIA offer to adjustment with a single covariate describing teachers' baseline (i.e., 2015-16) students' May 2016 APGOV/APES exam performance

	1 vs. 0 years of KIA offer				2 vs. 1 years of KIA offer			
	No covs.	Primary	Single prior AP perf cov.	N	No covs.	Primary	Single prior AP perf cov.	N
Took AP exam	-0.014	0.025	-0.068	3100	0.444*	0.222	0.356	2946
	(0.2)	(0.13)	(0.14)		(0.19)	(0.18)	(0.16)*	
QS (full sample)	-0.043	0.374	0.120	3100	0.062	0.212	0.198	2946
	(0.24)	(0.11)***[S]	(0.14)		(0.31)	(0.13)[S]	(0.17)	
QS (exam-takers only)	-0.059	0.44	0.121	2537	-0.048	0.087	0.094	2311
	(0.24)	(0.14)**[S]	(0.15)		(0.32)	(0.14)[S]	(0.18)	
Total score	0.044	0.186	0.096	1318	0.136	0.163	0.280	1424
	(0.21)	(0.12)	(0.17)		(0.22)	(0.11)	(0.17)	
Multiple-choice	0.086	0.204	0.149	1318	0.122	0.151	0.252	1424
	(0.21)	(0.12)	(0.16)		(0.22)	(0.11)	(0.16)	
Free-response	-0.002	0.154	0.029	1318	0.143	0.126	0.255	1424
	(0.2)	(0.11)	(0.18)		(0.21)	(0.11)	(0.16)	

*p<.05

On the qualifying score (full sample) outcome, the Year Two effect size with the one covariate describing teachers' baseline students' credit/no credit rate (ES=0.198, SE=0.17) is substantively the same as our primary model qualifying score (full sample) effect size of 0.212 (SE=0.13). We see the same in the magnitude of the Year Two qualifying score (exam-takers only) outcome in the outcome

an AP exam in 2015-16 (ES=0.018). In Year 2, they included student economic disadvantage (ES=0.044), proportion of Asian students (ES=0.044), and percentage of the school taking an AP exam in May 206 (-0.026).

including only teachers' baseline students' APGOV/APES qualifying score (exam-takers only) rate as a covariate (ES=0.094, SE=0.17) compared to our primary model (ES=0.087, SE=0.14). With the exam-taking outcome, the Year Two effect size is significant, with the one analog baseline covariate (ES=0.356, SE=0.16) higher than our primary model estimate (ES=0.222, SE=0.18)—but, in this case, lower than the unadjusted estimate (ES=0.444, SE=0.19). In continuous score outcomes models with the single covariate, estimated effect sizes are of higher magnitude, by more than a standard deviation, than primary model estimates.

We see a similar pattern within Year One estimates, as adjusted effect sizes are greater than unadjusted for AP exam performance measures, and lower than unadjusted for exam-taking. However, the differences are not generally as large, as the baseline imbalance was lower in magnitude in Year One.

Appendix W: Derivation of Inflated Experimental Indirect Variance

Define $\hat{\tau}_1$ to be the one-year impact from comparing T1 to C1, and $\hat{\tau}_{2-1}$to be the difference between two-year and one-year impact as estimated in Research Question One. Then, our estimator of interest for Research Question Two is defined as

$$\hat{\tau}_2 \equiv \hat{\tau}_1 + \hat{\tau}_{2-1}.$$

The variance of $\hat{\tau}_2$ is then

$$\begin{aligned}\text{var}(\hat{\tau}_2) &= \text{var}(\hat{\tau}_1 + \hat{\tau}_{2-1}) \\ &= \text{var}(\hat{\tau}_1) + \text{var}(\hat{\tau}_{2-1}) + 2\text{cov}(\hat{\tau}_1, \hat{\tau}_{2-1}) \\ &= \text{var}(\hat{\tau}_1) + \text{var}(\hat{\tau}_{2-1}) + 2\text{cor}(\hat{\tau}_1, \hat{\tau}_{2-1})\sqrt{\text{var}(\hat{\tau}_1)\text{var}(\hat{\tau}_{2-1})}.\end{aligned}$$

If $\text{var}(\hat{\tau}_1) \approx \text{var}(\hat{\tau}_{2-1})$—which they are, for the most part, because they are based on comparisons of the same sets of teachers on the same outcomes, just measured in different years—this simplifies to

$$\text{var}(\hat{\tau}_2) \approx [2 + 2\text{cor}(\hat{\tau}_1, \hat{\tau}_{2-1})]\text{var}(\hat{\tau}_1).$$

Calculating the correlation between the two estimates exactly is intractable without making additional assumptions, but the correlation will be positive because both estimates are based on comparisons of the same two groups of teachers, and any inevitable differences between these teachers, aside from the treatment, will persist across both estimates. Therefore, the variance of the indirect two-year impact estimate, $\hat{\tau}_2$, is at best double, and at worse quadruple, of the variance of the direct one-year impact estimate, $\hat{\tau}_1$. The truth is likely to be closer to quadruple than double due to the inevitably high correlation between the estimates.

Appendix X: Methods for Selecting Non-Experimental Comparison Teachers—First Round

To inform our non-experimental study of KIA teachers' students' performance two years after their original offer to participate in the KIA intervention, we constructed a counterfactual condition using statistical matching procedures, with propensity scores balancing the distribution of measured, school-, teacher-, and student-level pre-treatment and invariant covariates—that is, those that are unaffected by treatment assignment, such as race/ethnicity, gender, eligibility for free or reduced-price lunch, and prior academic performance—between experimental (T) and non-experimental conditions (NE).

We limited our sample to include teachers in experimental schools randomized into the treatment condition (T2), and non-experimental teachers (N2), who taught APGOV or APES in 2016-17 and 2017-18, and who were present in participating district schools between 2015-16 to 2017-18 (T0-2). In our "first-round" matching procedure, to reduce the risk of spillovers between experimental and non-experimental teachers, we excluded non-experimental teachers who did not teach APGOV or APES in both years, and those who taught in a randomized experimental school. We constructed a comparison group of non-experimental teachers who did not participate in the RCT, did not receive KIA professional development or support, and did not teach in any randomized school during the study period. Table X1 lists the criteria used to determine experimental and non-experimental teachers' eligibility for inclusion. After applying the study's exclusion criteria, remaining were 90 non-experimental and 23 treatment teachers.

Table X1. Study eligibility criteria, and sample loss and remaining eligibility counts, by non-experimental study condition

Group	Non-experimental teachers	Treatment teachers	Total teachers
All teachers who taught APGOV or APES for at least one year between 2015-16 and 2017-18	260	43	303
Not excluded by any condition	67	23	90
Excluded teachers by reason			
Deterministically randomized school	1	2	3
Principal denied consent	0	1	1
Non-eligible course	4	5	9
Did not teach in 2016-17	99	2	101
Did not teach in 2017-18	90	17	107
Taught at school with KIA teacher	70	0	70
Taught at school included in the 2015-16 KIA Pilot Study	25	0	25
Moved schools in 2017-18	1	1	2
Total excluded teachers	193	20	213

Note. Totals in the far-right column are sums by row. The bottom row totals equal the first minus second rows. Exclusion reason was not mutually exclusive, as a teacher could have met more than one exclusion criteria. Thus, the sum of teachers disqualified from inclusion is not equal to the total number of excluded teachers.

Matching methodology

Paralleling the baseline covariates used in the experimental arm of the study, covariates available for inclusion in the propensity score model included teacher-level baseline year (2015-16) average student performance on the outcomes of interest: AP exam-taking rates and scores. While pre-test information can substantially reduce bias in the estimated treatment effect in non-experimental studies[21], we supplemented with additional school and teacher covariates potentially related to teachers' KIA participation and students' AP performance outcomes.

Our selection of covariates for inclusion in the propensity score model reflected two sources of anticipated imbalance on measured covariates between T2 and N2 teachers. The first originates from teacher selection into KIA during the spring and summer of the 2016 school year, two years prior to measurement of student outcomes. Factors that may be related to teachers' decision to consent to participate in the KIA study, and for which we received data from participating districts, primarily include their 2015-16 (baseline) APGOV/APES students' AP exam-taking rates and scores.[22] The second source includes differences between T2 and N2 teachers' 2017-18 students, and may reflect student selection into a KIA course. To account for both sources of potential bias, we included teachers' 2017-18 students' baseline covariates in the propensity score model.

The underlying process determining teacher and student selection into KIA is not known, and whether the assumptions[23] underpinning matching methods for generating credible causal estimates of the impact of KIA on student outcomes are met—notably, that conditional on pre-treatment covariates, summarized by the propensity score, the potential outcomes between KIA and matched non-KIA sample are equivalent—is indeterminate.[24] Furthermore, because of the small sample size and the number of available covariates eligible for estimating the propensity score, we confronted a trade-off between maximizing the covariates included in the propensity score model, and the number of T2 and N2 teachers who could be successfully matched without endangering the covariate balance we sought

[21] For example, see Wong, Valentine, and Miller-Bains (2016) for an overview of the literature on the importance of pre-test information for reducing bias in observational studies of educational interventions.
[22] It is important to note that teacher-level baseline pre-test data are available only for school year 2015-16. If selection into KIA is influenced by dynamic factors—or other non-dynamic factors not included in the vector of covariates used to impose balance between conditions correlated with the outcome measure—the model will be mis-specified and the strong ignorability assumption will be violated.
[23] The risk of a SUTVA violation is minimal, as we disqualified from this study non-experimental teachers who were teaching the targeted course in schools with a KIA teacher.
[24] This is closely related to the notion of the "propensity score tautology," in which the efficacy of matching based on the propensity score is assessed based on the balance of raw covariates, irrespective of whether this balance truly achieves the requirements necessary to satisfy the unconfoundedness assumption (Ho et al., 2007).

to achieve through matching.[25] This trade-off occurred irrespective of the algorithm used to select matched controls with the propensity score.

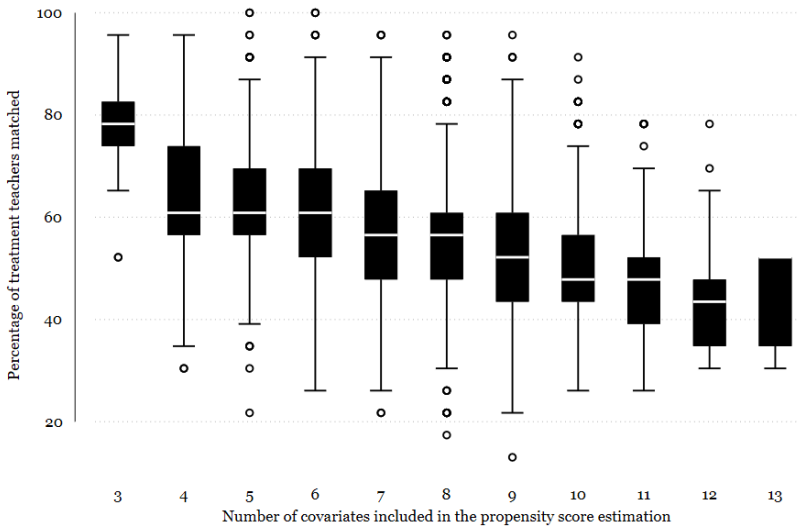
We illustrate the trade-off in Figure X1, which displays the relationship between the number of covariates included in a given propensity score model and the percentage of treatment teachers matched to at least one control teacher. The results are derived from approximately 13,000 unique matching solutions, with each varying the number of covariates[26] included, and the matching algorithm. The matching algorithm yielded 30,720 unique models, and its tuning parameters included:

1. Covariates, with 1,024 unique covariate combinations, ranging from as few as three covariates to a maximum of 13
2. Caliper width: between no caliper, 0.2, 0.3, 0.4, and 0.5 standard deviation of the logit of the propensity score
3. With and without control teacher replacement
4. Between one and three nearest neighbors

The dimensionality of the functional form for estimating the propensity score is negatively related to the number of treatment teachers matched to a control. For instance, the median number of treatment teachers successfully matched to a single control teacher was 78%, compared to 52% for models with 13 covariates. This pattern persists across the different matching algorithms, which are not shown, but are available upon request from the authors.

[25] We attempted to reduce dimensionality by combining some related covariates, including combining the national college readiness assessments into a single variable, and creating a single measure of racial composition.
[26] The total number of covariates ranged between three and 13. All three baseline covariates—which included two teacher-level pre-test covariates (percentage of students who took the respective AP exam, and the average APES/APGOV exam score) from the 2015-16 school year—were included in all permutations because of the measure's proximity to teachers deciding on KIA participation, the magnitude of the baseline imbalance between non-experimental and treatment teachers on these covariates, and the research team's qualitative and theoretical understanding of the factors influencing participation in KIA. Additionally, because of the importance of these baseline measures, no matching solution was included in the universe of acceptable matches if the absolute standardize difference between conditions was greater than or equal to WWC benchmark of 0.25 standard deviations.

Figure X1: Relationship between the number of covariates included in the propensity score model and the percentage of matched teachers



Note. Underlying data were obtained from 13,390 unique matching permutations with different covariates, calipers, maximum nearest neighbors, and allowing/disallowing replacement. Box height reflects the inter-quartile range, and the horizontal white bar represents the median percentage of matched treatment teachers. The whiskers represent extreme values 1.5 times greater than the upper and lower quartiles, and the dots reflect outliers beyond these ranges.

While our primary objective with matching was to maximize the equivalence of the joint distribution of the covariates available to the research team, we also wanted to maximize the number of teachers included in the analysis. This was important for statistical power; that is, our ability to detect a difference between KIA and non-experimental teachers, if one exists. It also was important as related to generalizability, due to the few KIA teachers in some districts who remained in the analytic sample after the sample inclusion rules were applied and lack of overlap in conditional probabilities of participating between the experimental and non-experimental conditions. Thus, to balance the objectives of maximizing both equivalence and sample size, we developed an index codifying these trade-offs and used it to select a single matching solution.[27] We specified the index as:

$$\max\{Totalmatches * (1 - Treatmatches) * Numcovs\}\ if\ cABS \leq .25\sigma \quad (1)$$

where *Totalmatches* is the total number of T2 and N2 teachers successfully matched and falling within the region of common support[28]; *Treatmatches* is the total number of unique treatment teachers for which at least one matched pair was found; *Numcovs* is the total number of covariates included in the

[27] Future sensitivity analyses could examine differences in the estimated effect of KIA after two years across different specifications.

[28] We globally enforced the common support by removing observations with a propensity score greater/less than the maximum/minimum of the other condition.

propensity score model; and c*ABS* is the maximum absolute standardized bias of all included covariates.

Full set of covariates considered for inclusion

Table X2 lists the school- and teacher-level covariates available to the research team. In addition, we provide the summary statistics, by condition, to assess the comparability of KIA and the complete pool of eligible non-experimental teachers before matching. Mean differences between the unmatched groups are reported in standard deviations. Covariates with stars were forced for inclusion in all iterations of propensity score estimation and matches were required to exhibit imbalance of less than 0.25 absolute standard deviations. We provide details about the transformation of variables measuring student achievement in Appendix E.

Table X2: Characteristics of non-experimental and experimental teachers in the full sample prior to matching

Covariate	N2	T2	Standardized difference
Average APES/APGOV score, baseline year (teacher)*	0.20	-0.57	-0.83
Percentage of 2015-16 students who took APES/APGOV exam (teacher)*	0.07	-0.11	-0.19
Average 2015-16 class size (teacher)*	-0.15	-0.02	0.11
Average 2015-16 students' eighth-grade Science scores (teacher)	-0.11	-0.33	-0.37
Average 2015-16 students' national Math and ELA combined (teacher)	-0.12	-0.65	-0.38
Average of whether 2015-16 students took any AP exam (teacher)	0.60	0.59	-0.03
2015-16 percentage of Black and Hispanic students (teacher)	0.45	0.50	0.14
Female (teacher)	0.58	0.56	-0.18
Average 2015-16 students' economic disadvantage status (teacher)	0.49	0.50	0.05
2015-16 school % students eligible free/reduced-price lunch (school)	0.08	0.20	0.12
2015-16 student-teacher ratio (school)	0.06	0.05	-0.01
2015-16 proportion of students taking AP exam (school)	-0.03	-0.44	-0.40
2015-16 average APGOV/APES grade level taught (teacher)	0.06	0.06	0.01

Note. We included starred covariates in all propensity score specifications.

Missing pre-treatment covariates

Baseline data are incomplete for some T2 and N2 teachers We imputed missing covariates using the multiple imputation method described in Appendix J.

Estimating the Propensity Score and Selecting Matched Comparison Teachers

To select a comparison group of teachers who did not consent to participate in the KIA RCT but who, conditioned on measured pre-treatment covariates, had similar probabilities of participating in KIA as a T2 teacher, we used *k*:1 nearest neighbor matching (NNM) to select *k* nonexperimental teachers, up to a maximum of three, using a "greedy" matching algorithm. We selected nearest

neighbors using the logit of the propensity score[29] estimated from a logistic regression. All analyses were performed at the teacher level. The model iteratively regressed a binary indicator of KIA participation ("1" for T2 teachers, "0" for N2 teachers) on all unique combinations of covariates derived from the full list of covariates detailed in Table Y2, although all models included the three starred baseline covariates. Further, to optimize both the number of covariates for which adequate balance was achieved and the number of teachers successfully matched, for each covariate combination, we produced matching solutions for several different matching algorithms and tuning parameters:

1. With and without replacement of control group teachers
2. Caliper-based matching, with calipers ranging between .10 and .5 standard deviations of the logit of the propensity score, in increments of .10
3. Nearest neighbor matching, between one and three nearest neighbors

In addition, we used exact matching on two covariates: district and course, and common support was enforced on all acceptable matching solutions.

We implemented the approach in three stages, following the procedures and notation of Becker and Ichino (2002):

Step 1: Fit a logistic regression:

$$\Pr(D_i = 1 \mid X_i) = \Phi\{h(X_i)\} \quad (2)$$

Where Φ is the propensity score, and $h(X_i)$ is a vector of pre-treatment (2015-16 teacher and 2017-18 teachers' students' pre-intervention data) teacher and school covariates. We fit models that included all combinations of covariates itemized in Table X2, forcing the inclusion of the three starred teacher-level baseline covariates in all fitted models.

Step 2: Find the nearest *k* non-experimental neighbor for T2 teacher within a caliper of *S* standard deviations of the logit of the propensity score, which ranged between 0.1 and 0.5, in increments of 0.10, for iterations with caliper-based matching, where a neighbor within the same district and who taught the same subject was selected irrespective of the difference in propensity scores.

$$C(i) = \left\| p_i - p_j \right\| \leq c_{s\sigma} \quad (3)$$

We selected the non-treated units (*j*) that satisfy the condition $(i) = \left\| p_i - p_j \right\| \leq c_{s\sigma}$; in other words, the N2 teachers with the smallest logit within 0sσ of each T2 teacher. We performed Step 2 separately by block (district) and by course, and picked non-experimental counterparts from within the same district and course to maximize balance on covariates, both observed and unobserved, varying across geography and districts. This procedure mimics the experimental approach of the Efficacy Study, which used a block-randomized design with school-level randomization by district, and is consistent with a large literature on the non-experimental design features yielding unbiased causal effects

[29] See Rosenbaum and Rubin (1985) for a discussion of the superior statistical properties and performance of the logit compared to the propensity score for bias reduction.

validated against experimental estimates.[30] All matching performed was blind to the outcome data, and the impact of a given matching solution on outcome differences between T and N conditions. The resulting matching solution only advanced to Step 3 if the absolute difference in the standard deviation of the propensity score of each baseline covariates was less than or equal to 0.25 standard deviations of the logit of the propensity score. From the full universe of 30,720 potential matching configuration, 13,390 matching solutions progressed to Step 3.

Step 3. Apply the index to select a single matching solution. To select a single matching solution from the approximately 13,000 satisfying the above criteria, we used formula in Step 1 to identify the matched sample. The selected matching solution parameters are itemized below, and we summarize matching diagnostics in the subsequent section.

1. No caliper
2. Two nearest-neighbors
    a. 22 of 23 treatment teachers were matched
    b. 24 of 90 control teachers were matched
3. With replacement
4. Eight covariates

Figure X2 provides the counts of control teachers matched to at least one treatment teacher, by course and district. Values on the y-axis, which range between 1 and 5, reflect the number of unique treatment teachers to which a control teacher was matched. For instance, the 11 control teachers in District E successfully matched to at least one treatment teacher, seven were matched only once, while two were matched to more than two treatment teachers.

[30] See Cook, Shadish, and Wong (2008) for within-study design evidence demonstrating the importance of local matching for replicating an experimental estimate using matching.

Figure X2. Distribution of matched non-experimental control teachers, by course, district, and number of treatment teacher matches

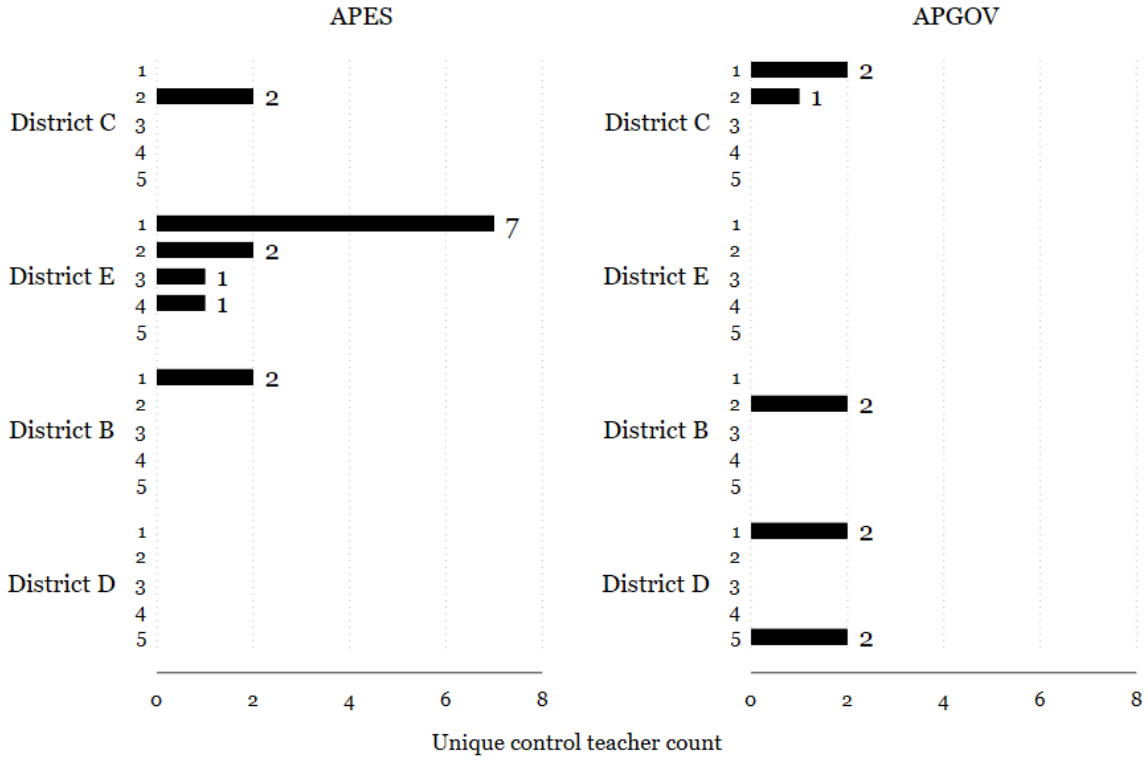


Figure X3's paired-plot illustrates the distances (based on the logit of the propensity score) between each treatment teacher (22) and matched control (24), by course and by district. The filled circles at the base of each arrow denote treatment teachers, and their positions on the y-axis reflects their estimated logit. The arrow barb is the predicted logit of the treatment teacher's matched pair, and the arrow's slope captures the absolute distance in the estimated logits of the propensity score between each match. Hollow circles indicate unmatched control teachers, while filled circles not linked by an arrow are unmatched treatment teachers. For instance, in District A, there were only two eligible control matches for a single APES teacher, and the distance between the nearest eligible match was nearly three logits. This figures further illustrates that, while the overall balance on the covariates of interest was within a tolerable range (demonstrated in Figure X5), varying markedly across district and course was the quality of matches, according to the absolute distance in the logit of the propensity score between a treatment teacher and her matched controls.

Figure X3. Distance between treatment teachers and matched control teachers based on propensity scores, by course and district

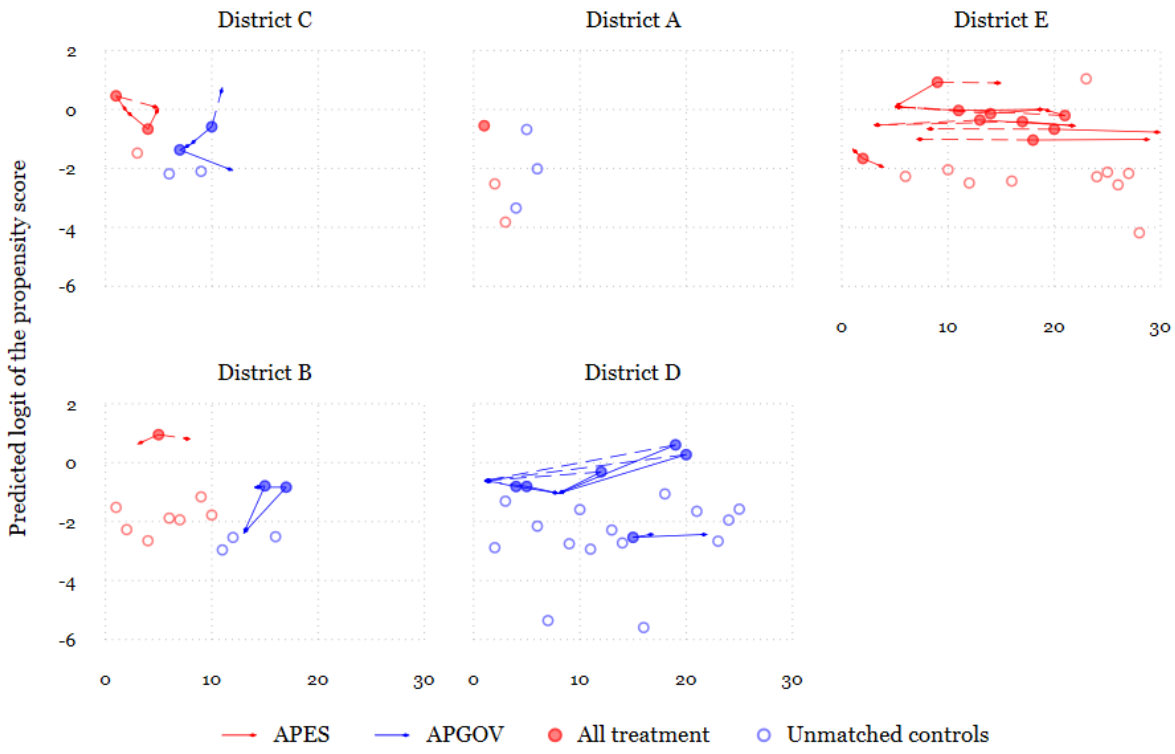


Figure X4 presents kernel density estimates of the propensity scores' predicted logits from the final matching solution for four groups: all 23 treatment and 67 control teachers, and the subsets of matched treatment (22) and control (24) teachers. Enforcing common support served to trim treatment and control teachers with estimated propensity scores below -.25 logits.

Figure X4. Overlap between treatment and control teachers, by matching status

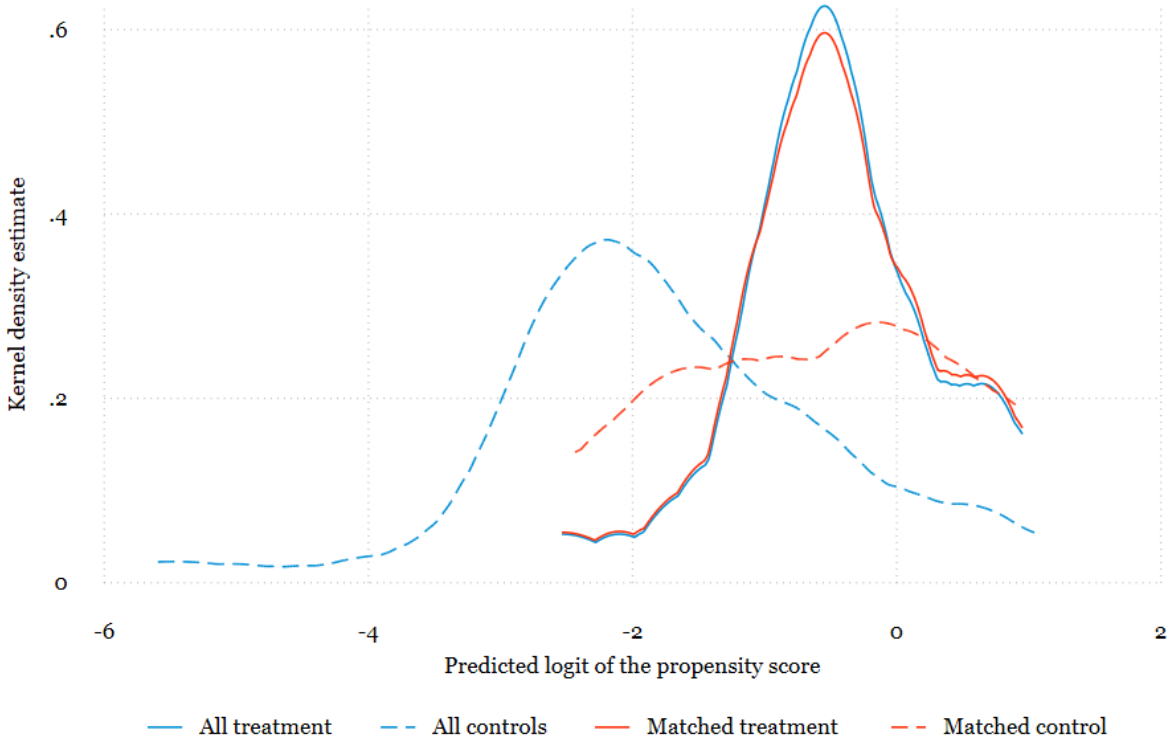
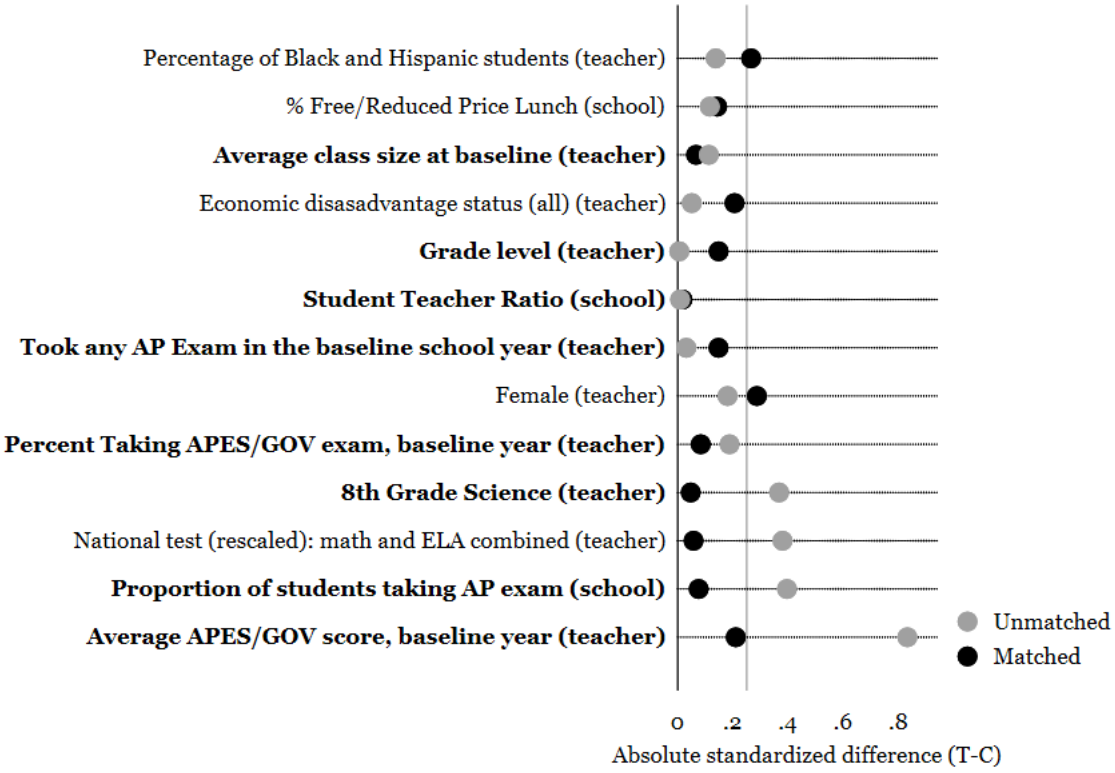


Figure X5 displays the baseline equivalence across all 13 covariates before and after matching, with covariates included in the final propensity score model bolded. Only two covariates in a teacher's 2017-18 classroom—the percentage of Black and Hispanic students, and the percentage of female students—exceeded the 0.25 absolute standardized value threshold. Neither covariate was included in the model estimating the propensity score.

Figure X5. Covariate balance between treatment and control teachers, matched and unmatched samples



Note. Bolded covariates indicate variables included in the selected propensity score estimation model. The vertical line denotes an absolute standardized difference of .25.

Appendix Y: Unmatched and Matched School, Teacher-, and Student-level Descriptive Statistics for Non-experimental Compared to Two-Year Experimental Treatment—First Round

School-level descriptive statistics

Table Y1: Year Two (2017-18) school-level baseline (2015-16) characteristics overall and by experimental treatment and non-experimental control status, across and by courses, unmatched and matched.

Variable	Overall Unmatched	Treatment Unmatched	Non-exp. Control Unmatched	Matched Overall	Matched Treatment	Matched Control
Counts						
Overall	66	21	45	44	20	24
APES	46	13	33	27	12	15
APGOV	28	9	19	18	9	9
School percentage FRPL						
Overall	59.49	54.06	62.02	58.46	56.76	60.00
APES	70.22	66.02	71.88	77.77	71.52	84.02
APGOV	36.26	36.85	35.98	33.86	36.85	31.16
Magnet school						
Overall	40.91	47.62	37.78	44.05	50.00	38.64
APES	41.30	46.15	39.39	54.17	50.00	58.33
APGOV	32.14	44.44	26.32	28.95	44.44	15.00
Title 1 school						
Overall	63.64	61.90	64.44	60.71	60.00	61.36
APES	82.61	92.31	78.79	95.83	91.67	100.00
APGOV	17.86	11.11	21.05	13.16	11.11	15.00
School percentage LEP						
Overall	10.08	10.29	9.98	10.57	10.11	10.98
APES	10.68	10.91	10.59	12.54	10.67	14.40
APGOV	7.64	9.45	6.79	8.09	9.45	6.87
School student-to-teacher ratio						
Overall	19.58	19.29	19.72	19.57	19.72	19.44
APES	21.46	21.90	21.29	22.25	22.83	21.66
APGOV	16.28	15.39	16.70	16.12	15.39	16.77
Urban school						

Overall	71.21	71.43	71.11	64.29	70.00	59.09
APES	84.78	84.62	84.85	79.17	83.33	75.00
APGOV	50.00	44.44	52.63	42.11	44.44	40.00
Charter school						
Overall	4.55	4.76	4.44	1.19	0.00	2.27
APES	6.52	7.69	6.06	2.08	0.00	4.17
APGOV	0.00	0.00	0.00	0.00	0.00	0.00
School proportion taking AP exams						
Overall	23.53	22.47	24.02	23.32	23.11	23.51
APES	22.40	21.41	22.79	20.66	22.38	18.95
APGOV	26.48	22.68	28.28	26.00	22.68	28.99
School enrollment counts						
Overall	1832.53	2046.38	1732.73	2007.37	2126.75	1898.84
APES	1765.02	1964.85	1686.30	1905.08	2092.00	1718.17
APGOV	2036.68	2213.00	1953.16	2161.76	2213.00	2115.65

Teacher-level descriptive statistics

Table Y2: Year Two (2017-18) teacher-level baseline (2015-16) characteristics overall and by experimental treatment and non-experimental control status, across and by courses, unmatched and matched.

Variable	Overall teachers across courses	Matched teachers across courses	Overall teachers APES	Matched teachers APES	Overall APGOV	Matched teachers APGOV
Counts						
Overall	90	46	48	27	42	19
Treatment	23	22	13	12	10	10
Non-exp. control	67	24	35	15	32	9
Female						
Overall	46.48	52.44	54.05	54.35	38.24	50.00
Treatment	52.17	54.55	53.85	58.33	50.00	50.00
Non-exp. control	43.75	50.00	54.17	50.00	33.33	50.00
Teacher baseline: average class size						
Overall	28.85	30.22	30.08	32.25	27.45	27.78
Treatment	29.46	29.98	31.56	32.69	26.72	26.72
Non-exp. control	28.65	30.45	29.53	31.80	27.68	28.84
Teacher average: student grade level						

Overall	11.48	11.54	11.30	11.40	11.68	11.72
Treatment	11.48	11.50	11.33	11.35	11.67	11.67
Non-exp. control	11.48	11.59	11.28	11.45	11.69	11.76
Teacher baseline: % of students taking exam						
Overall	90.91	88.47	86.65	82.84	95.78	95.24
Treatment	88.65	89.15	84.01	84.53	94.69	94.69
Non-exp. control	91.69	87.80	87.63	81.15	96.12	95.78
Teacher baseline: % of students earned qualifying score (full sample)						
Overall	46.27	34.05	28.75	14.48	66.30	57.54
Treatment	28.83	29.89	12.45	13.02	50.14	50.14
Non-exp. control	52.26	38.21	34.81	15.94	71.35	64.94
Teacher average: standardized national Math test						
Overall	0.22	0.05	-0.22	-0.40	0.72	0.58
Treatment	-0.00	0.04	-0.29	-0.25	0.38	0.38
Non-exp. control	0.29	0.06	-0.20	-0.55	0.83	0.78
Teacher average: standardized national ELA test						
Overall	0.25	0.12	-0.18	-0.32	0.75	0.66
Treatment	0.05	0.09	-0.28	-0.23	0.47	0.47
Non-exp. control	0.33	0.15	-0.14	-0.42	0.84	0.84
Teacher average: standardized state Math test						
Overall	317.73	311.85	305.34	298.22	331.88	328.20
Treatment	310.64	311.34	301.12	301.60	323.03	323.03
Non-exp. control	320.16	312.36	306.91	294.84	334.65	333.38
Teacher average: standardized state ELA test						
Overall	291.61	287.89	284.67	280.53	299.55	296.73
Treatment	287.06	287.40	281.70	281.88	294.03	294.03
Non-exp. control	293.17	288.38	285.77	279.18	301.27	299.43

Teacher average: standardized state science test						
Overall	176.34	172.60	167.26	162.02	186.73	185.29
Treatment	171.43	172.18	163.45	164.16	181.80	181.80
Non-exp. control	178.03	173.02	168.67	159.89	188.27	188.78
Teacher average: % students taking AP exam prior year						
Overall	59.88	62.84	52.11	55.56	68.75	71.56
Treatment	58.21	60.60	52.32	56.21	65.87	65.87
Non-exp. control	60.45	65.07	52.03	54.91	69.66	77.26
Teacher average: % of students with economic disadvantage						
Overall	48.76	52.75	66.64	71.46	28.32	30.29
Treatment	50.25	49.00	66.53	65.60	29.09	29.09
Non-exp. control	48.25	56.49	66.68	77.33	28.08	31.48
Teacher average: % of female students						
Overall	57.47	59.11	60.15	62.57	54.41	54.95
Treatment	56.18	56.71	57.26	58.33	54.78	54.78
Non-exp. control	57.92	61.50	61.23	66.82	54.30	55.12
Teacher average: % of Asian students						
Overall	16.57	12.12	11.22	8.78	22.69	16.13
Treatment	11.67	11.70	10.77	10.74	12.84	12.84
Non-exp. control	18.26	12.55	11.39	6.82	25.77	19.43
Teacher average: % of Hispanic students						
Overall	34.23	42.87	52.44	64.66	13.41	16.74
Treatment	41.40	41.77	56.34	58.26	21.99	21.99
Non-exp. control	31.76	43.98	50.99	71.05	10.73	11.49
Teacher average: % of White students						
Overall	32.08	30.74	20.10	14.24	45.77	50.53
Treatment	33.46	34.73	20.42	21.66	50.42	50.42
Non-exp. control	31.60	26.74	19.98	6.83	44.32	50.64

Student-level descriptive statistics

Table Y3: Year Two (2017-18) student characteristics overall and by experimental treatment and non-experimental control status, across and by course, unmatched and matched

Variable	Overall across courses	Matched across courses	Overall APES	Matched APES	Overall APGOV	Matched APGOV
Counts						
Overall	5,284	2,315	2,878	1,420	2,406	895
Treatment	1,186	1,168	675	657	511	511
Non-exp. control	4,098	1,147	2,203	763	1,895	384
Economic disadvantage						
Overall	41.78	50.22	56.91	66.81	23.69	27.31
Treatment	46.63	46.15	61.48	61.04	27.01	27.01
Non-exp. control	40.38	54.88	55.50	73.02	22.80	27.70
Female						
Overall	57.81	58.79	60.59	62.44	54.49	53.75
Treatment	56.66	56.85	57.48	57.84	55.58	55.58
Non-exp. control	58.14	61.01	61.54	67.38	54.20	51.47
Grade level						
Overall	11.45	11.49	11.28	11.38	11.66	11.65
Treatment	11.45	11.46	11.36	11.37	11.57	11.57
Non-exp. control	11.45	11.53	11.26	11.39	11.68	11.75
Asian						
Overall	17.78	12.64	13.66	10.37	22.71	15.78
Treatment	14.33	14.38	13.48	13.55	15.46	15.46
Non-exp. control	18.78	10.64	13.72	6.95	24.66	16.18
Hispanic						
Overall	30.49	43.79	44.67	62.16	13.52	18.44
Treatment	39.21	39.30	52.15	52.66	22.11	22.11
Non-exp. control	27.96	48.95	42.38	72.36	11.20	13.85
Black						
Overall	9.60	7.73	9.04	7.41	10.27	8.16
Treatment	7.34	6.68	7.11	5.94	7.63	7.63
Non-exp. control	10.25	8.93	9.63	8.99	10.98	8.82
White						
Overall	36.79	30.99	27.94	16.52	47.36	50.98
Treatment	33.98	34.42	22.81	23.29	48.73	48.73
Non-exp. control	37.60	27.07	29.52	9.24	47.00	53.80
Standardized national Math test						

Overall	0.35	0.05	0.00	-0.32	0.77	0.56
Treatment	0.08	0.09	-0.12	-0.10	0.34	0.34
Non-exp. control	0.43	-0.00	0.04	-0.56	0.88	0.84
Standardized national ELA test						
Overall	0.38	0.11	0.05	-0.25	0.78	0.61
Treatment	0.09	0.11	-0.13	-0.11	0.38	0.38
Non-exp. control	0.47	0.11	0.10	-0.40	0.89	0.89
Standardized state Math test						
Overall	321.59	311.59	312.07	299.97	332.97	327.62
Treatment	311.32	311.57	303.35	303.57	321.85	321.85
Non-exp. control	324.56	311.61	314.75	296.10	335.97	334.85
Standardized state ELA test						
Overall	294.21	288.09	289.35	282.20	300.03	296.23
Treatment	287.52	287.64	283.98	284.10	292.19	292.19
Non-exp. control	296.15	288.61	291.00	280.14	302.14	301.30
Standardized state Science test						
Overall	178.72	172.59	171.35	163.39	187.54	185.31
Treatment	172.53	172.80	166.10	166.40	181.02	181.02
Non-exp. control	180.52	172.36	172.96	160.15	189.30	190.67
Took AP exam in prior year						
Overall	63.91	64.00	58.48	57.23	70.41	73.34
Treatment	61.64	62.50	57.33	58.75	67.32	67.32
Non-exp. control	64.57	65.72	58.83	55.60	71.24	80.88

Table Y4: Counts of 2017-18 students per section, sections per teachers, and teachers per school, overall and by course, for non-experimental teachers, unmatched and matched

Count		Non-Experimental	Matched Non-Experimental
Students per section			
	Overall	28.04 (7.05)	28.62 (8.07)
	APES	28.19 (8.87)	29.74 (10.35)
	APGOV	27.87 (4.41)	27.20 (3.33)
Sections per teacher			

		2.10	1.55
	Overall	(1.20)	(0.70)
		2.09	1.58
	APES	(1.40)	(0.79)
		2.13	1.50
	APGOV	(0.94)	(0.63)
Teachers per school			
		1.49	1.59
	Overall	(0.89)	(0.91)
		1.33	1.00
	APES	(0.74)	(0.00)
		2.16	2.30
	APGOV	(1.07)	(0.95)

Appendix Z: Methods for Selecting Non-Experimental Comparison Teachers—Second Round

We used a second approach to inform our non-experimental study of KIA teachers' students' performance two years after their original offer to participate in the KIA intervention: The construction of a counterfactual condition using statistical matching procedures on propensity scores to balance the distribution of measured school-, teacher- and student-level covariates, pre-treatment and invariant—that is, those unaffected by treatment assignment, such as race/ethnicity, gender, eligibility for free or reduced-price lunch, and prior academic performance—between experimental (T) and non-experimental conditions (NE).

Once again, we limited our sample to include T2 and N2 teachers who taught APGOV or APES in 2016-17 and 2017-18, and who were present in the school system between 2015-16 to 2017-18, excluding non-experimental teachers who did not teach APGOV or APES in both years. This second time, however, we permitted inclusion of non-experimental teachers who had taught in the same school as experimental teachers. Another difference between our first and second rounds was the incorporation of student-level covariate information into the matching algorithm, using it to assess baseline equivalence between experimental and matched non-experimental teachers.

We then constructed a comparison group of non-experimental teachers who did not participate in the RCT and did not receive KIA professional development or support. Table Z1 lists the criteria used to determine experimental and non-experimental teachers' eligibility for inclusion. After applying the study's exclusion criteria, remaining were 106 non-experimental and 23 treatment teachers.

Table Z1: Study eligibility criteria, and sample loss and remaining eligibility counts, by non-experimental study condition

Group	Non-experimental teachers	Treatment teachers	Total teachers
All teachers who taught APGOV or APES for at least one year between 2015-16 and 2017-18	260	43	303
Not excluded by any condition	106	23	129
Excluded teachers by reason			
Deterministically randomized school	1	2	3
Principal denied consent	0	1	1
Non-eligible course	4	5	9
Did not teach in 2016-17	99	2	106
Did not teach in 2017-18	90	17	110
Moved schools in 2017-18	1	1	2
Total excluded teachers	154	20	174

Note. Totals in the far-right column are sums by row. The bottom row totals equal the first minus second rows. Exclusion reason was not mutually exclusive since a teacher may have met more than one exclusion criteria. Thus the sum of teachers disqualified from inclusion by exclusion reason is not equal to the total number of excluded teachers.

Matching methodology

Paralleling the baseline covariates used in the experimental arm of the study, covariates available for inclusion in the propensity score model included teacher-level baseline year (2015-16) average student performance on the outcomes of interest: AP exam-taking rates and scores. While pre-test information can substantially reduce bias in the estimated treatment effect in non-experimental studies[31], we supplemented these data with additional school and teacher covariates potentially related to teachers' KIA participation and students' AP performance outcomes.

Our selection of covariates for inclusion in the propensity score model reflected two sources of anticipated imbalance on measured covariates between T2 and N2 teachers. The first originates from teacher selection into KIA during the spring and summer of the 2016 school year, two years prior to measurement of student outcomes. Factors that may be related to teachers' decision to consent to participate in the KIA study, and for which we received data from participating districts, primarily include their 2015-16 (baseline) APGOV/APES students' AP exam-taking rates and scores.[32] The second source includes differences between T2 and N2 teachers' 2017-18 students, and may reflect student selection into a KIA course. To account for both sources of potential bias, we include teachers' 2017-18 students' baseline covariates in the propensity score model.

The underlying process determining teacher and student selection into KIA is unknown, and whether the assumptions[33] underpinning matching methods for generating credible causal estimates of the impact of KIA on student outcomes are met—notably, that conditional on pre-treatment covariates, summarized by the propensity score, the potential outcomes between KIA and matched non-KIA sample are equivalent—is indeterminate.[34] Furthermore, because of the small sample size and the number of available covariates eligible for estimating the propensity score, we confronted a trade-off between maximizing the number of covariates included in the propensity score model, and the number of T2 and N2 teachers who could be successfully matched without endangering the covariate balance

[31] For example, see Wong, Valentine, and Miller-Bains (2016) for an overview of the literature on the importance of pre-test information for reducing bias in observational studies of educational interventions.
[32] It is important to note that teacher-level baseline pre-test data are only available for a single school year, 2015-16. If selection into KIA is influenced by dynamic factors—or other non-dynamic factors not included in the vector of covariates used to impose balance between conditions that are also correlated with the outcome measure, the model will be misspecified and the strong ignorability assumption will be violated.
[33] The risk of a SUTVA violation is minimal since we disqualified from this study non-experimental teachers who were teaching the targeted course in schools with a KIA teacher.
[34] This is closely related to the notion of the "propensity score tautology", where the efficacy of matching based on the propensity score is assessed based on the balance of raw covariates, irrespective of whether this balance truly achieves the requirements necessary to satisfy the unconfoundedness assumption (Ho et al., 2007).

we sought through matching.[35] This trade-off occurred irrespective of the algorithm used to select matched controls with the propensity score.

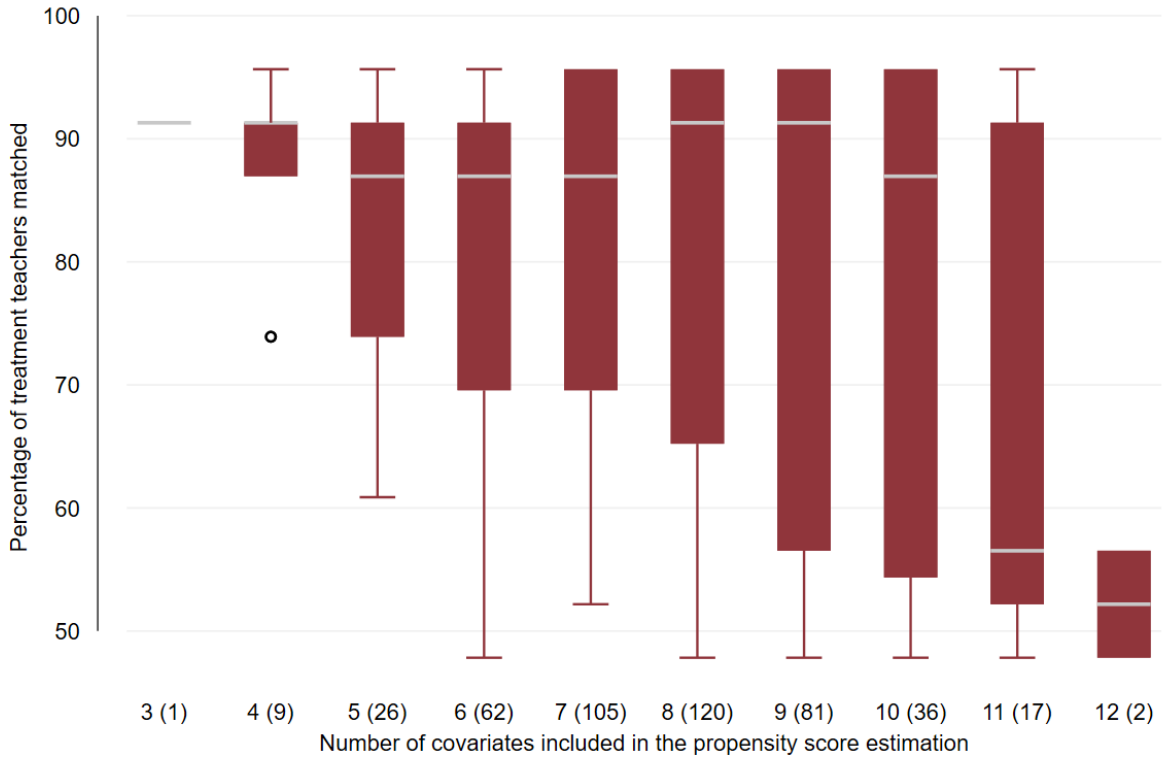
We illustrate the trade-off in Figure Z1, which displays the relationship between the number of covariates included in a given propensity score model and the percentage of treatment teachers who matched to at least one control teacher. The results are derived from approximately 450 unique matching solutions, with each varying the number of covariates[36] included, and the matching algorithm. The matching algorithm yielded 2,048 unique models, and the tuning parameters included:

1. Covariates, which included 1,024 unique covariate combinations, ranging from a minimum of three covariates, to a maximum of 13
2. Caliper width: no caliper and .4 standard deviations of the logit of the propensity score
3. With control teacher replacement
4. Up to three nearest neighbors

The dimensionality of the functional form for estimating the propensity score is negatively related to the number of treatment teachers matched to a control. For instance, the median number of treatment teachers successfully matched to a single control teacher was 80%, compared to 52% for models with 12 covariates.[37] This pattern persists across the different matching algorithms, which are not shown, but are available upon request from the authors.

[35] We attempted to reduce dimensionality by combining some related covariates, including combining the national college readiness assessments into a single variable, and creating a single measure of racial composition.
[36] The total number of covariates ranged between three and 13. All three baseline covariates—which included two teacher-level pre-test covariates (percentage of students who took the respective AP exam and the average APES/APGOV exam score) from the 2015-16 school year—were included in all permutations because of the proximity of the measure to the teacher's decision to participate in KIA, the magnitude of the baseline imbalance between non-experimental and treatment teachers on these covariates, and the research team's qualitative and theoretical understanding of the factors that influenced participation in KIA. Additionally, because of the importance of these baseline measures, no matching solution was included in the universe of acceptable matches if the absolute standardize difference between conditions was greater than or equal to .25 standard deviations.
[37] No acceptable matching solution was found for models that included all 13 covariates.

Figure Z1: Relationship between the number of covariates included in the propensity score model and the percentage of matched treatment teachers



Note. Underlying data were obtained from 2,048 unique matching permutations with different covariates and calipers. Box height reflects the inter-quartile range, and the horizontal white bar represents the median percentage of matched treatment teachers. The whiskers represent extreme values that are 1.5 times greater than the upper and lower quartiles, and the dots reflect outliers beyond these ranges. The x-axis values are the number of covariates included in the model fit to estimate the propensity score. Adjacent values in parentheses indicate the total number of acceptable matching solutions that met the eligibility threshold.

While our primary objective with matching was maximizing the equivalence of the joint distribution of the covariates available to the research team, we also wanted to maximize the number of teachers included in the analysis. This was important for statistical power; that is, our ability to detect a difference between KIA and non-experimental teachers, if one exists. It also was important as related to generalizability because of the small number of KIA teachers in some districts who remained in the analytic sample after the sample inclusion rules were applied, as well as the lack of overlap in conditional probabilities of participating between the experimental and non-experimental conditions. Thus, to balance the objectives of maximizing equivalence and sample size, we selected the matching solution that maximized the number of treatment teachers matched to a non-experimental teacher, and for which all covariates included in the propensity score specification exhibited an absolute standardized difference of less than or equal to .25.

Full set of covariates considered for balancing

Table Z2 lists the school-, student-, and teacher-level covariates available to the research team. We forced for inclusion three teacher-level, baseline covariates in all iterations of propensity score estimations: average APES/APGOV; the percentage of teachers' 2015-16 students who took APES/APGOV exam; and teachers' average class size in 2015-16. We also required matches to exhibit imbalance of less than 0.25 absolute standard deviations. (For details about the transformation of all variables measuring student achievement, see Appendix E.)

Table Z2. Characteristics of non-experimental and experimental teachers in the full sample prior to matching

Level	Variable	Treatment	Non-experimental	Raw mean difference
Student	Economic disadvantage	46.63	40.60	6.03
Student	Female	56.66	57.17	-0.51
Student	Grade level	11.45	11.49	-0.04
Student	Asian	14.33	16.73	-2.40
Student	Hispanic	39.21	28.05	11.16
Student	Black	7.34	10.10	-2.76
Student	White	33.98	39.70	-5.72
Student	Standardized national Math test	0.08	0.40	-0.32
Student	Standardized national ELA test	0.09	0.42	-0.33
Student	Standardized state Math test	311.32	323.35	-12.03
Student	Standardized state ELA test	287.52	294.95	-7.43
Student	Standardized state Science test	172.53	179.94	-7.41
Student	Took AP test in prior year	61.64	63.42	-1.78
Teacher	Teacher average: standardized national Math test	-0.00	0.27	-0.27
Teacher	Teacher average: standardized national ELA test	0.05	0.29	-0.24
Teacher	Teacher average: standardized state Math test	310.64	318.97	-8.33
Teacher	Teacher average: standardized state ELA test	287.06	292.13	-5.07
Teacher	Teacher average: standardized state science test	171.43	176.90	-5.47
Teacher	Teacher average: percent of students with economic disadvantage	50.25	48.75	1.50
Teacher	Teacher average: percent of female students	56.18	57.20	-1.02
Teacher	Teacher average: student grade level	11.48	11.51	-0.03
Teacher	Teacher average: percent of Asian students	11.67	15.70	-4.03

Teacher	Teacher average: percent of Hispanic students	41.40	33.40	8.00
Teacher	Teacher average: percent of Hispanic students	8.62	13.03	-4.41
Teacher	Teacher average: percent of White students	33.46	32.67	0.79
Teacher	Years teaching APES/APGOV	6.78	NA	NA
Teacher	Female	52.17	53.66	-1.49
Teacher	Teacher baseline: percent of students earning credit	28.83	46.71	-17.88
Teacher	Teacher baseline: percent of students taking test	88.65	90.53	-1.88
Teacher	Teacher baseline: average class size	29.46	29.00	0.46
Teacher	Teacher average: percent of students taking AP test prior year	58.21	60.61	-2.40
School	School percentage LEP	11.44	9.03	2.41
School	School students taking AP exams	416.06	476.90	-60.84
School	School enrollment	1929.29	1839.50	89.79
School	School student-to-teacher ratio	19.78	19.04	0.74
School	School percentage FRPL	58.52	57.61	0.91
School	Title 1 school	70.59	55.71	14.88
School	Urban school	70.59	70.00	0.59
School	Charter school	5.88	2.86	3.02
School	Magnet school	41.18	35.71	5.47
School	School proportion taking AP exams	24.59	25.87	-1.28

Missing pre-treatment covariates

Baseline data are incomplete for some T2 and N2 teachers. We imputed missing covariates using the multiple imputation method described in Appendix J.

Estimating the Propensity Score and Selecting Matched Comparison Teachers

To select a comparison group of teachers who did not consent to participate in the KIA RCT but who, conditioned on measured pre-treatment covariates, had similar probabilities of participating in KIA as a T2 teacher, we used 3:1 nearest neighbor matching (NNM) to select up to three nonexperimental teachers using a "greedy" matching algorithm. We selected nearest neighbors using the logit of the propensity score estimated from a logistic regression. All analyses were performed at the teacher level. The model iteratively regressed a binary indicator of KIA participation ("1" for T2 teachers, "0" for N2 teachers) on all unique combinations of covariates derived from the full list of covariates detailed in Table Z2, although all models included the three starred baseline covariates.

Further, to optimize both the number of covariates for which adequate balance was achieved and the number of teachers successfully matched, for each covariate combination we produced matching solutions for several different matching algorithms and tuning parameters. In addition, we used exact matching on two covariates: district and course, and common support was enforced on all acceptable matching solutions.

We implemented the approach in three stages. Following the procedures and notation of Becker and Ichino (2002):

Step 1: Fit a logistic regression:

$$\Pr(D_i = 1 \mid X_i) = \Phi\{h(X_i)\} \quad (1)$$

where Φ is the propensity score, and $h(X_i)$ is a vector of pre-treatment (2015-16 teacher and 2017-18 teachers' students' pre-intervention data) teacher and school covariates. We fit models that included all combinations of covariates itemized in Table W2, forcing the inclusion of the three starred teacher-level baseline covariates in all fitted models.

Step 2: Find the nearest three non-experimental neighbors for T2 teacher within a caliper of *S* standard deviations of the logit of the propensity score, which included 0 and .4 standard deviations, where a neighbor within the same district and who taught the same subject was selected irrespective of the difference in propensity scores.

$$C(i) = \left\| p_i - p_j \right\| \leq c_{s\sigma} \quad (2)$$

We selected the non-treated units (*j*) that satisfy the condition $(i) = \left\| p_i - p_j \right\| \leq c_{s\sigma}$. In other words, we picked the N2 teachers with the smallest logit within $c_{s\sigma}$ of each T2 teacher. We performed Step 2 separately by block (district) and by course, and selected non-experimental counterparts from within the same district and course to maximize balance on covariates, both observed and unobserved, varying across geography and districts. This procedure mimics the RCT experimental approach, which used a block-randomized design with school-level randomization by district, and is consistent with a large literature on the non-experimental design features that yield unbiased causal effects validated against experimental estimates. All matching performed was blind to the outcome data, and the impact of a given matching solution on outcome differences between T and N conditions. The resulting matching solution only advanced to Step 3 if the absolute difference in the standard deviation of the propensity score of each baseline covariate was less than or equal to 0.10 standard deviations of the logit of the propensity score. From the full universe of 2,048 potential matching configuration, 459 matching solutions progressed to Step 3.

Step 3: To select a single matching solution from the approximately 2,000 matching solutions satisfying the above criteria, we selected the matching solution that maximized the number of treatment teachers successfully matched to a non-experimental teacher, and for which all the absolute standardized difference between experimental and matched non-experimental teachers for all included covariates (including the covariates forced for inclusion) was less than or equal to .25ASD. Next, for these matches, we linked the teacher-level data to the student-level data, and calculated the difference in the Mahalanobis distance between treatment and matched non-experimental teachers for all available

covariates—even covariates not included in the teacher-level model fit to estimate the propensity score. This was done to account for differences in the number of students assigned to teachers, and to align with the level of analysis (student) of the Maturation Study impact model. We selected the matching solution with the minimum absolute difference in the Mahalanobis distance metric between treatment and non-experimental teachers at the student level. The selected matching solution parameters are itemized below, and we summarize matching diagnostics in the subsequent section.

1. No caliper
2. Three nearest-neighbors
    a. 22 of 23 treatment teachers were matched
    b. 44 four of 106 control teachers were matched
3. With replacement
4. Four covariates

Figure Z2 provides the counts of control teachers matched to at least one treatment teacher, by course and district. Values on the y-axis, which range between 1 and 4, reflect the number of unique treatment teachers to which a control teacher was matched. For instance, of the 15 control teachers in District E successfully matched to at least one treatment teacher, nine were matched only once, while six were matched to two or more treatment teachers.

Figure Z2: Distribution of matched non-experimental control teachers, by course, district, and number of treatment teacher matches

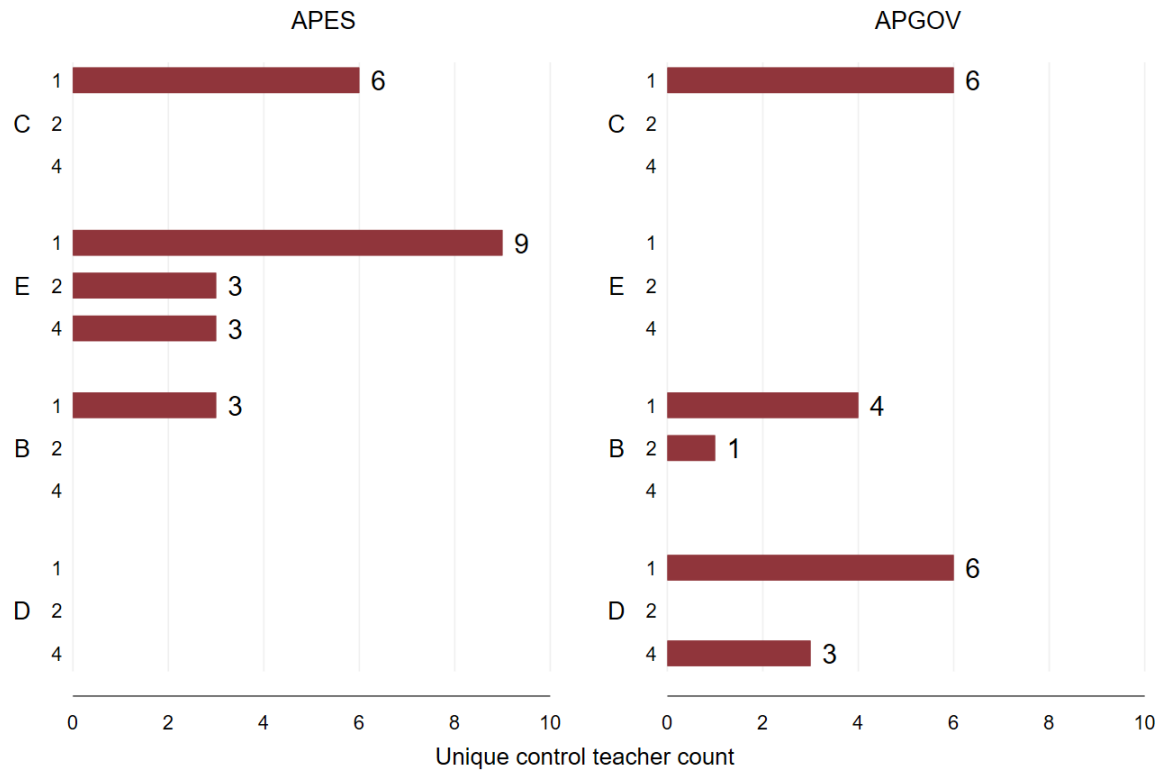
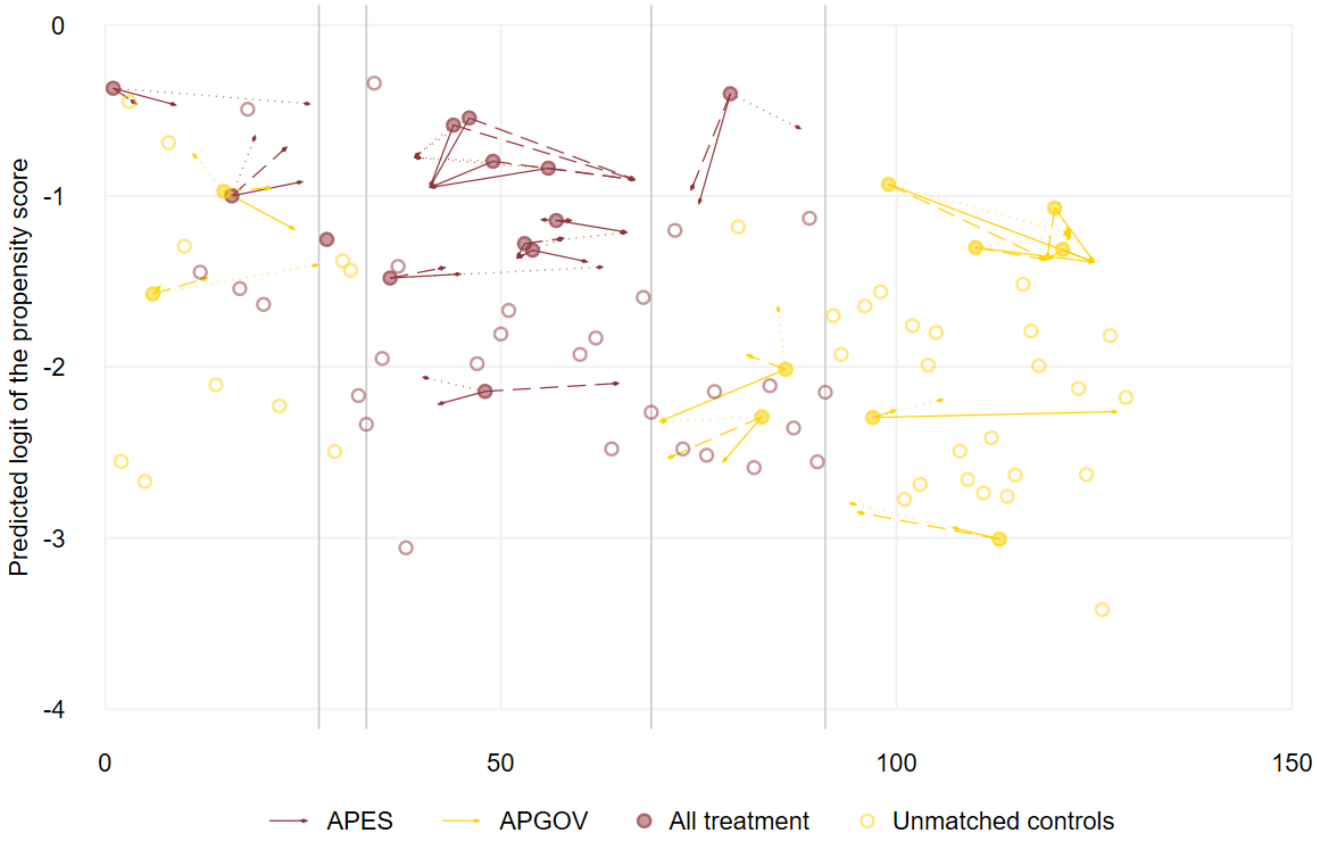


Figure Z3's paired-plot illustrates the distances (based on the logit of the propensity score) between each treatment teacher (22) and matched control (44), by course and by district. The filled circles at the base of each arrow denote treatment teachers, and their position on the y-axis reflects their estimated logit. The arrow barb is the predicted logit of the treatment teacher's matched pair, and the arrow's slope captures the absolute distance in the estimated logits of the propensity score between each match. Hollow circles indicate unmatched control teachers, while filled circles not linked by an arrow are unmatched treatment teachers. This figures further illustrates that, while the overall balance on the covariates of interest was within a tolerable range (demonstrated in Figure Z5), markedly varying across district and course was the quality of matches, according to the absolute distance in the logit of the propensity score between a treatment teacher and matched controls.

Figure Z3: Distance between treatment teachers and matched control teachers based on propensity scores, by course and district



Note. The x-axis contains a synthetic teacher identification number. Vertical gray lines demarcate unique districts, and the horizontal space between successive lines convey the number of unique teachers within a given district.

Figure Z4 presents kernel density estimates of the propensity scores' predicted logits from the final matching solution for four groups: all 23 treatment and 106 control teachers, and the subsets of matched treatment (22) and control (44) teachers. Enforcing common support served to trim treatment and control teachers with estimated propensity scores below -.3 logits.

Figure Z4: Overlap between treatment and control teachers, by matching status

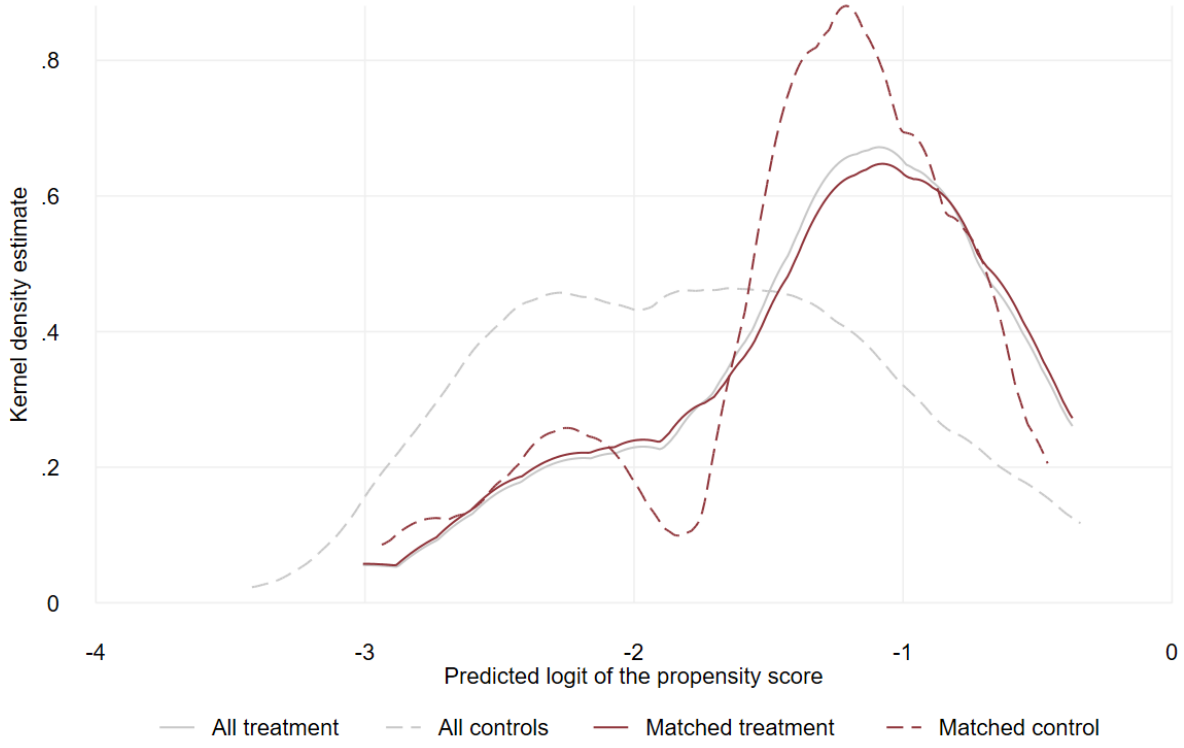
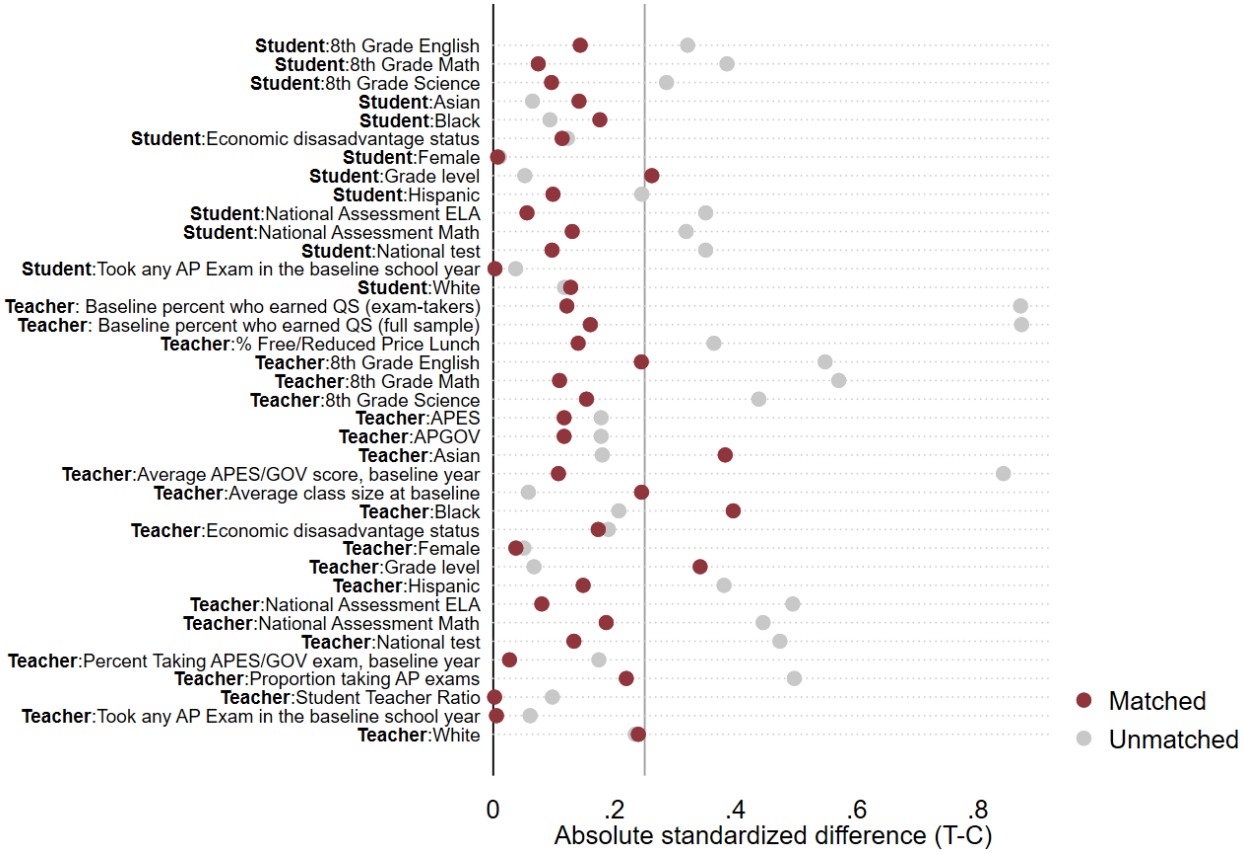


Figure Z5 displays the baseline equivalence across all available covariates before and after matching.

Figure Z5: Covariate balance between treatment and control teachers, matched and unmatched



Note. Covariates with the **Student** prefix indicate covariates calculated and weighted at the student-level. Covariates with the **Teacher** prefix denote covariates calculated and weighted at the teacher-level.

Appendix AA: Unmatched and Matched Non-experimental School-, Teacher-, and Student-Level Descriptive Statistics Compared to Experimental Treatment—Second Round

School-level descriptive statistics

Table AA1: Year Two (2017-18) school-level baseline (2015-16) characteristics overall and by experimental treatment and non-experimental control status, across and within courses, unmatched and matched

		Overall Unmatched	Treatment Unmatched	Non-Exp. Control Unmatched	Matched Overall	Matched Treatment	Matched Control
Counts							
	Overall	87	21	72	57	20	38
	APES	58	13	46	35	12	23
	APGOV	43	9	37	25	9	16
School percentage LEP							
	Overall	9.50	11.54	9.01	11.83	10.66	13.04
	APES	9.87	11.32	9.49	12.57	10.68	14.42
	APGOV	7.27	10.18	6.80	9.82	10.67	8.89
School enrollment							
	Overall	1857.05	1930.65	1839.17	1988.61	2134.34	1838.40
	APES	1738.78	1921.17	1691.20	1773.92	2122.31	1434.25
	APGOV	2093.14	1798.33	2140.95	2210.45	2210.21	2210.71
School student-to-teacher ratio							
	Overall	19.18	19.62	19.08	19.20	19.32	19.08
	APES	20.91	21.77	20.68	21.58	22.48	20.70
	APGOV	16.58	15.68	16.73	16.02	15.50	16.58
School percentage FRPL							
	Overall	57.79	58.70	57.57	59.10	55.40	62.92
	APES	69.13	66.47	69.83	74.82	70.23	79.30
	APGOV	40.49	44.36	39.87	39.05	38.26	39.92
Title 1 school							
	Overall	58.62	70.59	55.71	56.82	53.73	60.00
	APES	79.31	91.67	76.09	87.34	84.62	90.00
	APGOV	23.26	16.67	24.32	12.31	8.82	16.13
Urban school							
	Overall	70.11	76.47	68.57	63.64	64.18	63.08
	APES	86.21	83.33	86.96	82.28	76.92	87.50
	APGOV	53.49	50.00	54.05	38.46	38.24	38.71
Charter school							
	Overall	3.45	5.88	2.86	1.52	0.00	3.08
	APES	5.17	8.33	4.35	2.53	0.00	5.00
	APGOV	0.00	0.00	0.00	0.00	0.00	0.00
Magnet school							
	Overall	36.78	41.18	35.71	37.12	46.27	27.69
	APES	41.38	41.67	41.30	45.57	46.15	45.00
	APGOV	27.91	50.00	24.32	26.15	38.24	12.90
School proportion taking AP exams							

Overall	25.62	22.90	26.28	24.57	22.46	26.76
APES	24.57	21.87	25.28	23.24	21.47	24.97
APGOV	27.79	25.77	28.12	24.60	21.49	28.00

Teacher-level descriptive statistics

Table AA2: Year Two (2017-18) teacher-level baseline (2015-16) characteristics overall and by experimental treatment and non-experimental control status, across and by courses, unmatched and matched.

Variable	Overall Unmatched	Matched Overall	Overall unmatched APES	Matched APES	Overall unmatched APGOV	Matched APGOV
Counts						
Overall	129	66	65	36	64	30
Treatment	23	22	13	12	10	10
Control	106	44	52	24	54	20
Teacher average: standardized national Math						
Overall	0.22	-0.04	-0.17	-0.41	0.61	0.40
Treatment	-0.00	0.04	-0.29	-0.25	0.38	0.38
Control	0.27	-0.12	-0.14	-0.57	0.65	0.42
Teacher average: standardized national ELA						
Overall	0.24	0.02	-0.15	-0.37	0.64	0.50
Treatment	0.05	0.09	-0.28	-0.23	0.47	0.47
Control	0.29	-0.04	-0.12	-0.52	0.68	0.53
Teacher average: standardized state Math						
Overall	317.49	309.00	306.02	297.05	329.14	323.34
Treatment	310.64	311.34	301.12	301.60	323.03	323.03
Control	318.97	306.66	307.25	292.51	330.27	323.65
Teacher average: standardized state ELA						
Overall	291.23	285.19	285.35	278.52	297.20	293.20
Treatment	287.06	287.40	281.70	281.88	294.03	294.03
Control	292.13	282.99	286.26	275.17	297.79	292.37
Teacher average: standardized state science						
Overall	175.92	169.98	167.61	160.86	184.36	180.92
Treatment	171.43	172.18	163.45	164.16	181.80	181.80
Control	176.90	167.78	168.65	157.56	184.84	180.04
Teacher average: % of students with economic disadvantage						
Overall	49.02	53.90	68.62	72.54	29.12	31.54
Treatment	50.25	49.00	66.53	65.60	29.09	29.09
Control	48.75	58.81	69.14	79.49	29.13	33.99
Teacher average: % of female students						
Overall	57.02	56.91	59.53	58.61	54.47	54.87
Treatment	56.18	56.71	57.26	58.33	54.78	54.78

Control	57.20	57.11	60.10	58.90	54.42	54.96
Teacher average: student grade level						
Overall	11.50	11.55	11.32	11.37	11.69	11.77
Treatment	11.48	11.50	11.33	11.35	11.67	11.67
Control	11.51	11.61	11.32	11.39	11.69	11.87
Teacher average: % of Asian students						
Overall	14.98	10.44	9.32	7.36	20.73	14.13
Treatment	11.67	11.70	10.77	10.74	12.84	12.84
Control	15.70	9.19	8.96	3.98	22.19	15.43
Teacher average: % of Hispanic students						
Overall	34.83	43.68	51.85	63.65	17.54	19.73
Treatment	41.40	41.77	56.34	58.26	21.99	21.99
Control	33.40	45.60	50.72	69.03	16.72	17.48
Teacher average: % of Hispanic students						
Overall	12.24	10.79	13.02	9.89	11.45	11.87
Treatment	8.62	6.74	8.38	4.91	8.94	8.94
Control	13.03	14.85	14.19	14.88	11.91	14.80
Teacher average: % of White students						
Overall	32.81	30.14	21.41	14.31	44.39	49.13
Treatment	33.46	34.73	20.42	21.66	50.42	50.42
Control	32.67	25.54	21.66	6.96	43.27	47.84
Years teaching APES/APGOV						
Overall	6.78	7.00	5.69	6.00	8.20	8.20
Treatment	6.78	7.00	5.69	6.00	8.20	8.20
Control	NA	NA	NA	NA	NA	NA
Female						
Overall	53.33	58.54	60.00	60.87	47.27	55.56
Treatment	52.17	54.55	53.85	58.33	50.00	50.00
Control	53.66	63.16	62.16	63.64	46.67	62.50
Teacher baseline: % of students earn qualifying scores (full sample)						
Overall	43.52	30.60	26.59	12.47	60.71	52.35
Treatment	28.83	29.89	12.45	13.02	50.14	50.14
Control	46.71	31.30	30.13	11.92	62.67	54.56
Teacher baseline: % of students taking AP exam						
Overall	90.20	89.69	85.73	85.43	94.74	94.79
Treatment	88.65	89.15	84.01	84.53	94.69	94.69
Control	90.53	90.23	86.16	86.34	94.74	94.90
Teacher baseline: average class size						
Overall	29.08	30.29	30.04	32.37	28.10	27.81
Treatment	29.46	29.98	31.56	32.69	26.72	26.72
Control	29.00	30.61	29.67	32.05	28.35	28.89

Teacher average: % students take any AP exam prior year						
Overall	60.19	59.70	52.85	52.75	67.63	68.05
Treatment	58.21	60.60	52.32	56.21	65.87	65.87
Control	60.61	58.80	52.99	49.28	67.96	70.22

Student-level descriptive statistics

Table AA3: Year Two (2017-18) student characteristics overall and by experimental treatment (and non-experimental control status, across and within courses, unmatched and matched

Variable		Overall unmatched	Matched Overall	Overall APES	Matched APES	Overall APGOV	Matched APGOV
Counts							
	Overall	7,744	3,407	3,825	1,772	3,919	1,635
	Treatment	1,186	1,168	675	657	511	511
	Control	6,558	2,239	3,150	1,115	3,408	1,124
Economic disadvantage							
	Overall	41.52	49.02	59.94	68.40	23.55	26.91
	Treatment	46.63	46.15	61.48	61.04	27.01	27.01
	Control	40.60	51.82	59.61	76.40	23.03	26.83
Female							
	Overall	57.09	57.03	59.75	58.33	54.49	55.54
	Treatment	56.66	56.85	57.48	57.84	55.58	55.58
	Control	57.17	57.20	60.24	58.87	54.33	55.51
Grade level							
	Overall	11.48	11.55	11.28	11.36	11.68	11.75
	Treatment	11.45	11.46	11.36	11.37	11.57	11.57
	Control	11.49	11.63	11.27	11.36	11.70	11.91
Asian							
	Overall	16.36	12.11	11.62	9.33	20.99	15.27
	Treatment	14.33	14.38	13.48	13.55	15.46	15.46
	Control	16.73	9.89	11.22	4.75	21.82	15.11
Hispanic							
	Overall	29.76	41.76	44.48	61.63	15.39	19.09
	Treatment	39.21	39.30	52.15	52.66	22.11	22.11
	Control	28.05	44.16	42.84	71.37	14.38	16.49
Black							
	Overall	9.67	9.36	9.26	8.11	10.08	10.79
	Treatment	7.34	6.68	7.11	5.94	7.63	7.63
	Control	10.10	11.97	9.72	10.47	10.45	13.50
White							
	Overall	38.83	31.44	29.99	15.62	47.45	49.50
	Treatment	33.98	34.42	22.81	23.29	48.73	48.73
	Control	39.70	28.54	31.53	7.28	47.26	50.16
Standardized national Math test							

Overall	0.35	0.02	0.03	-0.33	0.67	0.43
Treatment	0.08	0.09	-0.12	-0.10	0.34	0.34
Control	0.40	-0.04	0.06	-0.58	0.72	0.50
Standardized national ELA test						
Overall	0.37	0.08	0.05	-0.29	0.69	0.50
Treatment	0.09	0.11	-0.13	-0.11	0.38	0.38
Control	0.42	0.05	0.09	-0.48	0.73	0.60
Standardized state Math test						
Overall	321.50	310.37	312.20	298.50	330.58	323.92
Treatment	311.32	311.57	303.35	303.57	321.85	321.85
Control	323.35	309.21	314.10	293.00	331.89	325.70
Standardized state ELA test						
Overall	293.81	285.91	289.52	280.08	298.01	292.56
Treatment	287.52	287.64	283.98	284.10	292.19	292.19
Control	294.95	284.23	290.71	275.72	298.88	292.89
Standardized state Science test						
Overall	178.80	171.49	171.66	162.55	185.77	181.69
Treatment	172.53	172.80	166.10	166.40	181.02	181.02
Control	179.94	170.22	172.86	158.36	186.48	182.27
Took any AP exam in prior year						
Overall	63.15	62.57	57.28	54.73	68.87	71.52
Treatment	61.64	62.50	57.33	58.75	67.32	67.32
Control	63.42	62.64	57.27	50.36	69.10	75.13

Table AA4: Counts of students per section (standard deviations in parentheses), sections per teachers, and teachers per school, overall and by course, for unmatched and matched teachers

Count	Overall	Treatment	Non-Experimental Overall	Matched Non-Experimental
Students per section				
Overall	28.40	28.40	28.89	30.19
	(6.16)	(6.16)	(6.49)	(7.55)
APES	29.76	29.76	29.18	29.58
	(6.71)	(6.71)	(8.31)	(9.40)
APGOV	26.89	26.89	28.64	30.78
	(5.25)	(5.25)	(4.40)	(5.29)
Sections per teacher				
Overall	1.74	1.74	2.08	1.71
	(0.86)	(0.86)	(1.15)	(0.70)
APES	1.62	1.62	1.94	1.53
	(0.87)	(0.87)	(1.30)	(0.66)

	1.90	1.90	2.20	1.93
APGOV	(0.88)	(0.88)	(0.98)	(0.70)
Teachers per school				
	1.33	1.33	1.52	1.67
Overall	(0.59)	(0.59)	(0.80)	(0.87)
	1.31	1.31	1.42	1.28
APES	(0.63)	(0.63)	(0.69)	(0.46)
	1.63	1.63	1.91	2.30
APGOV	(0.52)	(0.52)	(0.92)	(0.99)

Notes: Section was missing for 253 District E students (30 teachers at 27 schools).

Appendix BB: Non-Experimental Student-Level Baseline Equivalence—First Round

WWC reviewers will likely evaluate our results on AP outcomes using the Transition to College review protocol, which specifies that baseline equivalence must be established on measures of prior achievement (we have eighth-grade standardized test scores and national assessments) and SES (we have an indicator for economically disadvantaged).

In the Appendix BB tables, we show standardized mean differences between 2017-18 students of Round One matched treatment (n=22) and nonexperimental control (n=24) students on all student-, teacher-, and school-level covariates, for outcomes as specified. For our AP outcome models, we include as a covariate the 2015-16 AP score average corresponding specifically to that outcome, as opposed to the average AP score and exam-taking rate for all outcomes. For this reason, there are blank cells in Table BB1 for covariates describing 2015-16 AP performance. Student-level prior achievement variables assume missing at randomness (MAR) in Table BB1, and bound according to potential deviations from MAR in Table BB2.

Table BB1: Baseline standardized mean differences between treatment and matched non-experimental teachers' 2017-18 students on all student-, teacher-, and school-level covariates, for Research Question Three Approach 2 Round One analytic AP outcome samples

	AP continuous scores (n=1309)	AP QS (exam-takers only) (n=1,900)	AP QS (full sample) and exam-taking (n=2,315)
National Assessment Math	0.129	-0.041	-0.010
National Assessment ELA	0.092	-0.072	-0.052
Eighth-grade Math	0.051	-0.070	-0.059
Eighth-grade English	0.004	-0.123	-0.093
Eighth-grade Science	0.030	-0.094	-0.083
Economically disadvantaged	-0.163	0.067	-0.051
Took any AP exam in 2016	0.053	-0.039	-0.049
Female	-0.204	-0.040	-0.070
Grade	-0.023	0.040	0.046
Asian	0.082	-0.012	0.074
Hispanic	0.000	0.000	-0.023
Black	-0.478	-0.167	-0.233
White	0.188	0.166	0.280
2016 average AP score		-0.357	

2016 % earned QS (exam-takers only)		-0.415	
2016 % earned QS (full sample)			-0.379
2016 % taking AP exam			0.166
2016 Average total fine-grained score	-0.596		
2016 Average MC fine-grained score	-0.464		
2016 average FR fine-grained score	-0.737		
Course: APGOV	-0.497	0.020	0.093
2015-16 average class size	-0.105	-0.122	-0.173
% free/reduced-price lunch	-0.362	-0.043	-0.040
Student-teacher ratio	0.007	-0.117	-0.132
% taking an AP exam	0.256	-0.206	-0.110
Average national Math	0.375	-0.052	0.010
Average national ELA	0.202	-0.160	-0.112
Average eighth-grade Math	0.074	-0.206	-0.158
Average eighth-grade ELA	0.064	-0.154	-0.167
Average eighth-grade Science	0.171	-0.104	-0.088
Proportion school low SES	-0.421	-0.132	-0.108
Proportion school taking any AP exam in 2016	-0.044	-0.253	-0.167
Proportion female	-1.083	-0.511	-0.602
Average grade	-0.078	-0.190	-0.152
Proportion Asian	0.382	0.107	0.116
Proportion Hispanic	-0.239	-0.128	-0.122
Proportion Black	-0.243	-0.186	-0.219
Proportion White	0.226	0.193	0.200

Table BB2: Baseline standardized mean differences between treatment and matched non-experimental teachers' 2017-18 students' imputed student-level covariates, for Research Question Three Approach 2 Round One analytic AP outcome samples

Took AP exam

	D1	D2	D3	D4	Most extreme
National Math	-0.010	-0.023	-0.030	-0.003	-0.030
National ELA	-0.052	-0.067	-0.076	-0.043	-0.076
Eighth-grade Math	-0.059	-0.212	-0.172	-0.100	-0.212
Eighth-grade ELA	-0.093	-0.161	-0.182	-0.072	-0.182
Eighth-grade Science	-0.083	0.071	-0.104	0.091	-0.104

AP Qualifying score outcome (full sample)

	D1	D2	D3	D4	Most extreme
National Math	-0.010	0.007	-0.025	0.021	-0.025
National ELA	-0.052	-0.037	-0.065	-0.024	-0.065
Eighth-grade Math	-0.059	-0.062	0.020	-0.141	-0.141
Eighth-grade ELA	-0.093	-0.095	-0.012	-0.176	-0.176
Eighth-grade Science	-0.083	-0.016	-0.038	-0.061	-0.083

AP Qualifying score outcome (exam-takers only)

	D1	D2	D3	D4	Most extreme
National Math	-0.041	-0.021	-0.057	-0.006	-0.057
National ELA	-0.072	-0.055	-0.085	-0.041	-0.085
Eighth-grade Math	-0.070	-0.081	0.017	-0.168	-0.168
Eighth-grade ELA	-0.123	-0.135	-0.027	-0.230	-0.230
Eighth-grade Science	-0.094	-0.037	-0.060	-0.072	-0.094

AP total score

	D1	D2	D3	D4	Most extreme
National Math	0.129	0.127	0.131	0.126	0.131
National ELA	0.092	0.090	0.093	0.089	0.093
Eighth-grade Math	0.051	0.007	0.118	-0.060	0.118
Eighth-grade ELA	0.004	-0.046	0.062	-0.104	-0.104
Eighth-grade Science	0.030	0.059	0.044	0.045	0.059

AP multiple-choice score

	D1	D2	D3	D4	Most extreme
National Math	0.129	0.133	0.127	0.135	0.135
National ELA	0.092	0.094	0.090	0.096	0.096
Eighth-grade Math	0.051	-0.010	0.119	-0.078	0.119
Eighth-grade ELA	0.004	-0.063	0.068	-0.128	-0.128
Eighth-grade Science	0.030	0.055	0.056	0.029	0.056

AP free-response score

	D1	D2	D3	D4	Most extreme
National Math	0.129	0.119	0.136	0.113	0.136
National ELA	0.092	0.084	0.097	0.079	0.097
Eighth-grade Math	0.051	0.033	0.124	-0.040	0.124
Eighth-grade ELA	0.004	-0.024	0.059	-0.079	-0.079
Eighth-grade Science	0.030	0.075	0.028	0.077	0.077

Appendix CC: Non-Experimental Baseline Equivalence Results, Second Round

In Appendix CC, we show standardized mean differences and associated p-values between 2017-18 students of Round Two matched treatment (n=22) and nonexperimental control (n=44) teachers' students on all student-, teacher-, and school-level covariates, for outcomes as specified (i.e., second-round matching sample). For our AP outcome models, we include as a covariate the 2015-16 AP score average corresponding specifically to that outcome, as opposed to the average AP score and exam-taking rate for all outcomes. For this reason, there are blank cells in Table CC1 for covariates describing 2015-16 AP performance. Student-level prior achievement variables assume missing at randomness (MAR) in Table CC1, and bound according to potential deviations from MAR in Table CC2.

Table CC1: Baseline standardized mean differences between treatment and matched non-experimental teachers' 2017-18 students on all student-, teacher-, and school-level covariates, for Research Question Three Approach 2 Round Two analytic AP outcome samples

	AP continuous scores (n=1,668)		AP QS (exam-takers only) (n=2,833)		AP QS (full sample) and exam-taking (n=3,407)	
	SMD	p-value	SMD	p-value	SMD	p-value
National Assessment Math	0.284	0.075	0.163	0.244	0.151	0.258
National Assessment ELA	0.348	0.040*	0.217	0.139	0.221	0.119
Eighth-grade Math	0.070	0.655	-0.007	0.958	-0.007	0.953
Eighth-grade English	0.094	0.553	0.055	0.688	0.065	0.620
Eighth-grade Science	0.150	0.322	0.081	0.531	0.103	0.415
Economically disadvantaged	-0.051	0.843	0.003	0.989	-0.085	0.669
Took any AP exam in 2016	0.326	0.190	0.155	0.569	0.129	0.611
Female	0.023	0.833	-0.020	0.763	0.013	0.841
Grade	0.770	0.000***	0.879	0.000***	0.782	0.000***
Asian	0.388	0.167	0.160	0.420	0.219	0.246
Hispanic	0.016	0.945	0.050	0.798	0.035	0.848
Black	-0.597	0.048*	-0.435	0.038*	-0.519	0.000***
White	0.374	0.115	0.331	0.048*	0.375	0.029*
2016 average AP score			-0.070	0.809		
2016 % earning QS (exam-takers)			-0.087	0.751		

2016 % earning QS (full sample)					-0.119	0.634
2016 % taking AP exam					0.003	0.989
2016 average total score	-0.103	0.791				
2016 average MC score	-0.042	0.913				
2016 average FRQ score	-0.172	0.649				
Course: APGOV	-0.282	0.544	-0.166	0.730	-0.142	0.767
2015-16 average class size	-0.160	0.618	-0.212	0.416	-0.231	0.372
% free/reduced-price lunch	-0.453	0.143	-0.193	0.286	-0.190	0.281
Student-teacher ratio	0.171	0.470	-0.005	0.971	-0.010	0.950
% taking an AP exam	-0.039	0.911	-0.327	0.311	-0.238	0.432
Average national Math	0.540	0.194	0.152	0.589	0.193	0.465
Average national ELA	0.396	0.329	0.060	0.821	0.089	0.725
Average eighth-grade Math	0.265	0.373	0.088	0.704	0.099	0.658
Average eighth-grade ELA	0.360	0.246	0.260	0.304	0.233	0.350
Average eighth-grade Science	0.314	0.267	0.155	0.491	0.155	0.470
Proportion school low SES	-0.487	0.137	-0.238	0.172	-0.215	0.226
Proportion school taking any 2016 AP exam	0.253	0.561	-0.073	0.846	0.009	0.979
Proportion female	-0.231	0.496	0.067	0.827	-0.013	0.966
Average grade	-0.073	0.869	-0.356	0.349	-0.292	0.411
Proportion Asian	0.597	0.223	0.343	0.310	0.372	0.252
Proportion Hispanic	-0.317	0.313	-0.189	0.321	-0.169	0.380
Proportion Black	-0.325	0.148	-0.339	0.074	-0.387	0.045*
Proportion White	0.490	0.031*	0.272	0.076	0.266	0.088

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table CC2: Baseline standardized mean differences between treatment and matched non-experimental teachers' 2017-18 students imputed student-level covariates, for Research Question Three Approach 2 Round Two analytic AP outcome samples

AP total score

	D1	D2	D3	D4	Most extreme
National Math	0.284	0.274	0.287	0.270	0.287
National ELA	0.348	0.341	0.351	0.338	0.351
Eighth-grade Math	0.070	0.058	0.100	0.028	0.100
Eighth-grade ELA	0.094	0.075	0.121	0.048	0.121
Eighth-grade Science	0.150	0.186	0.179	0.157	0.186

AP multiple-choice score

	D1	D2	D3	D4	Most extreme
National Math	0.284	0.282	0.286	0.280	0.286
National ELA	0.348	0.347	0.350	0.346	0.350
Eighth-grade Math	0.070	0.047	0.104	0.013	0.104
Eighth-grade ELA	0.094	0.065	0.127	0.032	0.127
Eighth-grade Science	0.150	0.177	0.181	0.146	0.181

AP free-response score

	D1	D2	D3	D4	Most extreme
National Math	0.284	0.260	0.290	0.253	0.290
National ELA	0.348	0.329	0.354	0.324	0.354
Eighth-grade Math	0.070	0.079	0.102	0.048	0.102
Eighth-grade ELA	0.094	0.093	0.118	0.069	0.118
Eighth-grade Science	0.150	0.209	0.179	0.180	0.209

AP qualifying score (exam-takers only)

	D1	D2	D3	D4	Most extreme
National Math	0.163	0.182	0.156	0.189	0.189
National ELA	0.217	0.234	0.211	0.240	0.240
Eighth-grade Math	-0.007	0.028	0.063	-0.042	0.063
Eighth-grade ELA	0.055	0.093	0.139	0.009	0.139
Eighth-grade Science	0.081	0.147	0.119	0.110	0.147

AP qualifying score (full sample)

	D1	D2	D3	D4	Most extreme
National Math	0.151	0.169	0.140	0.180	0.180
National ELA	0.221	0.237	0.211	0.247	0.247
Eighth-grade Math	-0.007	0.025	0.061	-0.043	0.061
Eighth-grade ELA	0.065	0.098	0.145	0.019	0.145
Eighth-grade Science	0.103	0.173	0.141	0.135	0.173

Took AP exam

	D1	D2	D3	D4	Most extreme
National Math	0.151	0.176	0.102	0.225	0.225
National ELA	0.221	0.248	0.167	0.302	0.302
Eighth-grade Math	-0.007	0.036	0.017	0.011	0.036
Eighth-grade ELA	0.065	0.118	0.090	0.093	0.118
Eighth-grade Science	0.103	0.326	0.122	0.307	0.326

Appendix DD: Non-Experimental Impact and Sensitivity Results—First Round

The general critique of quasi-experimental methods is the lack of random assignment resulting in a comparison group biased by unmeasured characteristics driving selection and, hence, differing from the experimental group. In our non-experimental analysis, descriptive statistics comparing unmatched non-experimental (N2) to experimental second-year KIA teachers (T2) show they differed considerably along various measured dimensions, notably the extent to which baseline students of N2 teachers outperformed students of T2 teachers on the May 2016 APGOV/APES examination (by 0.83 SD's among unmatched N2 to T2 teachers).

Matching improved the comparability of T2 and N2 teachers, though imperfectly, with baseline equivalence differences between T2 and N2 teachers' 2015-16 students' May 2016 APGOV/APES exam performance exceeding WWC thresholds. In each sample of students with non-missing AP outcome measures, baseline students of T2 teachers underperformed those of N2 teachers to an extent surpassing WWC thresholds.

Given the extent to which T2 and N2 teachers differ considerably on measured characteristics, even after matching, they also likely had notable differences in unmeasured ways that we could not control for statistically or through research design. Therefore, we heavily caveat the results presented below as potentially biased based on: 1) measured characteristics not possible to balance via matching, which then requires covariate adjustment with model-dependent estimates; and 2) unmeasured characteristics related to teachers' propensity to enroll in the KIA RCT that may be associated with their students' AP performance outcomes. In Table DD1, we show adjusted estimated effect sizes and associated standard errors comparing AP performance outcomes among students of T2 teachers to matched N2 teachers.

Table DD1: Covariate-adjusted estimates of the overall impact of Knowledge in Action on AP outcomes among 2017-18 students of treatment and Round One non-experimental matched teachers

	Effect Size	95% CI	p-value	N
Took AP exam	-0.34 (0.17)*[S][C]	(-0.663, -0.016)	0.041	2315
AP QS (full sample)	0.411 (0.2)*[S]	(0.025, 0.797)	0.038	2315
AP QS (exam-takers only)	0.157 (0.24)[S]	(-0.318, 0.632)	0.52	1900
AP total score	-0.17 (0.15)	(-0.454, 0.115)	0.24	1309
AP multiple-choice	-0.154 (0.14)	(-0.433, 0.126)	0.28	1309
AP free-response	-0.135 (0.15)	(-0.434, 0.163)	0.37	1309

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit. [C]= model did not converge.

The exam-taking effect is of large negative magnitude (ES=-0.34), indicating T2 students took the APGOV/APES exam at lower rates than N2 students. However, this model failed to converge despite running for more than 20,000,000 iterations, so results are potentially not to be trusted.

The three models with binary outcomes all resulted in singularity; the model could not estimate the variance of the teacher-level random intercepts so it set variance to 0, effectively eliminating the random-effect for teacher. The primary implication is the p-values no longer reflect the clustering of students at the teacher level. P-values reported have not been adjusted, and thus are too small.

When we look at comparisons of matched T2 and N2 teachers' students on the continuous score outcomes, which are available for students in four of five districts, the picture is quite different. The direction of the effect sizes on the continuous AP scores measures (i.e., AP total score, AP multiple-choice, and AP free-response) is negative, albeit insignificant. This is opposite the direction of the experimental estimates of the same outcome, which also are insignificant.

A possible cause for the fairly drastic difference in the substance of the results between the four- and five-district samples relates to matching. While the overall balance on the covariates of interest was within a tolerable range, markedly varying across district and course was the quality of matches, according to the absolute distance in the logit of the propensity score between a treatment teacher and matched controls.

In Table DD2, we show sensitivity of the non-experimental estimates to covariates. Clearly, covariate adjustment was critical to the magnitude of obtained estimates, with large negative effect sizes, albeit not significant, in models fitting dichotomous outcomes without adjustments, and large positive differences in these models with covariate adjustments. P-values have not been corrected in singular models, so estimates noted as significant may not be after correction. For the four-district sample for which continuous outcomes were available, estimates are of lower magnitude insignificant, and in the negative direction across all models, adjusted and unadjusted.

Table DD2: Sensitivity to covariates of estimates of the overall impact of Knowledge in Action on AP outcomes between 2017-18 students of treatment and Round One matched non-experimental teachers

	No covariates	Covariates with ABE > 0.05	Primary	All covariates
Took AP exam	0.016 (0.2)	-0.283 (0.16)[S]	-0.34 (0.17)*[S][C]	-0.317 (0.16)[S]
AP QS (full sample)	-0.241 (0.26)	0.459 (0.2)*[S]	0.411 (0.2)*[S]	0.933 (0.34)**[S]
AP QS (exam-takers only)	-0.284 (0.27)	0.147 (0.24)[S]	0.157 (0.24)[S]	0.587 (0.36)[S]
AP total score	-0.01 (0.21)	-0.181 (0.14)	-0.17 (0.15)	-0.177 (0.14)
AP multiple-choice	0.022 (0.2)	-0.167 (0.14)	-0.154 (0.14)	-0.167 (0.13)
AP free-response	-0.048 (0.21)	-0.14 (0.15)	-0.135 (0.15)	-0.134 (0.15)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit. [C]= model did not converge.

Through the results of our matching and baseline equivalence analysis, we know differences exist on T2 and N2 teachers' measured teacher-level characteristics most associated with our outcomes of interest. While we can adjust in our impact models, with baseline differences this large, results will rely

heavily on accurately modeling the relationship between covariates and outcomes. Perhaps even more concerning, due to both lack of randomization and even larger observed differences present before matching, differences likely also exist on unmeasured teacher characteristics, which we could not control either statistically or by design. Therefore, we urge caution in interpreting these results.

Appendix EE: Non-Experimental Impact and Sensitivity Results, Second Round

Table EE1: Covariate-adjusted estimates of the overall impact of Knowledge in Action on AP outcomes among 2017-18 students of treatment and Round Two non-experimental matched teachers

	Effect Size	95% CI	p-value	N
Took AP exam	-0.883 (0.29)**[S]	(-1.455, -0.31)	0.003	3407
AP QS (full sample)	-0.624 (0.47)[S]	(-1.542, 0.294)	0.19	3407
AP QS (exam-takers only)	-0.477 (0.75)[S]	(-1.95, 0.996)	0.53	2833
AP total score	-0.095 (0.13)	(-0.341, 0.151)	0.45	1668
AP multiple-choice	-0.139 (0.11)	(-0.346, 0.068)	0.19	1668
AP free-response	-0.099 (0.13)	(-0.356, 0.157)	0.45	1668

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit.

Table EE2: Sensitivity to covariates of estimates of the overall impact of Knowledge in Action on AP outcomes between 2017-18 students of treatment and Round One matched non-experimental teachers

	No covariates	Covariates with ABE > 0.05	Primary	All covariates
Took AP exam	0.02 (0.16)	-1.02 (0.29)***[S]	-0.883 (0.29)**[S]	-1.291 (0.61)*[S][C]
AP QS (full sample)	-0.162 (0.21)	-0.308 (0.51)[S]	-0.624 (0.47)[S]	-0.55 (0.7)[S]
AP QS (exam-takers only)	-0.163 (0.22)	-0.48 (0.76)[S]	-0.477 (0.75)[S]	-0.43 (0.87)[S]
AP total score	0.171 (0.19)	-0.088 (0.12)	-0.095 (0.13)	-0.135 (0.13)
AP multiple-choice	0.186 (0.18)	-0.125 (0.1)	-0.139 (0.11)	-0.123 (0.12)
AP free-response	0.137 (0.19)	-0.093 (0.13)	-0.099 (0.13)	-0.131 (0.13)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. [S] denotes a possible underestimation of the true standard error due to a singular model fit. [C]= model did not converge.

Appendix FF: Year One Implementation Analysis, Methodology, and Results

Sample

We base reported results on data collected from teachers, and their students, school leaders, and coaches during the 2016-17 school year.

Schools and Teachers

For our Year One implementation analysis, we include responses from the 70 teachers—31 KIA and 39 control—who participated in the study as designed. Of the 70, all 31 assigned to the treatment condition used materials accessed through the Sprocket online curriculum portal. None of the 39 control teachers accessed Sprocket. KIA Summer Institute attendance among 30 participating treatment teachers ranged from 90-97%. KIA Professional Development Session attendance ranged from 90% for the first session to 68-74% attendance across the three subsequent sessions.

These 70 teachers were distributed across 65 schools that were predominantly urban (75%) and serving high proportions of economically-disadvantaged students, with 71% classified as Title 1 (NCES, 2015). Of all students enrolled in all 65 schools, 63% received free or reduced-price lunch (Table FF1) and 14% were identified as English Language Learners.

Table FF1: School characteristics

	Percent (%)	Obs. (N)
Charter	3	65
Magnet	35	65
Urban	75	65
Title I	71	65
Average % FRPL	63	65
Average % English learners	14	65

Note: Source is National Center for Education Statistics, 2015.

We drew from district administrative data to describe the demographic composition of our teacher sample. Given substantial missingness in these records, to the extent possible we supplemented with self-reported data gathered through teacher surveys. Among those for whom we have data, our teacher sample was predominantly White (79%) and female (61%). They were relatively experienced, as 67% had more than 10 years of experience, and 75% held at least a master's degree. Sample means and remaining missingness counts for each teacher demographic variables are presented in Table FF2.

Table FF2: Teacher characteristics

	Percent (%)	Obs. (N)	Missing (N)
Female	61	70	0
White	79	52	18
Black	8	52	18
Asian	13	52	18
10+ years of experience	67	66	4

MA+ Education	75	53	17

Of teachers participating in the study, more than half (63%) taught one APGOV or APES section, 17% taught two sections, and one-fifth (20%) taught more than two sections; seven teachers taught three sections, six teachers taught four sections, and one teacher taught five.

Students

We describe two samples of students who contribute data to this study. The first includes all 3,467 students enrolled in APES or APGOV classes taught by the 70 participating teachers. These are the students reflected upon by teachers in their self-report data. The second is a subset of those students: those who completed student surveys—a key source of data for describing classroom instructional practices. Among teachers who taught more than one APGOV or APES section, we collected survey data from one randomly-selected section. Completing surveys were 747 students from across 50 classrooms (n=22 KIA, 28 control). We present summary statistics for the two student samples in Table FF3. In addition, a subset of students (n=137) participated in interviews and focus groups.

Table FF3. Student characteristics

	Total Sample		Survey Sample	
	Percent (%)	Obs. (N)	Percent (%)	Obs. (N)
APES	54	3467	62	747
APGOV	46	3467	38	747
9th grade	5	3465	12	747
10th grade	9	3465	5	747
11th grade	25	3465	26	747
12th grade	61	3465	57	747
Female	56	3467	56	747
Asian	14	3467	14	747
Black	9	3467	6	747
Hispanic	39	3467	46	747
White	35	3467	31	747
Other race	3	3467	3	747
Economically disadvantaged	44	3467	44	747

Of the total student sample, 54% were enrolled in APES, and 46% in APGOV. More than three-fifths of all students were in 12th grade (61%), 25% were in 11th grade, and 14% were of grades 8-10. Almost half (47%) were from traditionally disadvantaged racial groups (i.e., non-Asian and non-White: Black, Hispanic, Native American/Islander)—a fact seen in the numbers of economically-disadvantaged students (44%), as measured by eligibility for free or reduced-price lunch.

The sample of students contributing survey data was mostly comparable to the overall student sample on key variables including gender, race, and economically-disadvantaged status. The largest difference between the two samples was the proportion in APES versus APGOV: 62% in APES in the survey sample compared to 54% in the total sample.

As a basis of comparison, our student sample was composed of a greater proportion of minority students relative to the population of students who took 2017 APGOV and APES exams. Overall, 65% of our student sample was non-White compared to 33% of the 2017 APES and APGOV exam-taking sample nationally (College Board, 2018). Our sample was 44% low-income compared, to 27% among AP 2017 exam-takers across subjects nationwide (Godrey, Wyatt, & Beard, 2016).

Data

To address our research questions, we collected administrative records as well as interview, survey, and/or instruction log data from students, teachers, and PD coaches. We also collected PD participation data and data describing teachers' use of the online curriculum portal.

Survey instruments

We developed four survey instruments informing this study: teacher surveys, pre-school year and post-testing (i.e., the summer preceding the intervention year and after the AP examination period); a daily instruction log for teachers to complete on 10 consecutive teaching days in spring 2017; and an end-of-school-year student survey. We adapted survey items from several open-access sources, including the Consortium for Chicago School Research, the International Education Association Civic Education Study, and the National Survey for Science and Mathematics Education.

Teacher surveys asked questions about teaching context and teaching supports, beliefs about student learning, and practices in curriculum and teaching. We posed the same questions to teachers before and after the KIA implementation year. Both measures were retrospective in nature: In the pre-survey, teachers reflected on their 2015-16 experiences, while in the post-survey, teachers looked back at their 2016-17 experiences. Due to the timeline of recruitment and randomization, teachers knew of their treatment status when they completed their pre-survey.

In addition to retrospective surveys, we also collected daily instruction logs from teachers over 10 consecutive class meeting days in March and April 2017. Each log posed questions about that days' class, including which student learning objectives were the focus of that day's instruction, how they spent the majority of class time (e.g., whole-group lecture, small groups, etc.), and what type of homework was assigned that day. Instruction logs offer two salient advantages over retrospective end-of year reports of classroom activities: 1) Measures are recorded closer in time to the actual event, reducing retrospection bias, and 2) repeated measures allows researchers to calculate and describe respondents' average experiences, which may better reflect a teachers' typical classroom than any single day's report (Iida, Shrout, Laurenceau & Bolger, 2012). We include the instrument log in Appendix B (online).

We used the student post-survey to capture two types of data: those relevant to curriculum and instructional practices, and those relevant to outcomes. (Survey results are referenced in separate publications.) To examine classroom practice, we asked students questions paralleling those asked of teachers, giving us a means to triangulate reports of what treatment classrooms "looked like" compared to control classrooms. Topics included students' perception of their teachers' instructional practices, the nature of homework, and prevalence of typical class activities (e.g., groupwork and classroom discussion).

Interviews

We interviewed individual students by telephone at the beginning of the year and again after the May 2017 AP examinations. After the exam period had concluded, we also conducted in-person group interviews with up to eight students per classroom among visited schools. Protocols for both were similar and focused on asking about students' background in AP, prior experience with PBL instruction, attitudes toward AP courses, engagement with learning, feelings of preparation for the AP examination, and their challenges and successes in the course. To elicit reflections on changes between the beginning and end of the year, we designed pre- and post- one-on-one interview protocols.

We also designed teacher interview protocols to allow teachers to reflect on changes over the year, as we conducted a pre-interview prior to exposure to the treatment and another at the end of the school year. Aligned to both the student interview protocol and the teacher surveys, teacher interview protocols focused on teachers' experience teaching AP and using PBL approaches, beliefs about PBL, and challenges and benefits of implementing AP curriculum.

Administrative records and publicly available data

Districts provided administrative student records, which we supplemented as needed with data from the National Center for Education Statistics, and state and district websites.

Analytic Methods

To address our research questions, we synthesized results from teacher and student surveys, teacher instruction logs, and interviews with students and teachers. Of the 70 complier teachers, 54 (27 each in treatment and control) completed both pre- and post- surveys, and 61 (27 treatment and 34 control) completed instruction logs. Students from 50 classrooms (22 treatment, 28 control) completed student surveys. We also drew from interviews with a qualitative subsample of 20 teachers (14 treatment and 6 control) and their students, as well as all 12 PD staff and coaches.

We did not use inference methods for this study for several reasons. First, given the sheer volume of implementation survey items and scales, appropriately adjusting for the Type 1 error rate would result in such extreme corrections that no differences would emerge. Second, the vast majority of implementation measures were ordinal in nature, requiring ordinal and/or multinomial models—which complicates interpretation of results. Instead, we examined results descriptively, relying on standardized mean differences (described below) and triangulated quantitative results with qualitative interviews. We defined meaningful standardized mean differences as those larger than 0.25 standard deviations (ES). We also point out differences of smaller magnitude when they are part of a measured pattern describing a trend across items or questions.

Teacher survey

The goal of our teacher survey analysis was to calculate standardized mean differences between treatment and control teacher responses to survey scales (n=18), and items that did not form scales. Among the 70 participating teachers who completed both pre- and post-surveys—the sample for which we report results—our overall response rate was 77%; broken down, it was 87% among treatment teachers and 69% among control.

We used Glass's delta (Ferguson, 2009), dividing the mean difference by the standard deviation of the control group, which represents the variability in the absence of the intervention (i.e., business-as-usual). Standardized mean differences allowed us to interpret differences between treatment and control groups on the same scale, regardless of the metric of the original variable. For survey questions posed only to treatment teachers, we presented descriptive statistics in terms of their original unit of measurement (e.g., frequency of selected response option).

Student survey

The student survey response rate was 71% among treatment classrooms, 72% among control. To analyze survey data, we aggregated students within classrooms to create classroom-level averages. We aggregated to the classroom level for two reasons. First, a greater number of student responses reflecting a single teacher is a more reliable reflection of that teacher's/classroom's activities than an individual response alone. Second, creating classroom-level means acts to weight each classroom equally. In contrast, using unaggregated student-level data weights each student equally, such that classes with large responding numbers of students will count more heavily than classrooms with small responding numbers of students. As with the teacher survey data, we used Glass's delta to calculate standardized mean differences between treatment and control classrooms.

Instruction logs

Of our analytic sample complier teachers, 61 of 70 (87%) submitted at least one log for a total of 471. On average, treatment teachers completed 8.9 logs while control completed 8.7. We first examined log data for patterns of missingness and compliance with instructions that the logs be completed within 48 hours of class. The number of logs completed per teacher ranged from one (n=two teachers) to 13. We excluded a total of 50 logs from 22 teachers because they were completed more than 48 hours after the day of instruction (or beyond Monday following a Friday class). We also excluded logs completed for school days committed to state testing, field trips, assemblies, or other anomalies.

To compare the daily activities of treatment and control teachers, and summarize their instruction, we collapsed observations within teachers to represent the percent of instructional days on which they reported an activity. For yes/no questions, collapsing was straightforward, and we then compared the average percentage of days reported "yes" on a given item between treatment and control groups. For questions with a frequency scale response, we calculated two summary statistics per teacher: 1) the percent of days on which the teacher reported any of the activity; and 2) an average for the extent to which the activity was implemented on a scale of 1 to 3 (limited, moderate, great extent). We then summarized these teacher-level averages by looking across all treatment teachers and all control teachers and reporting: 1) the mean percentage of days; or 2) the standardized mean difference of the extent to which the activity was implemented.

Qualitative Analysis

Before beginning our qualitative analysis, we looked to extant literature to define two constructs critical to our study: deeper learning and authenticity. Though there are a number of ways to think about what deeper learning means, we used the Hewlett Foundation's conceptualization to guide our understanding of the construct for qualitative coding and analysis purposes. According to Hewlett's definition (2010), it is through deeper learning that students learn how to learn, master core academic

content, work collaboratively, think critically and solve complex problems, and communicate effectively.

Another intended feature of the KIA approach is that students' learning opportunities should be authentic. We followed Polman (2015) in defining authentic learning to mean: 1) students find content and skills relevant to their lives; 2) they share products of their learning (e.g., a piece of written work or a presentation) with people outside of the classroom; and 3) they practice using "tools" people use outside of classrooms, such as letters to the editor or their government representatives.

We began the coding process with a structural approach (Saldaña, 2016), indexing our transcripts for topics of interest to the research study and for which we had designed our interviews to probe. For example, topical codes included: "learning objectives," "groupwork," and "student-centered." We then inductively coded the transcripts exhaustively for participants' perspectives regarding the primary topics. During this stage, as a means to capture participants' voices and feelings, we employed a high level of specificity using descriptive coding, emotion coding, and in vivo coding (Saldaña, 2016). Multiple coders participated in this process, with reliability checks performed by the lead qualitative researcher. This stage resulted in a large number of codes requiring further synthesis, in which we merged similar codes together into subcategories within the topical categories (e.g., teacher challenges facilitating groupwork, benefits of groupwork)

The subcategories formed the basis for our in-depth analysis across topics and groups by treatment status. To identify which domains and ideas were relatively more common and which rarely occurred, we generated code frequencies, determined by the number of unique participants who mentioned a particular idea (rather than by the total number of times an idea appears in the text). When interpreting code frequencies, we accounted for the different sample sizes of treatment and control groups in the qualitative sample. We then analyzed and annotated resulting data to investigate any differences between control and treatment groups. We also identified disconfirming evidence (Erickson, 1986)—less-frequently occurring domains that contradict those occurring more frequently.

We report on the most common emergent themes overall, and by treatment/comparison group, as well as use qualitative data to help interpret and illustrate quantitative findings. When participants' responses vary based on their role (e.g., student versus teacher), we report on divergent perspectives.

Triangulation of results across data sources

We used teacher instruction log, survey, and qualitative data to address our first research question, relying heavily on quantitative data to identify differences of meaningful magnitude, and qualitative data to help explain, illustrate, or further expand our understanding of those differences. We addressed our second research question with the same three sources of data; however, we relied more heavily on qualitative self-report data from treatment participants to answer "how" questions regarding implementation successes and challenges. To attenuate self-reporting bias or bias incurred through participants' knowledge of their treatment status, we triangulated teachers' self-reports with students' own reports and coaches' reports of classroom activities. We used qualitative data to explain quantitative results, and also to provide context, exemplars, and divergent or disconfirming evidence.

Results, Year One Implementation Research Question #4a

Our first implementation-related sub-research question was, "Between APGOV and APES teachers randomly assigned access to the KIA intervention (i.e., course-specific PBL curriculum, instructional materials, and professional development supports) versus those continuing business-as-usual instruction, what were the differences in teachers' and their students' self-reported classroom experiences with curriculum and instructional approaches?"

As expected, due to teachers' self-selection into the study, control teachers used some curriculum and instructional practices that were consistent with the KIA approach, as did all teachers prior to their involvement with KIA. Since shifting to PBL is demanding on several levels, treatment teachers may well have abandoned their efforts at any point throughout the year. As such, we did not expect to find differences between treatment and control teachers on every domain measured, though it was realistic to expect to see more KIA-like practices taking place in treatment classes than in control classrooms.

We found that relative to control, treatment teachers used student-centered, KIA-aligned practices more frequently and to a greater extent. We also found that treatment teachers, on average, did indeed sustain their KIA use throughout the year. Results did not differ across courses, except when explicitly noted. In several areas, highlighted at the end of this section, student survey data suggested differences in practices between courses.

Treatment teachers' student learning objectives were more focused on deeper learning

When we interviewed teachers at the beginning of the year, we asked about their most important learning objectives for AP students. The objectives most commonly described as most important, in both treatment and control groups, related to civic engagement, particularly community activism, and environmental awareness/activism.

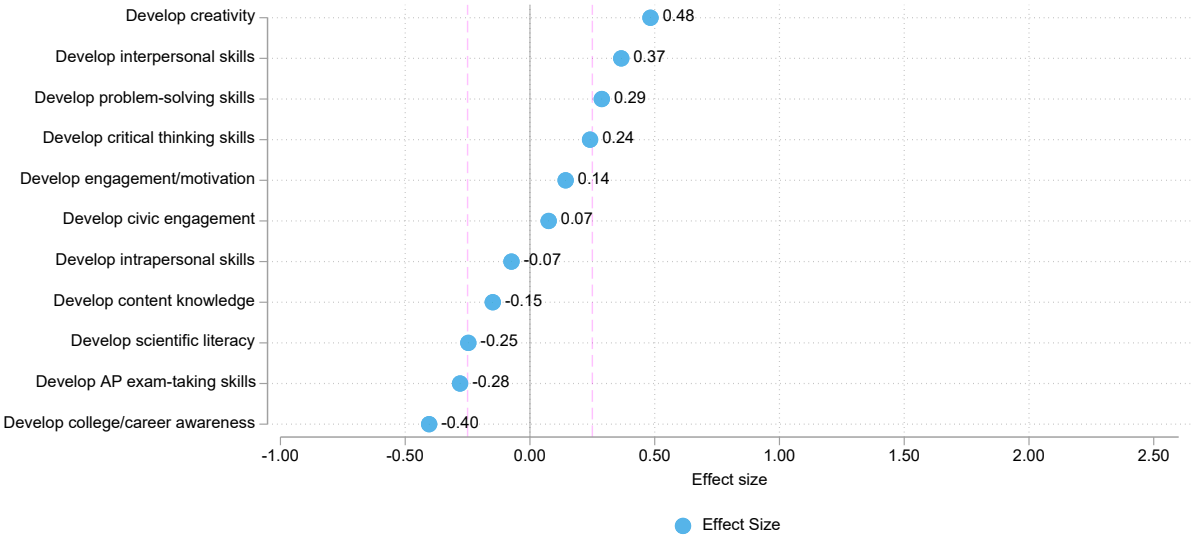
By mid-spring 2017, treatment teachers reported over the course of consecutive instruction log days that their instruction more frequently focused, than their control peers, on deeper learning objectives. As we show in Table FF4, treatment teachers more frequently reported focusing on developing: creativity (by 14 percentage points), interpersonal skills (by 9), college/career awareness and problem-solving skills (both by 8), and civic engagement (by 6). Both groups generally focused on the other learning objectives with similar frequency.

Table FF4: Difference in the percent of days on which treatment and control teachers reported focusing on various student learning objectives over consecutive instruction log days in spring 2017 (n=61)

	Control (%)	Treatment (%)	Difference (%)
Develop creativity	67.0	81.1	14.1
Develop interpersonal skills	73.0	81.6	8.6
Develop college/career awareness	48.9	57.3	8.4
Develop problem-solving skills	83.1	91.3	8.2
Develop civic engagement	73.5	79.9	6.4
Develop intrapersonal skills	79.5	83.2	3.7
Develop engagement/motivation	90.1	93.5	3.4
Develop critical thinking skills	95.3	97.1	1.8
Develop content knowledge	97.8	97.5	-0.3
Develop scientific literacy* (only APES teachers)	95.9	94.5	-1.4
Develop AP exam-taking skills	82.2	80.4	-1.8

In addition to frequency, instructional logs also showed treatment teachers focusing in greater depth on these skills. Figure FF1 shows the difference between treatment and control teachers' instruction log ratings of the extent to which they emphasized various learning objectives on the days on which they reported having focused on them at all. Effect sizes (ES) to the right of zero indicate the treatment group average was higher than the control group, and vice-versa. Red dashed lines mark a difference of one-quarter of a standard deviation (SD). Treatment teachers more strongly emphasized developing creativity, interpersonal skills, and problem-solving compared to control, with effect sizes of, respectively, 0.48, 0.37, and 0.29 on those days. An effect size of 0.24 on developing critical-thinking skills approached the threshold of one-quarter of a standard deviation. On the other hand, though treatment and control teachers reported developing students' AP exam-taking skills on a similar percentage of days, on those days, treatment teachers focused on this skill to a lesser extent than control teachers (-0.28 ES). Treatment teachers also reported less emphasis on college/career awareness on the days in which this was an objective (-0.4 ES).

Figure FF1: Standardized mean differences between treatment and control teachers' reported extent to which instruction focused on various student learning objectives over consecutive instruction log days in spring 2017 (n=61)

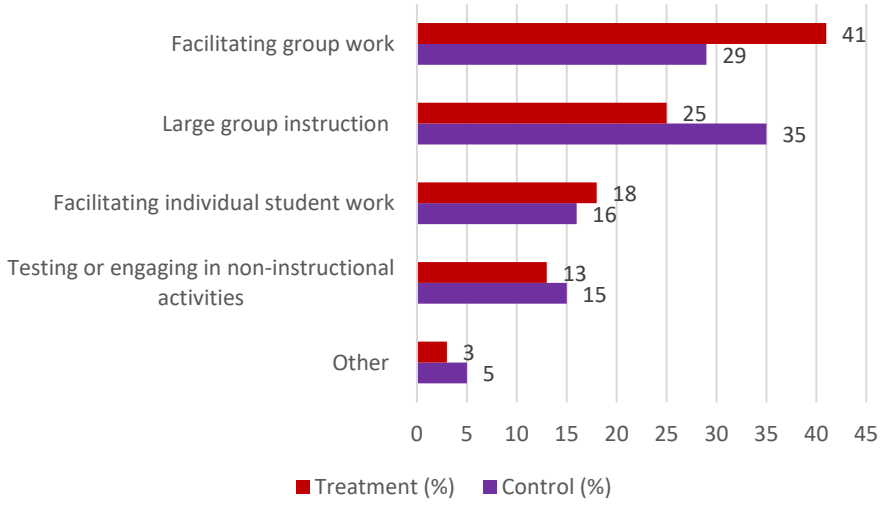


Treatment teachers used student-centered instructional practices more frequently

In mid-spring 2017, treatment and control teachers reported similar rates of active student engagement with course learning (87% of instruction days among both groups). However, emerging through teacher and student reports were fundamental differences between how treatment and control teachers balanced transmission and student-centered instructional practices.

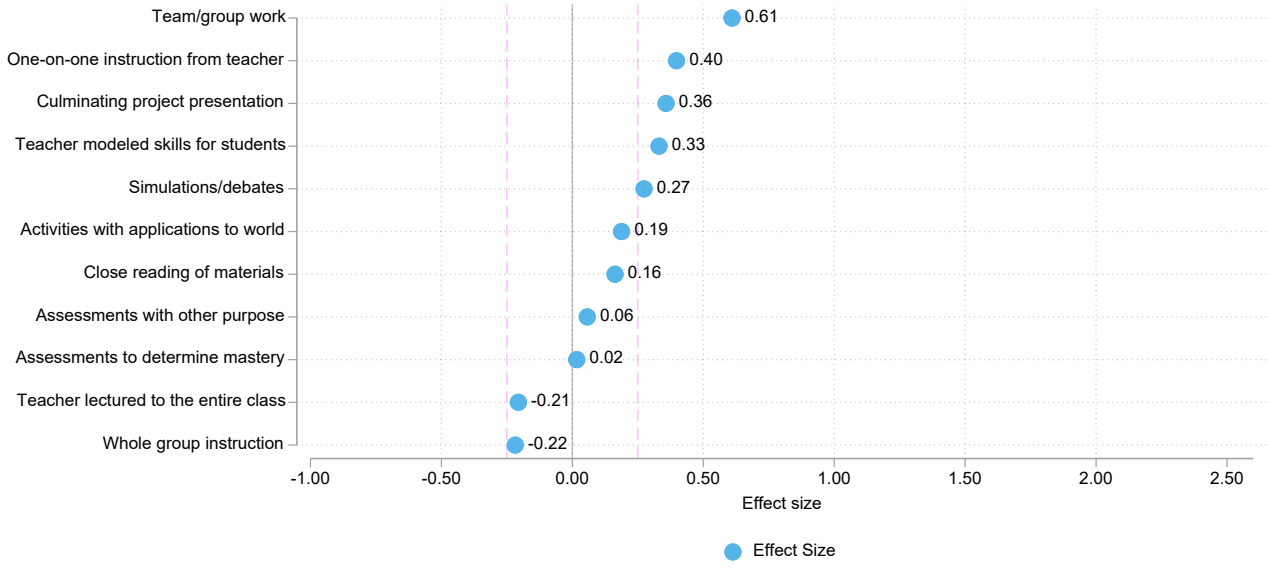
Treatment teachers, in their end-of-year surveys (Figure FF2), reported spending the most time facilitating groupwork (41% of the time) followed by delivering large-group instruction (25% of the time). In contrast, control teachers reported most often delivering large-group instruction (35% of the time) followed by facilitating groupwork (29% of the time)—an average of 12 percentage points lower than treatment teachers.

Figure FF2: Treatment and control teachers' reported proportions of class time spent facilitating groupwork and delivering large-group instruction at the end of the year (n=54)



Log data also revealed treatment teachers' greater use, compared to control teachers, of student-centered instructional practices. As shown in Figure FF3, KIA teachers reported more heavily emphasizing groupwork by more than half a standard deviation. With standardized mean differences of greater than 0.25, relative to control they also more heavily emphasized one-on-one instruction, culminating project presentation, teams presenting materials, teachers modeling skills for students, and simulations/debates.

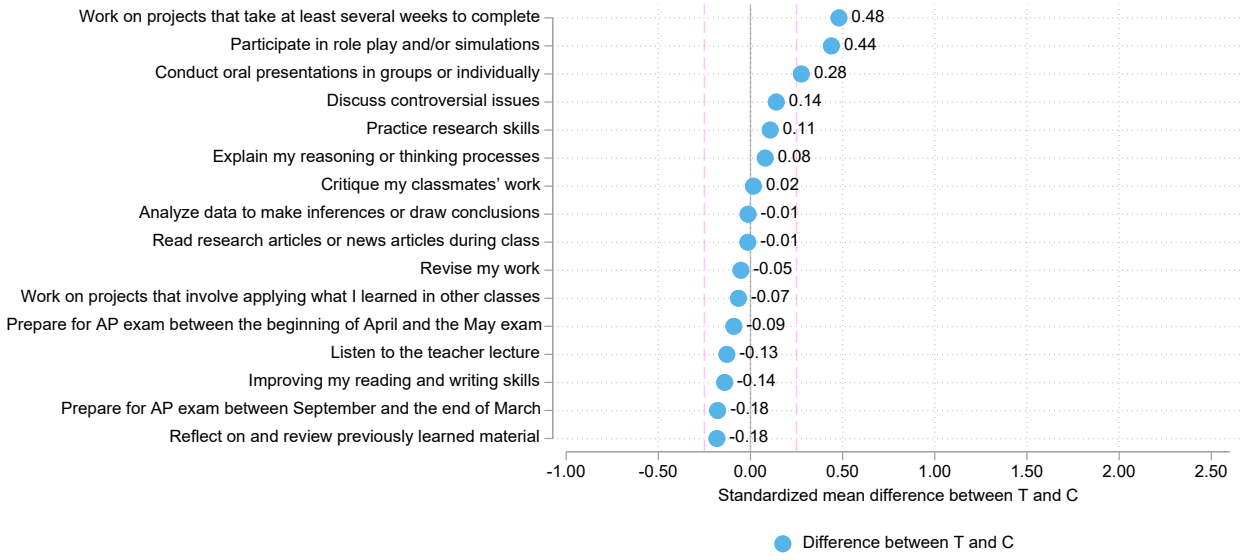
Figure FF3: Mean difference between treatment and control teachers' reported extent to which instruction included various activities over consecutive instruction log days in spring 2017 (n=61)



In terms of out-of-class assignments, treatment and control teachers both reported assigning such work on 64% of instruction log days. However, the nature of what teachers asked students to do was meaningfully different: treatment teachers assigned ongoing project work 65% of the time, compared to but 35% of the time among control teachers. The proportions were reversed for quick turnaround work, with control teachers assigning this 65% of the time and treatment teachers doing so 35% of the time. This difference of 30 percentage points indicates a fundamental change in the homework assigned by participating teachers.

Comparisons between students' end-of-year survey-based reflections on teacher behaviors did not indicate as many pronounced differences between treatment and control classrooms—though the pattern of responses indicated more student-centered activities, and less lecture and AP exam preparation. The most pronounced differences (Figure FF4) were that compared to students in control classrooms, students in treatment teachers' classrooms reported to a greater extent working on projects taking several weeks to complete (ES = 0.48), participating in role play and/or simulations (ES = 0.44), and conducting oral presentations (ES = 0.28). Treatment students also reported more frequent use of performance-based assessments compared to control (ES = 0.63). Albeit of lesser magnitude, students in treatment classes reported less exam preparation (ES = -0.18) and lecturing (ES = -0.13).

Figure FF4: Standardized mean differences between students' end-of year reports on extent of engagement in inquiry- and transmission-based instructional activities in treatment compared to control classrooms (n=747)



Differences between KIA APGOV and APES students' reports about instruction

Student surveys revealed meaningful differences between instruction in KIA APGOV as compared to KIA APES. Regarding several items and scales on average across courses, there was no meaningful difference between treatment and control classrooms—but there was a meaningful difference in one

course but not the other. In other cases, what looked like a meaningful difference was driven by a large difference in one course, along with a smaller one in the other.

As an example of the former, whereas the overall difference between treatment and control students' reported frequency of teacher lecture was small (ES = -0.13), KIA APGOV students reported considerably less lecture (ES = -0.29) relative to control, whereas there was virtually no difference between APES KIA and control classes (ES = 0.02). Similarly, though the overall treatment-to-control difference in frequency of discussion of controversial issues was small (ES = 0.14), the difference was large in APGOV classes (ES = 0.30)—yet almost nonexistent in APES classes (ES = -0.01). On the other hand, treatment APGOV students reported less frequently than control students reading research or news articles during class (ES = -0.19), while the difference was in the opposite direction in APES classrooms (ES = 0.21). And, though there was little difference overall between treatment and control students' reports on the frequency of taking quizzes and tests (ES = -0.14), time detracting from "important learning activities," students in KIA APES classrooms reported spending less time taking quizzes and tests compared to non-KIA APES classrooms (ES = -0.38), while in KIA APGOV classrooms the difference was small (ES = 0.08).

In two cases, a difference that looked meaningful across all participating students was actually large in one course and small in the other, and in both, the larger differences were in AP GOV. First, on average, treatment students reported a greater frequency of roleplay and/or simulations (ES = 0.44)—but in APGOV classes the difference was almost one full standard deviation (ES = 0.94), while in APES classes it was almost zero (ES = 0.06). Second, the extent to which students conducted oral presentations (ES = 0.28) was again larger in APGOV classes (ES = 0.48) than APES classes (ES = 0.16). Overall, the student post-survey results showed larger pedagogy-related treatment differences in the KIA APGOV course than the KIA APGOV course, suggesting that students in KIA APGOV classrooms engaged in more PBL activities than their KIA APES counterparts.

Results, Year One Research Question #4b

Our second implementation sub-research question was, "How did AP teachers and their students in treatment schools describe their PBL experiences including perceived challenges and benefits?" As noted above, we relied primarily on qualitative data collected from teachers and students in treatment schools, as well as from coaches, to address this "how" question, though we refer to teacher and student treatment-control differences measured through surveys as relevant.

Adjusting to student-centered instruction

According to concurring end-of-year interview reports from treatment teachers, students, and coaches, acclimating to KIA's student-centered approach was one of the greatest challenges. Groupwork on projects, which is key to the KIA approach, presented difficulties for teachers and students, though students also found groupwork and projects to be beneficial and enjoyable. Troublesome for both instructors and students was pacing—the rate at which they progressed through the KIA curriculum.

Balance between student-centered instruction and lecture was drastic yet teachers persisted

The shift from their prior AP experience to the KIA approach felt like a major change for both teachers and students. For example, reflecting back on the year, one treatment teacher explained that upon enrolling s/he, "had no idea of the scope and how much it would impact the classroom, and how much it [affects] students." This teacher described KIA as, "definitely changing the classroom functioning, atmosphere, planning… everything—in a positive way, though."

Interview data from teachers, coaches, and students aligned regarding students' experiences with the student-centered approach: It was difficult to strike a successful balance between student-centered instruction—including projects and groupwork—and typical transmission instructional approaches. During beginning-of-year interviews, treatment teachers anticipated the challenge of getting their students to "buy in" to the AP-level work, including the shift from being "spoon-fed" information to more of an independent approach to learning. In the end-of-year interviews, treatment teachers talked even more about students' resistance and pushback. As one treatment teacher explained, the student-centeredness of KIA required a fundamental and difficult "shift in mindset" for students.

From the perspective of students in treatment teachers' classrooms, many described their struggle as feeling they had to teach themselves content and skills, and they wanted more guidance from their teachers. For example, in one group interview, students discussed how they "didn't really get anything" from their teacher. One student felt, "We basically taught ourselves this course, the whole thing," while another clarified, "I mean, we did get a small portion of [the teacher] helping, but mainly it was just us doing the work and trying to learn it." A treatment teacher described hearing comments like these from students after the first project cycle, and that they helped him/her work harder to balance transmission and student-centered activities.

The shift to more student-centered approaches seemed most uncomfortable for students who were the most comfortable (and successful) with traditional transmission approaches. One coach noted hearing about "pushback from kids" through teachers in which, "traditionally academically successful kids are being asked to do something that's not comfortable when they've been always really good at school and they understood the rules of school."

Parker et al. (2011) described this type of student pushback to the student-centeredness of the KIA approach as "the two worlds" challenge. Some students, familiar and successful at preparing for AP examinations through the transmission method of instruction, find the shift to the student-centered KIA approach to be uncomfortable; they worry their recipe for success may no longer work. One student interviewed for the present study summarized this perspective, explaining, "I wish that sometimes in class she would just stop with the projects [and] just go with the traditional class group setting—just teaching us like a regular class would."

Treatment teachers found grouping students and distributing work challenging

In end-of-year interviews, KIA teachers most frequently described two difficult-to-implement features of groupwork on projects: 1) effectively grouping students, and 2) ensuring all group members participated and shared an equal distribution of the work. While teachers described different strategies for tackling these challenges, they also explained that groupwork success depended on students'

efforts—or lack thereof. One teacher explained how groupwork can help meet the needs of diverse learners, yet "stronger personalities" can still "take over." Another teacher spoke of changing grouping strategy throughout the year. Treatment teachers most commonly mentioned varying student grouping combinations and assigning specific roles to group members as effective means of facilitating groupwork.

Students in treatment classrooms valued groupwork

Though students in KIA group interviews talked about groupwork challenges in ways similar to KIA teachers (e.g., the challenge of equally distributing contributions), in other ways students' perspectives on groupwork were positive. In response to the open-ended end-of-year survey question about what they enjoyed most about their class, groupwork was one of the most common responses. One student "had a really great time working together as a group to achieve different goals," crediting "other classmates [who] helped us to learn the topic more easily." Another stated, "The thing I like most in this class is that we were given the opportunity to work in group many times and learn different skills from others." Students also frequently reported that groupwork helped them better communicate and collaborate with their peers—"getting to hear their opinions" and "learning together and having that kind of communication"—which added to their learning experience. One student wrote that groupwork "helped me to stay focused and enjoy the class even more." Some responses compared KIA groupwork to the work they do in other classes: "My other classes mainly [were] independent work, so it felt good to work with other students." and "I've never had a class that is so group-heavy before, and I enjoyed working with that style of learning."

Nearly two-thirds of KIA students interviewed, either one-on-one or among others, also described groupwork as enjoyable and helping them to learn. One student concluded, "Even though I complained" about groupwork, "I enjoyed working in groups and doing projects with other people... I just wish we had a little more time to complete certain things." This sentiment connects the advantages of groupwork with the challenges of KIA pacing.

Pacing challenges were intrinsic to the first year and also specific to KIA

Teachers', students', and coaches' end-of-year feedback about KIA pacing fell into two categories. The first related to teachers' learning curve for adjusting to appropriate pacing, which is generally intrinsic with any new curriculum and implies a particularly heavy workload for teachers in the first year. Most KIA teachers interviewed expressed the feeling that using KIA as a new curriculum for the first time was time-intensive and involved a heavy workload for the teachers themselves. The second was teachers' and students' perception that the KIA curriculum included too many activities with insufficient time built in for reflection and review, resulting in students' "project fatigue." In combination, the two-fold pacing challenges were the most pronounced of all difficulties discussed by KIA teachers in their end-of-year interviews.

One teacher explained, "projects were too long," and units "always took twice as long as it said it was going to take," which was "frustrating" and led to "fatigue from the students." Another attributed some of the "time challenge" to "learning some of the things while the students were learning at the same time." As both alluded, teachers felt that too much material—and new student-centered PBL material, at that—was challenging for students. A coach who thought the KIA curriculum was "pretty

solid" concurred, sharing the perspective that teachers "have a hard time getting through the curriculum, if [they] want to get through all the project cycles." One teacher summarized feelings changing as the year progressed: "In the beginning our students were a lot more engaged in the project … I thought it was really effective." However, towards the end of the year, "senioritis kicked in" at which point the teacher described the KIA approach as "very ineffective and almost impossible to do." At that point, the teacher admitted, they "just lectured."

Students themselves also felt the pacing was too fast; this sentiment emerged in group interviews with treatment students, as well as a handful of individual treatment student interviews at year's end. As one student said, "[If] we could just review it once in a while and break things down, like slow the pace just a little bit, it would be a little bit better." Students also described projects feeling tedious because there were too many. One student in a group interview described fatigue with projects that came "one after another after another." This student wished for more time to "just go over things." In another group interview, a student expressed a similar feeling including desire for "almost 50/50" balance between projects and direct transmission modes of instruction. This student described projects as fun but, "It was just too much, too many, and not enough time."

Pacing was very challenging for both teachers and students. As one student summarized in a group interview, "The ideas of the projects are good. It's just some of us—or maybe all of us—just need to pause just a moment just to breathe in because of all the stress and anxiety." The solution of cutting curriculum and combining lessons was one of teachers' most frequent adaptations.

Persistence with Knowledge in Action despite challenges

Though the transition felt difficult, treatment teachers sustained their use of KIA through the year. On the end-of-year survey, of the 27 out of 31 treatment teachers responding, all reported using KIA project units throughout the year, with or without adaptation. More than half (52%) taught all five units that compose the year-long course, and nearly 90% used four of five. None taught less than three. As another perspective on the extent to which they persisted with the approach, on 82% of instructional log teaching days, treatment teachers reported using KIA curriculum.

Perceived benefits of KIA for students

Our end-of-year interview protocols for both treatment teachers and students included a general question asking for reflections on how they believe KIA affected them and/or their students. The most prominent theme to emerge from responses related to deeper student learning. Other common responses from KIA teachers and students referenced students' increased awareness of civic and environmental issues. Students also felt prepared for the AP examinations despite reporting less frequent explicit AP exam-preparation activities.

Deeper learning through KIA: perceived improvements in students' ability to learn

Among the five Hewlett aspects of deeper learning, the most pronounced was teachers' perception that through KIA, students learned how to learn. Approximately half the interviewed KIA teachers talked about students' growing persistence, responsibility for their own learning, accountability to others, research and discussion skills, note-taking skills, etc. They also referenced students' newfound appreciation of the necessity for daily effort, attendance, and not procrastinating to produce high-

quality work. One treatment teacher said, "KIA really pushed them to do high-quality work." Opportunities to iteratively submit drafts/components of work product helped students develop their time management, and understanding of and accountability for producing high-quality work. This teacher felt that by requiring "redesigns" and "not allowing them to settle for mediocre work," students learned "accountability for time management" and work quality.

Students also referenced their development of time management, efficiency, and organizational skills. As an example, in one group interview, one student commented that learning to manage time can help to deal with "stress and anxiety" associated with juggling a heavy academic workload. However, students' feeling that they had to teach themselves content and skills, as well as their desire for more guidance from teachers (as described above), suggests the KIA curriculum and/or KIA teachers' instructional practices could make learning-to-learn goals more explicit.

Whereas greater proportions of teachers than students reflected on the process of "learning how to learn," treatment students commonly referred to their mastery of core academic content as the most prominent benefit of KIA. Treatment students, in most group interviews and about one-third of individual interviews, described becoming interested in the subject field, said they felt informed and knowledgeable about the subject, and were interested in future study in the subject area. A treatment student compared the broad relevance of APGOV for all students—"Every two years, there's going to be an election so you've got to be prepared for that."—to AP Calculus, which may be relevant only to those students who "go into engineering or something like that." As another student described, many students experienced KIA as a more hands-on, authentic way to learn, which helped them understand the content more so than the standard transmission approach.

> "I take a couple of AP classes, and I think this is one of the easier ones for me to understand and grasp more because I am a hands-on learner. All the other ones, it's kind of like you sit in a class and you take notes, and then you don't understand those notes, and then you fail the test and so on. I think this class made us more involved in what we were learning, so it was easier to grasp."

Approximately one-quarter of treatment teachers discussed how KIA students gained more understanding of the course content through projects, relative to business-as-usual AP instruction.

Students also referenced another aspect of deeper learning, gaining experience working with peers and honing collaboration skills, as a benefit of KIA. One-third of interviewed KIA students, and participants in nearly two-thirds of the student group interviews, talked about groupwork having an effect on them besides being enjoyable. Working with their peers helped them learn, and also helped them practice their communication and collaboration skills. These results relate to students' and teachers' more general perspectives on groupwork.

KIA teachers viewed ample opportunities for authenticity as a KIA strength

One of the intended features of the KIA approach is that students' learning opportunities should be authentic. PBL can be authentic in several ways: Students can find content and skills relevant to their lives; they can share products of their learning (e.g., a piece of written work or a presentation) with

people outside of the classroom; and they can use "tools" seen outside of the classrooms, like letters to the editor, podcasts, and films (Polman, 2015).

Compared to control teachers and students, at the end of the year, treatment teachers and students both reported more frequent "authentic" learning opportunities, such as sharing their work with outside audiences (by 2.3 ES's and 0.23 ES's, respectively, in surveys). Interviewed treatment teachers felt the KIA curriculum's greatest strength was ample opportunities to make it authentic, though teachers needed to adapt to make it so. In end-of-year interviews, most treatment teachers referenced the real-world relevance of the curriculum to students' lives (i.e., one of the Polman's features of authenticity) as a positive and none referenced it as a challenge. In contrast, control teachers only talked about the challenges of planning authentic learning experiences.

Though teachers had to work at teaching the KIA curriculum in a way that was authentic for their students, they described KIA tasks as "encouraging students to think about things that are going on in the real world and help make those connections." One teacher described how "campaigning and role-playing" led to students becoming "very politically active themselves."

Two of the KIA curriculum's strengths, authenticity and opportunities to develop students' civic skills and engagement, were quite congruent. Teachers discussed authenticity and civic focus similarly, in that the curriculum's strengths lie in the opportunities to make students' learning processes authentic and civic-oriented. One teacher reflected during an end-of-year interview that while, "the traditional approach" to APES instruction can be "very abstract and it's hard to relate to abstract things," KIA students "can put things in perspective … [They can] see meaning and purpose."

At the end of the year, a KIA student also related authenticity to engagement:

> "I thought [the KIA approach] was really effective, especially in government, because [our teacher was] preparing us to work in the government and be policy advisors … and they don't really do tests, so our projects were mainly based on real things that you might have to do as a policy advisor. I thought that was really cool and interesting."

At the same time, half the treatment teachers interviewed acknowledged that, without adaptation, the curriculum could be outdated or incompatible with some contexts. One teacher recommended projects set in urban environments so that urban students "could see themselves living in the projects as opposed to having them envision themselves in somewhere where they never see themselves." This teacher referred to the Oceans in Action project set in the Pacific Northwest, and the Farm project set in Iowa, as harder for students to relate to.

KIA students perceived civic engagement and environmental activism benefits

In response to the end-of-year interview question about how KIA affected them, almost one-third of treatment students talked about how their increased knowledge (i.e., deeper learning) effectually raised their awareness of real-life issues in politics and the environment. In addition, treatment teachers reported emphasizing development of students' knowledge necessary to engage civically (e.g., voting, petitioning, campaigning for a political candidate) more so than control, by 0.8 ES's.

Civic engagement was an expected benefit that we explored in more depth with treatment students. When students were questioned about how they benefited from KIA, what from KIA they could take into the future, and what effect KIA might have on their civic engagement, they spoke to the impact they perceived on their future behaviors and intentions as related to civic engagement and activism.

As important context, the 2016-17 school year wrapped around the historic November 2016 presidential election, in which Donald Trump won more electoral votes than Hilary Clinton to win the presidency. For many reasons, this election and aftermath were unlike any other in modern U.S. history, and provided atypical salience and relevance for APGOV students' civic engagement-related responses.

KIA APGOV students referenced civic engagement in terms of voting, citizen behaviors, rights, and political viewpoints. In contrast, when control APGOV students spoke about civic engagement—rather than talking about understanding why to vote and how students can enact their power by making their voices heard, as treatment students did—they referenced understanding of civil liberties and rights, becoming better people, and knowing how to be an involved citizen.

The majority of treatment APGOV students' civic engagement-related comments, discussed predominantly in the student group interviews, referenced increased likelihood of voting because of a newfound awareness of why to vote. For example, per one student:

> "This class definitely made me want to vote more. If young people voted more, then politicians would put more focus on the things that young people would be interested in. I might as well do my part."

Another student explained,

> "A lot of people that I talked to that said they don't vote because they think that their vote doesn't count, and I'm trying to explain to them how it works and how it does … It helps a lot more to make an argument of why you should do it."

APGOV students also described their course as helping them understand that individuals have the power to affect the political system. A student shared a newfound understanding that, "one person can make a difference," and another described wanting to "work on campaigns or contact local government in my area to see if I can become a little more active." This student described learning about the power of the individual in U.S. government through the KIA class: "I guess I just didn't realize that I actually do play an important role in what happens in my government." In another group interview, a treatment APGOV student similarly shared a newfound understanding that students "have to have our voice."

> "If we don't speak up and show what we want, then they're not going to change anything. This year, I've gone to more marches and been more active in my government, wanting to have things change."

When APES students in treatment group interviews discussed civic engagement, they spoke about becoming more resourceful and cutting down on resources use, as well as encouraging others to care about the environment. APES students in both the treatment and control groups referenced adopting

conservation behaviors—actions driven by increased awareness of how humans affect the environment. Students talked about changes they were trying to make in their own lives, as well as changes they planned to make in the future. Reflecting on the Eco-Footprint activity, one student described "cutting down on certain things I use, like with water, with oil and all of that. I cut down a lot."

Other APES students acknowledged that though they are not at the point in their lives where they make their household's consumer decisions, they plan for future changes. For example, one student was considering buying an electric car to "move more efficiently and cleaner rather than having coal mines and oil drilling and all these things that are really, really harmful to the environment." Other students talked about similarly weighing environmental considerations regarding electricity use and grocery shopping.

Students in KIA classrooms felt prepared for the AP examinations

When the school year began, interviewed treatment students across courses said they expected to feel prepared for their AP exams in May. After students had taken the exams, most treatment students continued to describe themselves as prepared. As context, compared to control, treatment students reported on the end-of-year survey somewhat less frequent AP exam preparation between September and the end of March (-0.18 ES's), though essentially the same frequency of preparation in April and May (-0.08 ES's).

Treatment students had various opinions regarding whether they were more prepared for the multiple-choice versus the free-response subsections of the exams; some said that FRQ questions were more difficult because they did not cover parts of the topics in class, while others said the multiple-choice items were more difficult because of the terminology and vocabulary they needed to memorize.

In response to questions about what helped them to feel prepared, most interviewed treatment students referenced "learning for the test" in the two to three months prior to the examination period through AP practice exams, study guides, and exam-taking strategies provided by their teachers. KIA projects were the second-most frequently referenced activity students described as helpful for exam preparation. Even while describing practice exams as the most helpful, some students acknowledged that tested content and skills were woven into the projects and labs throughout the year.

One student articulated the purpose of his/her APGOV class as, "less preparing us for the AP test but more preparing us to be politicians and go out into the [political] world."

One treatment teacher reflected on how students felt about the AP exam, describing how the KIA instruction approach and student engagement contributed to their preparation: Students "taking an active role in what they're learning" helped them remember content so that after the examination they understood the purpose of what they had been taught. The teacher said students' comments included, "Now I get it. Now it all fits together. I feel like I was really prepared for the test."

Treatment teachers felt KIA is aligned to AP frameworks though preparation was a concern

At the end of the 2016-17 school year, there were no differences between treatment and control teachers in reports of the extent to which the curriculum they had used was aligned to the AP

curriculum frameworks and AP examinations (0.01 and 0.09 ES's, respectively). This was in contrast to the year's start, when treatment teachers had anticipated incorporating AP curriculum frameworks into their 2016-17 AP instruction to a considerably lesser extent than control teachers (-0.52 SD). At the end of the year, treatment teachers also reported more alignment between their curriculum and Common Core State Standards for English Language Arts, Math, and the Next Generation Science Standards than did control teachers, with differences ranging from approximately 1.0 to 2.5 ES's. These differences did not exist at the beginning of the year.

However, despite reporting that the KIA curriculum aligned well with the exam, at the end of the year, one-third of interviewed treatment teachers said preparing students for the AP exam was a challenge. Being vigilant about covering everything for the exam was one of teachers' biggest concerns. A teacher described needing to be "purposeful" about covering all of the AP Curriculum Framework content, noting that "in project-based learning, we might not be learning every single piece of content that you're going to have to know."

Covering everything expected to be on the AP exam was a common concern, even when teachers felt like the curriculum was well aligned. In addition to planning connections between the AP exam and KIA projects, another strategy teachers described was double-checking that KIA students were retaining the "factual material" (e.g., vocabulary and concepts) they would need to know for the exam. One teacher described uncertainty about whether students are learning what they need to learn as a "consistent challenge." This teacher described a tension between the students "enjoying the experience," and having a "running inner monologue [asking] do they know what divided government means?" As the examination date drew closer, such concern or doubt may have driven the extra preparations and AP materials (e.g., practice exams) teachers provided for students.

In both classrooms of the teachers previously quoted, students in the end-of-year group interviews talked about the AP exam being easier than they expected. One student even said, "I feel like the exam was a lot easier than any test we've ever taken in this class." They used standard test resources, like study guides and practice exams, and their teachers helped them identify which content and skills most likely would be covered on the exam. Yet as one KIA group interview participant stated, "We were actually reviewing. In a lot of classes, your review is you [learn] new stuff."

Perceived benefits of KIA for teachers

Through KIA, teachers' understanding of PBL deepened, expanding their instructional repertoire. Treatment teachers' perceptions of KIA were positive, and almost all planned to continue using KIA and recommend it to others.

KIA improved teachers' understanding of PBL, expanding and deepening their tools

As expected, by the end of the year, treatment teachers' understanding of PBL had clarified and deepened. For example, a teacher described better understanding the role of driving questions as open-ended: "The kids really have to really create and do versus find these answers." Another teacher described an "eye-opening" shift in understanding of how to provide feedback: The realization that students could evaluate each other's work "really changed the way I did things in the classroom."

In response to a question about how KIA affected them as teachers—including their AP course, teaching, professional growth, and themselves—some treatment teachers responded by pointing out how KIA challenged them to develop their own critical thinking and apply it to their teaching. One teacher described KIA as "beneficial, because it does pull us out of our comfort zone," requiring teachers to "be more resourceful … not just regurgitate." Another described how "being able to work through someone else's curriculum"—rather than creating their own in which "you never really know if you're getting to the true heart of project-based learning"—was "the best thing this year."

The most prominent perceived benefit for treatment teachers was a new lens on curriculum and instruction, and the new set of "tools" for their "toolbox." While KIA was challenging for teachers, with both perceived advantages and disadvantages, most KIA teachers' big-picture takeaway was that KIA helped make their teaching more authentic and engaging for students, with students driving their own learning. One APGOV teacher said, "I definitely think [KIA] helps me think more about how to engage students, because I think one of the best parts of this was the fact that students are so engaged with the projects." Another APGOV teacher, speaking specifically about the "options and tools" KIA provided, gave the example of a coach suggesting "something simple like seating kids in a different way to encourage group participation."

KIA inspired teachers to start thinking about approaching their classes with more emphasis on student-centeredness. A KIA teacher, also a department chair, described KIA giving experienced teachers a "PBL lens" they could apply to both AP and non-AP courses "at least in little bits and pieces." This teacher described taking a "step back and looking with fresh eyes at some stuff that I'd been doing for a long time," and feeling "I can apply the PBL criteria and perhaps make this better." For example, the teacher felt that students publicizing their work makes them care about it much more so than if the teacher is the sole audience.

Many KIA teachers expressed the intention of using PBL more in their teaching moving forward. For these teachers, KIA demonstrated an impactful connection between curricular authenticity and student engagement:

> "I think like I'm going to look for PBL, either units or curriculum, in anything I teach from here on out … [because KIA] … really emphasizes how important it is to make what you do in the classroom real-world and engaging for students."

Even veteran teachers who have taught AP for a number of years, or who have been professional development trainers for other teachers, found they learned and benefited from KIA. A teacher with 10 years of experience described the paucity of professional development in contrast to KIA: "[It] feels like I learned a lot of new stuff this year." This teacher hoped to adopt a PBL approach in physics courses as well as APES.

Teachers' perceptions of KIA were positive

In alignment with their holistically-positive sense of the benefits of KIA for themselves as teachers, at the end of the year, all but one KIA teacher in our sample (96%) reported that they planned to use elements of KIA in their non-APES or APGOV courses. All but two (93%) said they would encourage non-AP teachers to use elements of the KIA approach to curriculum and instruction in their courses,

and plan to use KIA the next time they teach APGOV or APES, and 89% would encourage their school to adopt KIA curriculum for all AP classes.

Interview results conveyed the same message: When asked if they would recommend KIA to other teachers, most treatment teachers said they would. As a treatment teacher explained, KIA "definitely intrigued lots of people in the Social Studies department." Another teacher said, "I would absolutely recommend it," further explaining they would be "happy to sit down and talk about … what makes it PBL versus 'Let's do projects.'"

During end-of-year interviews, most treatment teachers said they incorporated KIA practices into other courses to a limited extent, with a smaller proportion making radical changes to their other courses. Reasons for using KIA in other courses included other students hearing about activities and wanting to take part, and teachers wanting to engage students.

Only 11% of teachers did not recommend encouraging their school to adopt a KIA-like curriculum for all courses. They described limiting factors to incorporating PBL into other classes, including students not being ready for PBL, the considerable amount of time necessary for preparation, and the KIA approach not working well with non-APGOV or APES course content. As an APGOV teacher explained the challenges of using the KIA approach in World History, "It's just not part of the curriculum. I couldn't see myself doing it successfully … The students in World History are not advanced students." While this teacher felt less-advanced students have difficulty doing the "higher-level tasks" required of PBL, they also described "successfully doing some debates with them."

Limitations to Year One implementation results

There are several limitations to our Year One implementation analysis. A primary limitation of our results is generalizability. The KIA RCT study was, by definition, a test of the efficacy of KIA under ideal conditions. The five participating districts were not representative of all districts offering AP courses. Rather, they are districts: 1) supporting a teaching and learning approach, and philosophy that align with the KIA approach; 2) offering AP courses at enough individual high schools to warrant inclusion in the RCT; 3) showing enough interest in KIA to agree to participate; and 4) requiring open-access AP course enrollment. Further, the teachers and schools choosing to participate in the KIA Efficacy Study were volunteering "early-adopters," and may not have delivered the courses in a way representative of large-scale implementation.

The second limitation is that all data from interviews, focus groups, instruction logs, and surveys is reported from the perspectives of individual principals, teachers, and students, and so is subject to the drawbacks of self-reported data. Teachers' self-reported responses about their KIA implementation may be particularly subject to potential over-reporting bias, as teachers could have said they implemented more KIA practices than they actually did. For this reason, we cross-reference, or "triangulate," data across sources to verify self-reported responses from one group against other groups' responses. Though we do triangulate responses, which highlights discrepancies, self-response bias is a concern.

Appendix GG: Year Two Implementation Analysis, Methodology and Results

Year Two implementation results were intended to align with two separate arms—experimental and non-experimental—of the Year Two portion of study. The experimental arm compares treatment teachers in their second year of KIA implementation during 2017-18 to control teachers in their first year of implementation in 2017-18. The non-experimental arm compares treatment teachers in their second year of KIA implementation during 2017-18 to non-experimental teachers in 2017-18 who were never exposed to KIA.

Experimental Arm

Teacher sample

In Year Two, for treatment teachers in their second year of KIA implementation, we carried over complier status from Year One, as this indicator reflects their participation in the intervention when it was offered to them. That is, a non-complier in 2016-17 could not become a complier in 2017-18, and vice versa. For control teachers, all "complied" with control status in the Year One because none received the treatment offer. In Year Two, we created a complier flag based upon teachers' participation in the KIA intervention in 2017-18. Of the 23 treatment teachers in their second KIA year in 2017-18, 20 were compliers (87%). Of the 30 control teachers in their first year with the KIA offer in 2017-18, 26 were compliers (87%).

Of the 55 experimental teachers who persisted the Year Two sample, 22 (9 in the treatment condition and 13 in control) completed teacher surveys at all three time points: 1) at the end of the 2015-16 school year (baseline); 2) at the end of the 2016-17 school year—after one year of implementation for treatment teachers, and business-as-usual conditions for control teachers—and; 3) at the end of the 2017-18 year—the second year of implementation for treatment teachers and the first year of implementation for control teachers. This subgroup of survey respondents was mostly comparable to the Year Two experimental teacher sample on measured characteristics, particularly in the areas of teaching experience and degree held (Table GG1). More than one-third of those surveyed taught APGOV (36%), while for the overall Maturation sample 42% taught APGOV. There were small differences on gender and race, with more women among the survey completers compared to the overall sample (73% compared to 64%), no Black teachers in the survey sample (versus 7% overall), and more Asian teachers in the survey sample (29% versus 14%).

Table GG1. Sample characteristics of teachers with surveys at all three time-points compared to the overall Year Two teacher sample

		Experimental teachers with surveys (n=22)	Overall Year Two teacher sample (n=53)
Course			
	APES	14 (64%)	31 (58%)
	APGOV	8 (36%)	22 (42%)
Education		5 of 22 missing education	5 of 53 missing education
	Bachelor's	3 (18%)	10 (21%)

Master's	12 (71%)	30 (63%)
Doctorate	2 (12%)	8 (17%)
Gender		
Female	16 (73%)	34 (64%)
Teacher ethnicity	5 of 22 missing ethnicity	11 of 53 missing ethnicity
White	12 (71%)	33 (79%)
Black	0 (0%)	3 (7%)
Asian	5 (29%)	6 (14%)
Hispanic	11 of 22 missing Hispanic or not	30 of 53 missing Hispanic or not
Yes	0 (0%)	2 (6%)
Years of experience		5 of 53 missing experience
Less than 2 years	0 (0%)	0 (0%)
2 to 5 years	1 (5%)	4 (8%)
6 to 10 years	5 (23%)	8 (17%)
More than 10 years	16 (73%)	35 (73%)

Analytic approach

Given the small sample sizes in both treatment and control groups, we limited our approach to a descriptive analysis, calculating means and standard errors for each survey item and scale. (See Appendix FF for a comprehensive description of our teacher survey analytic methods.) Because small samples can produce unstable mean estimates with wide standard deviations, all survey analyses are exploratory and must be interpreted with caution. We also limited our reporting to results of survey scales—that is, composite measures that included multiple survey items. Though descriptive statistics for survey scales from small samples still will be relatively unstable, composite measures are more reliable than single items. We examined means for each group at each time point, and included standard errors around the mean estimates to remind the reader of their lack of precision. No inferences about differences between groups or differences over time should be drawn from these results.

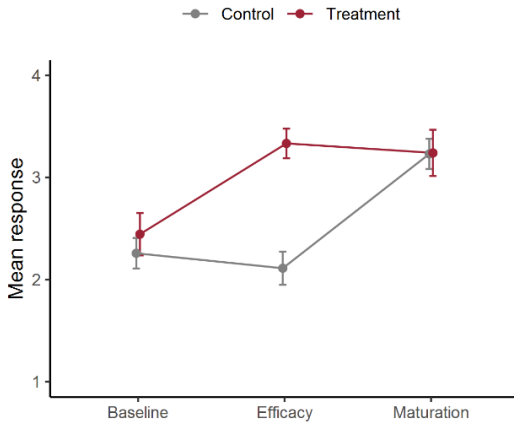
Results

Year One implementation results showed that treatment teachers meaningfully shifted their instructional practices in their first year of KIA implementation (Appendix FF). Implementation results from Year Two suggest that teachers in their second year implementing KIA did not revert to their pre-intervention practices, nor did they continue to further change their practices. Rather, for several key KIA-aligned instructional practices measured through self-report, teachers in their second year sustained their changed practices, while teachers implementing for the first time in 2017-18 changed their practices similarly to how treatment teachers did so in their first year of implementation. These trends resulted in similar levels of self-reported use of instructional strategies and practices between treatment and control teachers at the end of the 2017-18 school year.

This pattern, though descriptive and based on small samples, is exemplified by teacher responses to survey questions comprising a scale titled "KIA-like projects." The scale is a composite of questions about the frequency with which teachers assigned projects that extended over several weeks or

months, assigned interdisciplinary projects, implemented projects requiring students to draw from skills gained in other classes, and used role play and/or simulations. Figure GG1 shows the average response for each group of teachers (treatment and control) at each of the three time points (baseline, end of 2016-17, end of 2017-18). The greatest shift observed is in a teachers' first year of implementation, and treatment and control groups made similar shifts in each of their first years. Further, treatment teachers did not self-report returning to pre-intervention levels in their second year of implementation when they were no longer receiving professional development.

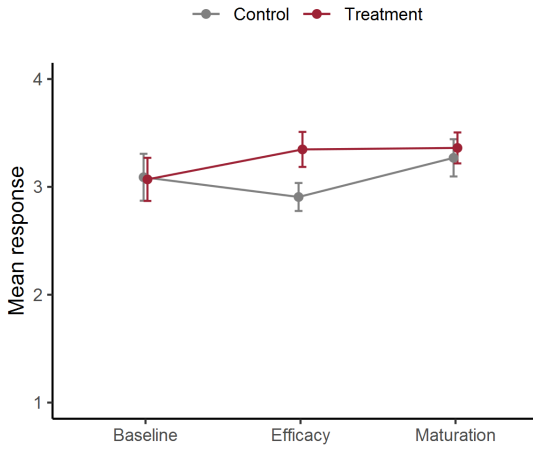
Figure GG1: KIA-like projects



Note. Error bars represent standard errors of the mean.

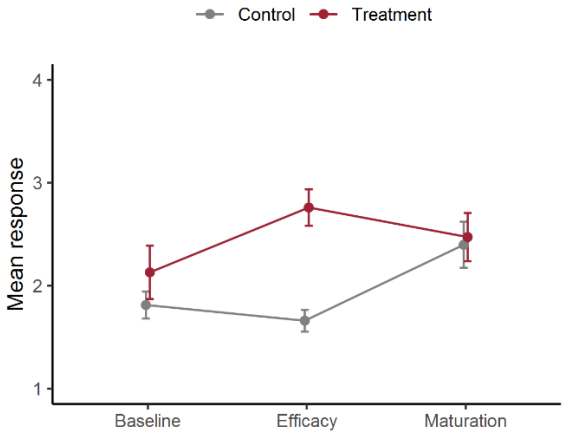
We observed a similar pattern in two other scales critical to PBL practice. The "relevance of curriculum to students" scale, a composite of eight items, included agreement with items such as, "students learn how to apply skills and course material to the world outside the classroom," "students work on learning tasks that are relevant to their life," and "students identify concerns and priorities for our community, state and nation that we all share." The scale of "sharing learning with outside audiences" includes presenting solutions for community challenges, presenting work to an audience of community members or adults, and working on a project that addresses a community problem or seeks to make a real change in the world. Figures GG2a and GG2b show a similar pattern for both of these constructs, with instructional practices changing in a teacher's first year of implementation, and treatment teachers sustaining those practices in their second year.

Figure GG2a: Relevance of curriculum



Note. Error bars represent standard errors of the mean.

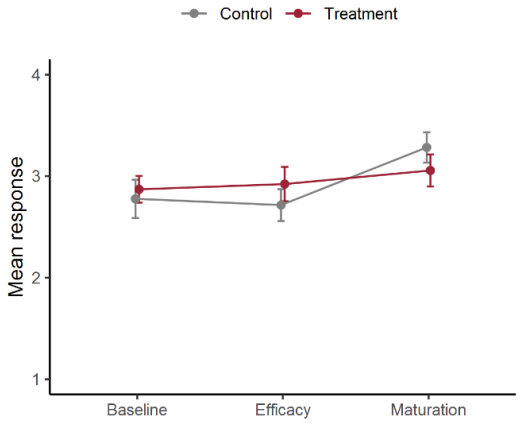
Figure GG2b: Share learning



Note. Error bars represent standard errors of the mean.

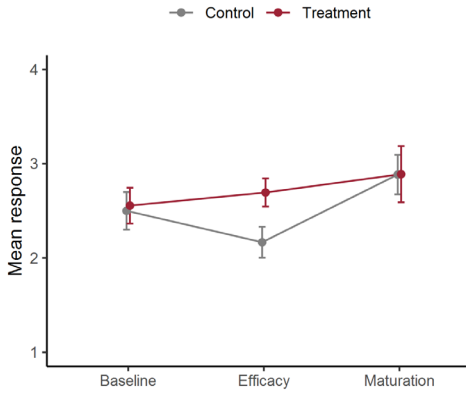
For "quality of groupwork," a 12-item scale composed of questions asking about groupwork-related practices (e.g., agreement with the statement, "I provided feedback to students on how well they work together."), the treatment group's responses increased marginally year to year, while the control group showed a more noticeable change in their first implementation year. Though a different pattern, the end result was similar—both control and treatment teachers reported similar practices that were aligned to KIA's core practices—at the end of the Maturation year. As we show in Figures GG3a and GG3b, the "practice research" scale—composed of items asking about frequency with which students worked on projects involving gathering information within and outside of school, and practicing research skills—also demonstrated a pattern in which treatment and control teachers reported similar levels of practice at the end of the Maturation year.

Figure GG3a: Quality of groupwork



Note. Error bars represent standard errors of the mean.

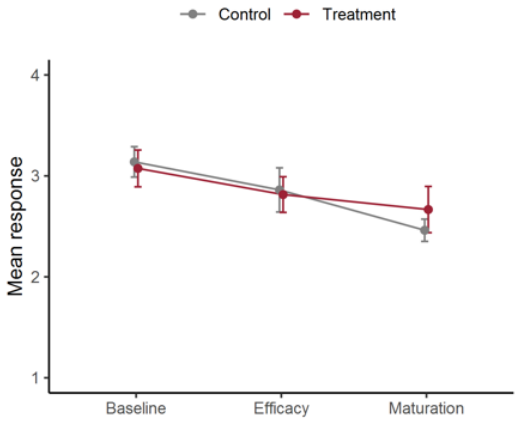
Figure GG3b: Practice research



Note. Error bars represent standard errors of the mean.

The transmission instruction scale was composed of three items: "I provided direct instruction," "Students listened to teacher present material and explain concepts," and "Students take notes about presented content/material." These results demonstrate decrease in transmission instruction over time for both groups (Figure GG4). This pattern of self-reported decreasing reliance on transmission instructional methods may exemplify a combination of KIA professional development with an unmeasured characteristic among teachers who volunteered for the KIA RCT—all were likely interested in incorporating more PBL and less transmission instruction. In 2016-17, both treatment teachers (under optimal professional development conditions) and control teachers (operating under business-as-usual conditions but knowing they would receive the treatment the following year) similarly reduced their self-reported use of transmission instruction. In 2017-18, under optimal conditions of professional development, control teachers' use of transmission instruction decreased even more. Also important, treatment teachers' instructional practices did not revert to pre-intervention levels in their second year; rather, their self-reports further aligned to KIA's principle of reduced transmission instructional practices.

Figure GG4: Transmission instruction



Note. Error bars represent standard errors of the mean.

Overall, the implementation results are consistent with the intensity of the professional development provided to teachers during their first year participating in the KIA intervention—including four full days in the summer, four full days throughout the year, and on-demand one-on-one coaching support—and lack of continued intensive PD in their second year. They also are consistent with the Year One impact analyses demonstrating KIA students outperformed non-KIA students on the AP examination by a significant and meaningful magnitude. Even so, given the complexity of shifting towards a PBL approach in the AP setting, none of this observed consistency was a given. Based on "implementation dip" literature, we might not have expected teachers to change so many of their practices during their first year using the approach.

Non-Experimental Arm

Implementation results from the experimental arm descriptively examined mean patterns over time; for example, the extent to which changes continued at the same pace into a teachers' second year and if teachers regressed to previous levels in their second year. Comparing responses between experimental and non-experimental teachers shed light the extent to which classrooms of KIA teachers differed from non-KIA classrooms. Though the Year One implementation analysis examined this same question, those analyses detailed the ways in which treatment classes differed from comparable teachers who volunteered to participate in KIA. This non-experimental implementation analysis examines differences between KIA teachers with two years of experience and teachers in the same districts teaching under business-as-usual conditions who did not participate in KIA in the first place for various, and mostly unmeasured/unknown reasons.

Teacher sample

Ten treatment teachers who persisted into the Year Two sample completed the required baseline (2015-16) and end-of-maturation-year (2017-18) teacher surveys.[38] From the universe of non-experimental teachers across all five districts, 22 completed baseline (2016-17) and end-of-year (2017-

[38] We did not require teachers to also have completed the 2016-17 survey because we only examined pre and post school year responses for the two groups of teachers.

18) surveys. We first explored how the sample of treatment teachers with surveys (contributing data to implementation analyses) compare to the full group of T2 teachers (used in impact analyses). Then we examined how the sample of non-experimental teachers with surveys (contributing data to implementation analyses) compare to the Round One matched comparison group of teachers in the non-experimental impact analysis.[39] Finally, we compared the surveyed T2 teachers with the surveyed N2 teachers.

The 10 treatment teachers with both baseline (2015-16) and end-of-Year Two (2017-18) survey results looked much like the overall group of T2 teachers on measured characteristics. As shown in Table GG2, among teachers with surveys and the overall T2 group, approximately 60% taught APES while 40% taught APGOV. Both groups were highly educated, with approximately 85% holding master's degrees or higher, and almost all had at least six years of experience teaching APGOV or APES as applicable. The survey sample was composed of more women than the overall T2 sample (70% compared to 52%). On race and ethnicity, there were some differences, with the surveyed sample not including any Black teachers yet including more Asian teachers proportionately than the overall treatment group in the Year Two sample.

Table GG2: Sample characteristics of treatment teachers in their second KIA year with surveys at baseline and end of Year Two compared to overall sample of teachers in their second KIA year

	T2 teachers with surveys (n=10)	Overall T2 teachers (n=23)
Course		
APES	6 (60%)	13 (57%)
APGOV	4 (40%)	10 (44%)
Education	2 of 10 missing education	4 of 23 missing education
Bachelor's	1 (13%)	3 (16%)
Master's	6 (75%)	14 (74%)
Doctorate	1 (13%)	2 (11%)
Gender		
Female	7 (70%)	12 (52%)
Teacher ethnicity	2 of 10 missing ethnicity	5 of 23 missing ethnicity
White	6 (75%)	15 (83%)
Black	0 (0%)	1 (6%)
Asian	2 (25%)	2 (11%)
Hispanic	4 of 10 missing Hispanic or not	14 of 23 missing Hispanic or not
Yes	0 (0%)	1 (11%)
Years of experience		2 of 23 missing experience
Less than 2 years	0 (0%)	0 (0%)
2 to 5 years	0 (0%)	2 (10%)
6 to 10 years	4 (40%)	7 (33%)
More than 10 years	6 (60%)	12 (57%)

[39] We did not repeat the Year Two non-experimental implementation analysis to include teachers from our Round Two matching.

In Table GG2, we compare the 22 non-experimental teachers who completed surveys to the 24 non-experimental teachers selected for the matched comparison group in the Round One non-experimental impact analysis. Demographic characteristics for the surveyed non-experimental teachers are self-reported and suffer from a high level of missingness that could not be resolved. Importantly, we could not link survey responses to the non-experimental group's administrative records. So, while we can compare the demographic characteristics of the non-experimental teachers with surveys (used in implementation analysis) to those in the matched comparison group (used for non-experimental outcomes analyses), we do not know the extent to which these two groups overlap.[40]

Table GG3 shows that N2 survey completers were somewhat different on measurable characteristics compared to matched N2 teachers used in the non-experimental outcomes analyses, though a high level of missingness on administrative records makes the comparison difficult. More surveyed N2 teachers taught APGOV compared to APES (59% to 41%), while more N2 teachers in the matched comparison group taught APES compared to APGOV(63% to 38%). Teachers in the two groups mostly were similar on years of teaching experience, as more than half in both groups possessed at least 11 years of teaching experience. Roughly half of both groups were female. Because of missing data in the matched teacher group, it is difficult to draw comparisons on other variables.

Table GG3. Sample characteristics of non-experimental teachers with surveys compared to the Round One matched non-experimental sample

	N2 teachers with surveys (n=22)[a]	N2 matched teachers (n=24)[b]
Course		
APES	9 (41%)	15 (63%)
APGOV	13 (59%)	9 (38%)
Education		14 of 24 missing
Bachelor's	5 (23%)	0 (0%)
Master's	15 (68%)	9 (90%)
Doctorate	2 (9%)	1 (10%)
Gender		4 of 24 missing
Female	9 (41%)	10 (50%)
Teacher ethnicity	1 of 22 missing	5 of 24 missing
White	20 (95%)	14 (74%)
Black	1 (5%)	1 (5%)
Asian	0 (0%)	1 (5%)
Other[c]	0 (0%)	3 (16%)
Hispanic		13 of 24 missing
Yes	1 (5%)	2 (18%)
Years of experience		5 of 24 missing
Less than 2 years	0 (0%)	0 (0%)

[40] In other words, the non-experimental teachers providing implementation data may be entirely non-overlapping, somewhat overlapping, or mostly overlapping with those contributing to the non-experimental impact analyses.

2 to 5 years	5 (23%)	5 (26%)
6 to 10 years	4 (18%)	2 (11%)
More than 10 years	13 (59%)	12 (63%)

[a] Based on self-report
[b] Based on administrative data
[c] Includes Native American (1), Hawaiian/Pacific Islander (1), and Multiple ethnicities (1)

There were similarities and differences between T2 teachers and N2 teachers who returned surveys (Table GG4). Both groups had similar variation in education levels, with most holding a master's degree. There is only one Hispanic teacher across both groups. Teachers in both groups were highly experienced, such that more than half in both groups have at least 11 years of experience teaching AGPOV or APES. In the T2 group, all had at least six years of experience, while in the N2 group, three-quarters (77%) had as much experience. A greater proportion of the T2 group (60%) taught APES, while 59% of the N2 group taught APGOV. In terms of differences, T2 teachers completing surveys were mostly female (70%), compared to 41% in the N2 group. And while the N2 group is 95% White with one Black teacher, the T2 group was 75% White with two Asian teachers and two for whom race/ethnicity is unknown.

Table GG4: Characteristics of teachers contributing to the implementation analysis

	T2 teachers with surveys (n=10)	N2 teachers with surveys (n=22)
Course		
APES	6 (60%)	9 (41%)
APGOV	4 (40%)	13 (59%)
Education	2 of 10 missing education	
Bachelor's	1 (13%)	5 (23%)
Master's	6 (75%)	15 (68%)
Doctorate	1 (13%)	2 (9%)
Gender		
Female	7 (70%)	9 (41%)
Teacher ethnicity	2 of 10 missing ethnicity	1 of 22 missing
White	6 (75%)	20 (95%)
Black	0 (0%)	1 (5%)
Asian	2 (25%)	0 (0%)
Hispanic	4 of 10 missing Hispanic or not	
Yes	0 (0%)	1 (5%)
Years of experience		
Less than 2 years	0 (0%)	0 (0%)
2 to 5 years	0 (0%)	5 (23%)
6 to 10 years	4 (40%)	4 (18%)
More than 10 years	6 (60%)	13 (59%)

Analytic Approach

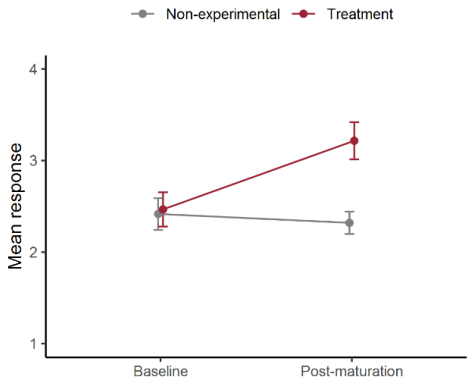
Again, given the very small sample sizes, we restricted our approach to a descriptive analysis, calculating means and standard errors for each survey item and scale. And for the same reasons stated above, we limited our exploration of results only to survey scales. We note that the amount of time between baseline and end-of-Year Two surveys varied for the two groups: two years for treatment teachers, one year for non-experimental teachers. Audiences should not draw inferences about differences between groups or over time.

Results

Two consistent patterns are revealed when comparing survey responses between treatment teachers with two years of KIA experience against non-experimental teachers never exposed to KIA. First, treatment teachers (who returned surveys) changed their instructional approaches to align with KIA's over the course of two years, while non-experimental teachers mostly did not change their instructional approaches from the baseline to the end-of-year survey. Second, in alignment with descriptive sample characteristics shared throughout these appendices (2015-16 for treatment, 2016-17 for non-experimental), treatment teachers and nonexperimental teachers differed on two key self-reported measures of teaching behaviors—frequency of activities intended to develop students' critical thinking and quality of classroom discussion—that may shed light into T2 teachers' motivations for originally enrolling in the KIA Efficacy Study in 2016.

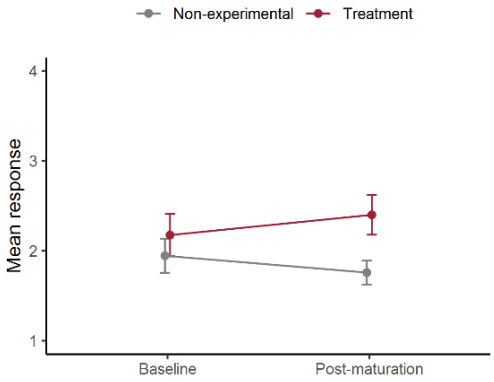
Instructors' responses on several scales exemplify how treatment teachers changed their practices over the two-year period compared to the lack of change among non-KIA teachers. As shown in Figures GG5a through GG5d, this pattern emerges on the scales describing "KIA-like projects," "sharing learning with outside audiences," "practice research," and "relevance of curriculum to students." Because these constructs are central to the KIA intervention—including curriculum and professional development—these patterns are expected. How treatment and non-experimental teachers (who may have already been using project-based learning pedagogies in their classrooms) would compare was unknown.

Figure GG5a: KIA-like projects



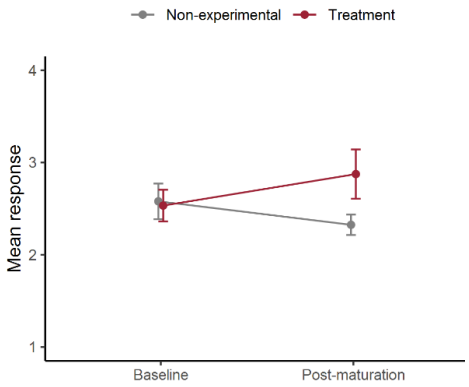
Note. Error bars represent standard errors of the mean.

Figure GG5b: Share learning



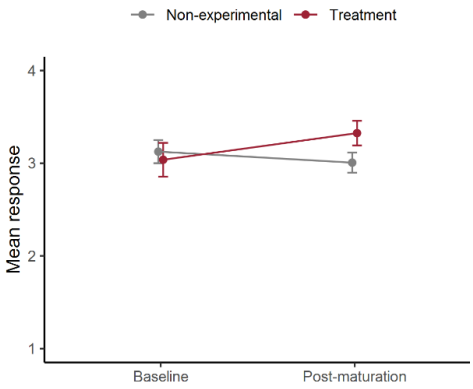
Note. Error bars represent standard errors of the mean.

Figure GG5c: Practice research



Note. Error bars represent standard errors of the mean.

Figure GG5d: Relevance of curriculum to students

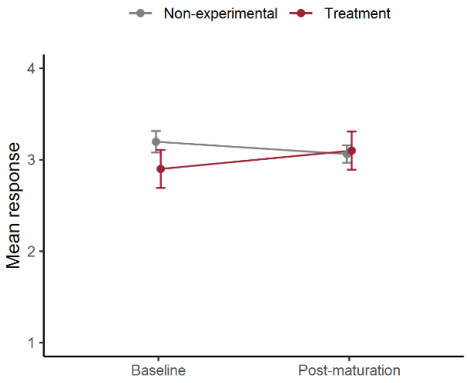


Note. Error bars represent standard errors of the mean.

On other scales, results show treatment teachers changing their instruction in a way that makes them look more like the non-experimental teachers. We see this pattern in scales measuring frequency of activities intended to develop "critical thinking" (how often teachers ask students to critique peer

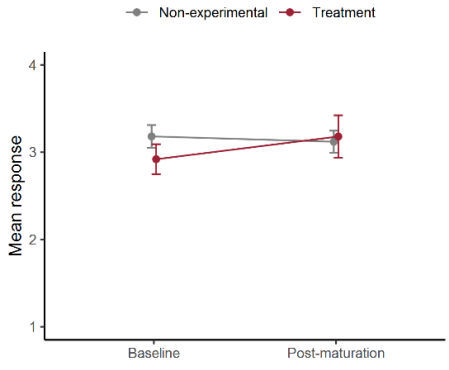
work, explain their reasoning or thinking, analyze data to make inferences or draw conclusions, and develop thinking skills) and "quality of classroom discussion" (how often students build on each other's' ideas during discussions, use data to support their ideas, provide constructive feedback to peers, participate in a discussion, and how often teachers successfully foster productive conversations, arguments, or discussions). These scale results, seen in Figures GG6a and GG6b, suggest two ways in which experimental teachers self-reported deficiencies in their teaching proficiency, relative to non-experimental, for which they improved—meeting N2 teachers' self-reports—after two years participating in the KIA intervention.

Figure GG6a: Critical thinking



Note. Error bars represent standard errors of the mean.

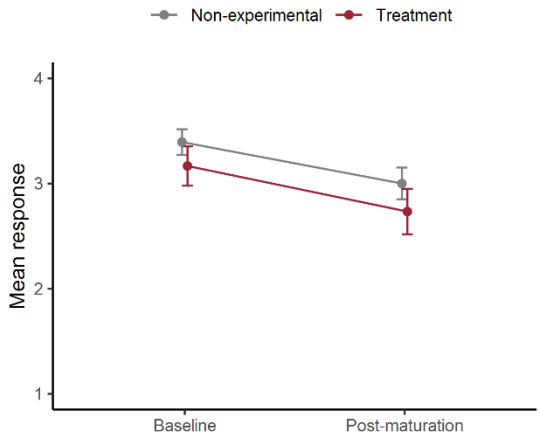
Figure GG6b: Quality of classroom discussion



Note. Error bars represent standard errors of the mean.

In a third pattern, relevant to the "transmission instruction" scale we see, from baseline, both groups reducing their reliance on transmission practices by the end of 2017-18.

Figure GG7: Transmission instruction



Note. Error bars represent standard errors of the mean.

Limitations to Year Two Implementation Results

Implementation results are based on small samples of teachers who responded to survey requests. For the experimental arm, teachers responded to surveys at three time points; in the non-experimental arm, teachers responded twice. Responding teachers may have differed in unmeasured ways from those who did not respond. For example, motivation may have been related to their instructional practices.

Means for the treatment group are based on nine teachers in the experimental arm and 10 teachers in the non-experimental arm. Means for the control group are based on 13 teachers in the experimental arm and 22 teachers in the non-experimental arm. Such small numbers of individual responses lead to greater uncertainty in reported means. We caution readers against interpreting differences in the means between groups or over time. Despite this caution, the patterns observed in the scales presented above emerged across multiple measures and align with other results presented in this report.

An additional limitation relevant to the non-experimental arm is that baseline means reported are calculated from different years: at the end of 2015-16 for treatment teachers; for non-experimental teachers, the end of 2016-17. As such, end-of-year self-reports reflect two years of changes in teaching for treatment teachers against one year for non-experimental teachers. Though we visually align these means in figures GG6a through GG8, time may confound interpretation of these patterns.

Appendix HH: Teachers' Use of Sprocket

KIA teachers used the Sprocket online curriculum portal during the 2016-17 school year, as described in this appendix. Of note, Lucas Education Research has continued to develop and iterate upon Sprocket since then.

Treatment teachers received access to the Sprocket online curriculum portal one or two days prior to the first day of their district's Summer Institute. We examined teachers' Sprocket activity in Year One from the date their access began, differing across districts from June through August, until the end of May 2017. Out of the 31 teachers using Knowledge in Action supports in some way, either by participating in professional development and/or using Sprocket, all used Sprocket. Our description of Sprocket usage is based on these 31 teachers: 16 teaching APES and 15 APGOV. Control teachers did not have access to Sprocket; therefore, we do not make comparisons between treatment and control teachers.

Sprocket offered teachers a number of tools to support their Knowledge in Action approach to APGOV and APES curriculum and instruction. Through the portal, teachers could look at curriculum pages, download curriculum and/or instructional materials, upload materials to share with others, adapt existing materials to share with others, participate in an online forum discussion, request support, and/or organize their calendar.

On average, teachers logged activity on 58% of all possible weeks of the school year, between the week of their summer 2016 Institute and the end of May 2017. Looking at the distribution of usage based on the percent of weeks in which activity was logged, four teachers accessed Sprocket between one-third of weeks or less (light usage), 14 teachers did so between 34-66% of weeks (moderate usage), and 13 accessed Sprocket on two-thirds or more of available weeks (heavy usage). Light-users, on average, used Sprocket for 15% of all possible weeks. Of the four light-users, two taught APGOV and two taught APES. Moderate users, on average, used Sprocket for 54% of the possible weeks. Of the 14 moderate users, nine taught APGOV, five APES. Heavy-users used Sprocket for an average of 76% of possible weeks; of the 13 heavy-users, four taught APGOV and nine APES. We summarize these descriptions of light, moderate and high user profiles in Table HH1.

Table HH1: Profiles of low, medium, and high usage of Sprocket

	Range of percentage of weeks used	Number of teachers		Average percentage of weeks accessed Sprocket from all weeks possible
		APGOV	APES	
Light	0-33	2	2	15
Medium	34-66	9	5	54
Heavy	67-100	4	9	76

In this section, we describe teachers' use of Sprocket based on metrics calculated from system usage data. We describe usage overall across all teachers, and separately for light- and heavy-users to contrast those groups. Teachers' feedback to Sprocket was predominantly positive; indeed, they rated it as the most helpful of the supports provided through participation in the KIA RCT.

Curriculum page views

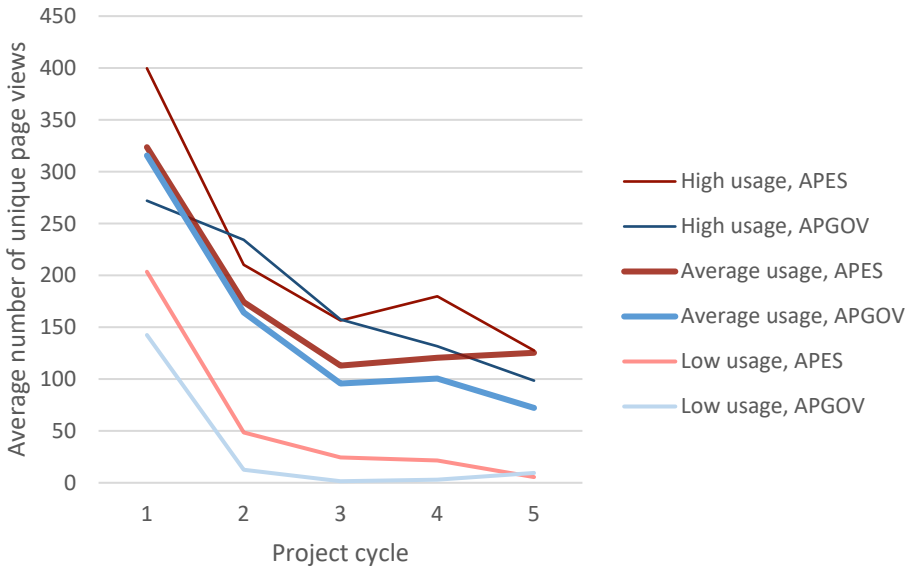
Teachers may have visited a single curriculum page (for example, one providing driving questions) multiple times—navigating through a scope and sequence, revisiting a page for more information, using different computers to access information, etc. To calculate a meaningful user metric conveying something about the depth and breadth to which teachers used the system, we count a page as viewed or not, no matter how many times the teacher viewed the page. This way we can count the total number of pages viewed without overcounting when teachers viewed the same page repeatedly. We refer to this metric as "unique" page views.

There was considerable variation in teachers' time spent on unique curriculum page views by course: APES teachers, on average, spent nearly twice the time as APGOV teachers. APES teachers spent, on average, 11.3 hours (678 minutes) viewing curriculum pages, compared to 5.8 hours (350 minutes) among APGOV teachers.[41]

Teachers' average viewing of curriculum pages was highest at the beginning of the year, for the first project, and followed a downward trend throughout the year and across the subsequent four projects (see Figure HH1). Note that teachers used Sprocket during the Summer Institute, which contributed to the highest number of views for the first unit. All teachers viewed at least one curriculum page of the first and second units, 30 of 31 teachers viewed at least one page for the third unit, and 29 of 31 viewed curriculum pages in units four and five. The average number of pages viewed per unit started at 320 for the first, declined to 157 for the second, and to less than 120 for the final three. These patterns indicate Year One complier treatment teachers were using Sprocket less as the year progressed.

[41] The Sprocket portal, after 10 minutes of inactivity, times out.

Figure HH1: Average number of unique page views per project cycle for light-usage, average-usage, and heavy-usage teachers



Light APES users averaged 53 unique curriculum page views, while light APGOV users averaged 44, compared to 134 unique curriculum page views among heavy APES users and 119 among heavy APGOV users. Light APES users also spent more time than light APGOV users, 4.7 hours compared to 3.2 hours—both of which were less than half the time spent by heavy-users: 13.9 hours for APES teachers and 8 for APGOV teachers.

Downloading, uploading, and adapting instructional material files

To fully engage with all that Sprocket offers, teachers should be uploading and downloading documents from Sprocket's extensive library of "instructional materials" (e.g., day-by-day project lesson plans, assessment rubrics, lecture PowerPoints), and making curricular adaptations within the platform. All teachers viewed these files, with no teacher viewing fewer than six throughout the year. However, usage was proportionally low, with 55 (out of hundreds) as the average number of unique files viewed throughout the entire school year. One teacher viewed 228 unique files, but the median (a metric more representative of the typical teacher) was 46. Among light-users, the average number of files viewed was 36; for heavy-users, 58.

Though 28 of 31 teachers downloaded at least one file, the median number of downloads was 18. One light-user downloaded 483 files, inflating the mean number of downloaded files for that group to 150 (and overall to 46). Notably, while the typical heavy-user downloaded 17 files, the typical light-user downloaded 56. This downloading pattern in the light-user profile group, in conjunction with their below-average metrics of files viewed, suggests these teachers downloaded batches of materials sight unseen before sorting through the files offline.

While 30 of 31 teachers uploaded at least one file to share with other teachers and coaches, and one teacher uploaded 16, on average teachers uploaded 5.8. Light-users uploaded, on average, three files; heavy-users more than six.

Few teachers completed a full adaptation cycle through Sprocket, which involves downloading files, editing them, uploading the edited version, and explaining the reason for the adaptation (Table HH2). However, interview and survey data indicate teachers frequently adapted outside of the Sprocket context. Across both courses, 35% (n=11) of teachers completed an adaptation through Sprocket, though there was a course-specific difference, with 20% (n=3) of APGOV teachers completing an adaptation cycle compared to half (n=8) of APES teachers. Among teachers making an adaptation, the average number of adaptations was three. Light-users never made an adaptation, while heavy-users averaged 2.6.

Table HH2: Instructional material file views, downloads, uploads, and adaptations.

	Files Viewed		Files Downloaded		Files Uploaded		Sprocket Adaptations	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Average user	55	46	46	18	5.8	5	3	2
Low-user	36	29	150	56	3.0	2	0.0	0
High-user	58	46	31	17	6.3	6	2.6	2

Sprocket online community forum

Teachers also could use Sprocket to participate in the online community forum. Given the intent to use the forum to build teacher community, overall participation was low. All teachers but one posted to the forum at least once—but on average, teachers only posted 10 times throughout the year (seven times among light-users, 14 among heavy-users).

Sprocket support requests

Teachers also could use Sprocket to request support from coaches. Only one teacher never requested a support ticket, but among the 30 who did, the average number of tickets requested was 11—indicating the option to request support through Sprocket was a popular feature. Of those tickets, most were to request curriculum support. Light-users requested, on average, eight—seven of which were to request curriculum support from coaches, while two of the four low-users each requested one ticket. Heavy-users requested an average of 14 tickets, 10 of which were for curriculum support. All except for one of the top users requested to submit a video, indicating considerable overlap between Sprocket high-users and teachers participating in virtual instructional improvement coaching cycles.

Sprocket calendar

Approximately one quarter (26%) of teachers used the Sprocket calendar feature. However, there was a course-specific difference in its use, with 44% (n=7) of APES teachers using the calendar, compared to only one of 15 APGOV teachers. One teacher used the calendar heavily, logging 45 events. Of all teachers using the calendar, most only logged 2 or 3 events. Low-users did not use the calendar feature.

Course-specific differences in Sprocket frequency of use

Overall, teachers' frequency of use of Sprocket's resources did not differ in a meaningful way between APGOV and APES, save for the three instances explicitly referenced. Regarding time spent on curriculum pages, frequency of adaptations, and calendar use, APES teachers' Sprocket use was more frequent.

Sprocket usage summary

- APES teachers spent nearly twice as much time on the portal (11.3 hours) than APGOV teachers (5.8 hours), predominantly through viewing curriculum pages. APES teachers also adapted through Sprocket and used the calendar feature more frequently, though with the caveat that calendar use was minimal across all teachers.
- The typical teacher did not frequently download, upload, or adapt files. Notably, low-usage teachers downloaded more files than other teachers, indicating that downloading characterized how they used Sprocket.
- Teachers' use of Sprocket's online community and calendar features was minimal.
- Teachers frequently used Sprocket to request curriculum support.

Appendix II: KIA Professional Development Description, Participation, and Lessons Learned

KIA teachers participated in PBLWorks' KIA professional development program during the 2016-17 school year, as described in this appendix. Of note, PBLWorks has continued to develop and iterate upon its program since then.

The Buck Institute for Education (BIE), founded in 1987 and headquartered in Novato, California, is a non-profit organization devoted to developing instructional practices for project-based learning (PBL) and supporting teachers, schools, and districts worldwide in their use of PBL. In 2015, Lucas Education Research (LER) contracted the Buck Institute—now named PBLWorks—to develop and provide Knowledge in Action (KIA) professional development. LER selected PBLWorks from among just a few existing providers of PBL professional development with the capacity, at that time, to support the number of teachers and schools projected to take part in the KIA Efficacy Study. To develop teachers' ability to use research-driven PBL teaching practices, support their curriculum and lesson planning, and integrate new KIA teachers into the larger PBL community, the two organizations collaboratively designed the KIA Summer Institute, ongoing Professional Development Sessions, and a coaching program.

This Appendix describes the elements of the KIA professional development program, documents teachers' participation in each element, and describes lessons learned about implementation of the program. Data sources informing this chapter include: written professional development materials; teacher surveys administered by the research team at the end of the 2016-17 school year; teacher satisfaction surveys administered by PBLWorks at the end of the Summer Institute and each Professional Development Session day; observations conducted in one district; data about teachers' use of the Sprocket online curriculum portal; and interviews with teachers, LER staff, and PBLWorks coaches and staff. We conducted interviews with teachers at the school year's beginning and end, with coaches at the middle of the school year and end, and with LER staff three times throughout the year.

Implementation of Knowledge in Action Professional Development

We begin by describing the KIA professional development program components: the KIA Summer Institute; Professional Development Sessions held four times throughout the school year; and coaching. As part of the description of each element, we document teacher participation rates.

The overarching objectives of KIA professional development, in all forms, were to: 1) familiarize teachers with the KIA design principles, curriculum, and resources; 2) support teachers' planning for KIA curriculum and instruction; and 3) develop teachers' PBL instructional abilities. Critical to the first and second objectives, KIA professional development emphasized to teachers how the curriculum is not "scripted." Rather, successful KIA teaching and learning depends upon instructors adapting the curriculum and materials to their specific classroom contexts and students. To inform the third objective of developing teachers' PBL instructional practices, instead of simply preparing them to teach the KIA curriculum, PBLWorks integrated the Buck Institute's "Gold Standard" definitions of PBL design and practice into all aspects of the professional development program.

Summer Institute

Between June and August 2016, the PBLWorks offered a KIA Summer Institute in each of the five districts, all held locally in school classrooms and district office spaces over the course of four eight-hour weekdays. On average, there was one PBLWorks coach for every three teachers, allowing for a high degree of personalized attention during one-on-one and small group activities. To deepen coach-teacher relationships and help the coaches learn about individual teachers' instructional practices, the same coach supported the same three teachers during the Summer Institute, subsequent Professional Development Sessions, and one-on-one support. With a total of three teachers to support, the reasonable ratio gave coaches ample time to work with each teacher.

Summer Institute attendance was strong, with daily attendance among the 31 participating teachers ranging from 90 to 97%, as shown in Table II1.[42]

Table II1: Summer Institute Attendance

Day	Number in Attendance	Percentage of Total
1	30	97
2	30	97
3	29	94
4	28	90

Though there was variation across districts, individual group sizes within each district were small in general, with an average of approximately six teachers per district taking part in each Summer Institute day.

Summer Institute learning objectives

The driving question guiding the Summer Institute was: "How can we adapt and implement KIA curriculum to *meaningfully engage our students*?" Within the context of the driving question, the purposes of the Summer Institute were to: 1) introduce teachers to the KIA design principles, curriculum, and resources; 2) support teachers' planning for KIA curriculum and instruction; and 3) develop teachers' PBL teaching capacity.

A considerable focus in the Summer Institute—approximately 12 hours across the four days—was familiarizing teachers with PBLWorks' "Gold Standard PBL Design Elements" and "Project Based Teaching Practices." The Gold Standard Design Elements are a comprehensive, research-based model for best practices in PBL instruction. The seven elements include the following (from Larmer, Mergendoller, & Boss, 2015):

1) Projects pose a "driving question" that is challenging and relevant to students' lives.
2) Projects promote a sustained inquiry that takes time, requires multiple forms of research (e.g., reading a book, doing field interviews, or searching on the Internet) and involves an iterative process in which answering a question leads to more questions and a deeper learning experience.

[42] Of the 31 analytic sample treatment teachers participating in the KIA program (described as "compliers"), 30 took part in at least one form of professional development while one used Sprocket to access KIA curriculum but never participated in a professional development activity.

3) Projects involve engagement in authentic, "real-world" learning experiences.
4) Students have "voice and choice"—input into many aspects of projects with the goal of developing student ownership for their learning.
5) Teachers build time into instruction for students to reflect on their learning, how they are learning, and why they are learning.
6) Students create high-quality work through critique and revision, including from peers, teachers, and outside experts.
7) To meet the PBL Gold Standard, projects should result in a public product (e.g., presentation, debate, display, etc.) to present to students' families and/or community.

Summer Institute activities

On the first day of the Summer Institute, coaches provided teachers with an overview of KIA, describing its history, design principles, and the five project units. On subsequent days, coaches led teachers through deeper exploration of the first project and engagement in project planning. PBLWorks expected teachers to produce two work products by the end of the institute: one individual and the other done as part of a group. As individuals, teachers planned their first KIA unit, and with their groups they adapted the unit in two ways to align with PBL instructional best practices. Also during the Summer Institute, LER staff introduced teachers to the Sprocket platform, explaining the its structure and features, including editing, searching, organizing, sharing, and community tools.

PBLWorks and LER designed the Summer Institute to have a different focus each day as a way to develop teachers' understanding of PBL and teaching abilities. The four days' foci were, respectively, "experiencing," "finding," "sharing," and "refining." On the first "experiencing" day, so teachers could experience a KIA lesson from their classes' perspective, they assumed the role of students while coaches assumed the teachers' roles. Teachers-as-students worked in small teams of approximately 3-4 on an abbreviated KIA project related to their subject area, then presented to the larger group. Teachers had multiple opportunities throughout the Summer Institute to view PBL practices from the student perspective and consider how to apply the experiences to their own classroom. After most of the activities, coaches asked teachers how they might use the activity in their classroom with students.

The focus of the second day of Summer Institute was "finding" or identifying, PBL elements within the KIA curriculum. On the third day, focused on "sharing," coaches introduced teachers to PBLWorks Gold Standard PBL Teaching Practices. On the fourth day, "refining," teachers spent approximately three hours preparing to teach the first KIA unit to their students.

The Summer Institute schedule on days 2-4 included dedicated planning time as a means for teachers, while in immediate proximity to supportive coaches and peers, to prepare for their 2016-17 school year AP U.S. Government (APGOV) or AP Environmental Science (APES) classes. Coaches suggested ways for teachers to structure their planning time, including through provision of a task list of the deliverables they should aim to draft by the end of each block of planning time. They also assisted teachers' identification of personal time-management goals—particularly important, as shifting from traditional lecture-based AP instruction to KIA AP instruction placed intensive time demands on prior cohorts of KIA teachers. Suggested deliverables included pacing calendars on Sprocket for the first project unit, documentation of daily learning targets for students, and adaptations

for the first unit plan. Coaches also encouraged teachers to use the Understanding by Design (UBD, 2018) framework to "backwards plan" based on desired results, evidence of student learning, and a plan for student learning. On the final day of the Summer Institute, teachers presented to, and received feedback from, their peers and coaches on the KIA planning and adaptation work produced during the week.

Appendix JJ provides greater detail on the Summer Institute's day-by-day activities.

Professional Development Sessions

To build upon the Summer Institute foundation, PBLWorks and LER designed four full-day, in-person KIA Professional Development Sessions, held locally in each district throughout the school year. PBLWorks coaches led Professional Development Sessions in district classrooms or offices on weekdays, except in one district where sessions took place on Saturdays (following district norms). Districts paid for the substitutes or Saturday pay, as applicable. The coaches continued working with the same teachers with whom they worked during the Summer Institute.

Out of the 31 teachers participating in the KIA program, Professional Development Session attendance ranged from 90% for the first session to 68-74% attendance in subsequent sessions (Table II2). Overall numbers of teachers across districts ranged from 21 to 28, translating into small group sizes, an average of approximately 4-6 teachers within each district.

Table II2: Professional development session attendance

Session	Number in attendance	Percentage of total
1	28	90
2	21	68
3	21	68
4	23	74

PBLWorks staff said the explanations for absences varied, such as less teacher "buy-in" in one district and, in another district, teachers having personal conflicts and administrative needs. Teachers from three of the four districts holding Professional Development Sessions on weekdays said being out of the classroom challenged their pacing, as substitutes did not follow their KIA lesson plans, with several citing this reason for not attending.

Professional Development Session learning objectives

The driving question across the four Professional Development Sessions was: "How can we adapt and implement the KIA curriculum for students to produce high-quality work?" To address the driving question, during each of the four KIA Professional Development Sessions, teachers engaged, with support from their coaches and peers, in the following three processes:

1) Reflecting on previous projects to inform instruction for future work—Teachers reflected on their teaching and students' learning of the prior project unit as a way to guide planning and delivery of their next unit.
2) Implementing effective instructional strategies—Teachers augmented their understanding of the PBLWorks PBL Gold Standard Teaching Practices ("Building the Culture," "Scaffolding

Student Learning," and "Sustained Inquiry") to further develop their abilities to teach KIA material.

3) Adapting KIA projects—Teachers adapted KIA curriculum and instructional materials to meet their students' needs based on their reflections and development of their instructional practices.

Professional Development Session activities

As part of these three processes, the sessions included teachers' production of "deliverables," including lesson scaffolds, assessments, and curricular adaptations.

The focus of the first and second sessions was building a classroom culture that supports high-quality student work. The third session focused on how to promote high-quality student work by "scaffolding," or providing specific learning supports. And the fourth and final session—taking place 4-6 weeks before the AP examinations—focused on how to use inquiry-based instructional and assessment approaches to support student learning.

The Professional Development Sessions continued to build teachers' knowledge about PBL, with an emphasis on producing a culture of high-quality student work. In the Appendix, as with the Summer Institute, we provide day-by-day detail about the four sessions.

Coaching

Coaching provided as part of the KIA Efficacy Study included:

- Four opportunities for each teacher to engage in one-on-one virtual sessions with his/her coach as part of a three-part "improvement cycle."
- Four opportunities for each teacher to engage with his/her coach in one-on-one virtual "planning meetings" devoted to planning, adapting, and implementing the curriculum.
- Individual and group in-person coaching during the Summer Institute and the four Professional Development Sessions.

Coaches also encouraged teachers to request support at any time via email, phone, or Sprocket. As mentioned, on average, each coach worked with just three teachers over the course of the year.

Preparation and support for KIA coaches

Prior to the first Summer Institute, held in early June 2016, PBLWorks offered a three-day "KIA coach boot camp" designed to acclimate coaches to the KIA curriculum and professional development program. Buck Institute staff required the five (out of 10 total) coaches who had not previously taught KIA to complete 20 hours of curriculum study and submit an artifact of their learning.

The PBLWorks provided additional informal training during meetings with coaches. Coaches worked with the manager and colleagues to "troubleshoot" issues encountered in the field and share successful practices.

Virtual coaching improvement cycles

Prior to the KIA Efficacy Study, coaching via virtual means (i.e., teleconference and telephone) was not a standard feature of PBLWorks' professional development approach. Rather, LER and PBLWorks together developed a virtual coaching approach specifically for the Efficacy Study.

In its intended design, the first step of each virtual coaching "improvement cycle" (Knight, 2007) was a pre-observation meeting, during which the teacher and coach reflected on elements of the teacher's practice in need of improvement and considered how to address the targeted area(s). Next, the coach and teacher developed a "Theory of Action" for improving the teacher's practice. The coach then observed a lesson the teacher had recorded using an LER-provided video camera. Guided by the Theory of Action, the coaches identified, or "tagged," moments to review with teachers, then transcribed, or "scripted," segments of the lesson to discuss (Aguilar, 2013). According to Aguilar, the purpose of tagging and scripting is to provide "low inference" data to review with the teacher. PBLWorks also offered teachers an alternative to video, such as submitting student work for analysis. Finally, during the post-observation review, the teacher and coach analyzed the teacher's personal Theory of Action in the context of the tagged segments and relative to the PBL Gold Standard elements.

In practice and despite coaches' efforts, the teachers did not engage with virtual coaching improvement cycles as intended. Table II3 displays the completion rates for each of the four Improvement Cycles. More than half (55%) the teachers completed the first and second full cycles; that is, they participated in the pre-observation meeting, submitted a video or a student assignment, and participated in the post-observation feedback. Participation declined as the year progressed, with only two teachers (6% of the full sample of 31) finishing all four cycles; those two also were the only ones to complete the fourth cycle. Thirteen teachers did not complete any full cycles, two completed one full cycle, six completed two full cycles, and eight completed three.

Table II3: Teachers' completion of improvement cycles

Cycle	Number of teachers completing	Percentage of total
1	17	55
2	17	55
3	10	32
4	2	6
All	2	6

Teachers used both video and alternative forms of data for analysis with their coaches, with video used less in cycles three and four. The "Lessons Learned" section of this chapter discusses challenges to teachers' participation in virtual coaching.

Virtual planning meetings

Virtual planning meetings provided teachers with an opportunity to plan and adapt the KIA curriculum with coaches, and otherwise seek coach support. Nearly all (94%) teachers participated in at least one of the four virtual planning meetings. Participation among the 31 teachers decreased over the course of the year, from a high of 87% in the first meeting to a low of 16% in the fourth meeting

(Table II4). Most teachers (n=11) participated in two planning meetings, followed by nine participating in one meeting, three in six meetings, and three in four. Two teachers never participated in a planning meeting.

Table II4: Virtual planning meeting attendance

Planning meeting	Number of teachers completing	Percentage of total
1	27	87
2	20	65
3	9	29
4	5	16

According to coaches, the teachers needed the additional planning support at the beginning of the year when they were first familiarizing with the KIA curriculum, but their need decreased over time as teachers became more familiar with KIA's scope and sequence.

In-person coaching

The goal of the one-on-one, in-person coaching meetings held during the Summer Institute was to build trust between the coach and teacher prior to conducting virtual coaching. Coaches asked teachers about their learning processes, prior coaching history, teaching strengths, and areas of needed growth. At the end of the Summer Institute, teachers self-assessed their practices with help from their coaches, following a PBL teaching rubric. Based on their self-assessments, teachers set personal development goals intended to guide subsequent coach feedback and inform the focus of future virtual coaching cycles.

During the professional development sessions, coaches also learned about teachers' work plans, answered questions, and gave constructive feedback through small group and one-on-one conversations. Starting in the second Professional Development Session, coaches incorporated the beginning of the Improvement Cycle—creating a Theory of Action—into the in-person meeting.

Informal coaching interactions

In addition to virtual and in-person coaching during the Summer Institute and Professional Development Sessions, coaches encouraged teachers in need of immediate support to contact them at any time via email, text, phone, or Sprocket. Through the course of the year, coaches recorded 43 instances outside of the virtual improvement cycles or planning sessions in which they had informal, substantive interactions with teachers lasting 30 minutes or more. These 44 instances were distributed across 17 teachers: six teachers had one interaction, one teacher had seven, and the other 30 instances were split among 10 teachers.

Building Professional Community

To kick off teachers' KIA experiences, PBLWorks and LER hosted dinners for teachers and district staff during each of the five district-specific Summer Institutes. In the districts where we conducted observations, the morning after the dinner the atmosphere was notably more familiar, with teachers and coaches talking and laughing together. Catered breakfast and lunch also provided time for informal conversations and networking, with teachers and coaches discussing a range of topics, from weekend happenings, to Betsy DeVos' nomination for U.S. Secretary of Education.

The Summer Institute and Professional Development Sessions featured daily community-building activities, such as community "energizers" and "icebreakers." On the first day of the Summer Institute, coaches led an activity in which teachers shared their stories, both professional and personal, and questions about KIA. On the fourth day, the coaches facilitated an interactive activity requiring teachers to assemble into "human bar graphs," prompting jokes and laughs.

The daily schedules of the Summer Institute and Professional Development Sessions also included informal time for community-building activities. During the Summer Institute, teachers exchanged ideas on how to use newly-adopted APES texts with KIA, helped each other navigate Sprocket, strategized about how best to meet the needs of students in their AP courses performing below their grade level, and shared student learning supports, such as a "sentence starter template."

Teachers' collaboration with their peers continued into the school year during Professional Development Sessions and, to less of an extent, via Sprocket and email. For example, teachers helped others who needed extra support getting "up to pace" in the curriculum, loaned each other resources, and shared contacts for local field trips. Coaches also checked in with teachers outside of formal, in-person coaching sessions to offer encouragement. In the district that participated in the KIA pilot study, teachers who started KIA the year earlier joined professional development again, serving as an additional "resource in the room."

Lessons Learned About Implementation of Knowledge in Action Professional Development

This section shares lessons learned—areas of both strength and challenge—about implementing KIA professional development during the 2016-17 school year as part of the KIA Efficacy Study. Substantiating these findings are interviews with PBLWorks staff and coaches, LER staff, and teachers, as well as observations in one district.

Overview of Lessons Learned

The lessons learned include the following:

- KIA professional development required ongoing adaptation to meet teacher and district needs, teacher by teacher and district by district.
- The definition of high-quality PBL was a point of tension among PBLWorks coaches, with ramifications for teachers.
- The balance between too much KIA professional development structure and too little needs continual refinement.
- Teachers' participation in virtual coaching opportunities was lower than anticipated. When they did take part in virtual coaching, teachers preferred to focus on curricular adaptations and lesson planning rather than on developing their instructional practices.
- Finding coaches with the necessary skill set for the position was a challenge—though with effort and training, coaches developed skills.
- Training for coaches was useful, though it could be improved through more personalization to meet coaches' unique needs.

- Coaches valued their ownership and agency, although a few still desired more autonomy.
- Community and collaboration between coaches facilitated success.

We provide further detail in the sections below.

Development of Teachers' PBL Instructional Practice Required Ongoing Adaptation to Meet Teachers' and Students' Needs

A key KIA professional development principle is to develop teachers' PBL instructional practices in addition to preparing them to teach the KIA curriculum. From the coaches' perspective, teachers found the focus on teaching practices (e.g., groupwork facilitation, assessment, scaffolding) helpful and relevant to their immediate needs.

Development of teachers' practices necessitated adaptation of KIA professional development objectives, agendas, and materials to unique teacher needs and district contexts. For example, during the professional development days, teachers often were at different stages in their curriculum due to variations in their pacing. Although LER staff pointed out that the KIA professional development helped keep teachers at roughly similar places in the curriculum, this still required adaptation to meet teachers where they were.

In a few instances, districts' own professional development overlapped with topics covered in KIA professional development, such as scaffolding and Depth of Knowledge, requiring adaptation to avoid too much duplication. There also was variation across districts' testing cultures, such that teachers in a district with a particularly strong testing culture requested more focus on AP examination preparation than did other teachers.

Coaches responded to teachers' developmental needs "in the moment" based on their growing knowledge of individual teachers' practices and their learning progress during KIA professional development. Coaches solicited, documented, and addressed teachers' questions or wonderings (called "need to knows") on a daily basis during the Summer Institute and Professional Development Sessions. For example, during the second session, teachers and coaches shared anecdotes about building a culture of high-quality student work in their classrooms. One coach customized a story to meet the needs of a teacher who was moving through curriculum too quickly. A second coach customized a story to speak to a teacher whose ideas about high-quality student work did not align with PBL best practices. At the end of the session, coaches discussed with one another how they could address specific areas of needed development for each teacher through helping teachers outline pedagogical goals and steps to achieving the goal as part of their Theory of Action.

PBLWorks, as a way to address variation in expertise within and across districts, also harnessed the assets teachers brought to professional development. For example, during the "Collective Wisdom" protocol on Day 3 of the Summer Institute, PBLWorks had teachers share their successful teaching strategies with others in the community.

Despite pacing disparities and differences in teachers' needs, LER and virtually all PBLWorks staff spoke positively of their ability to facilitate professional development in a way that developed teachers'

practice and was responsive to their needs. A coach described attentiveness to teachers' unique needs across districts:

> "I think PBLWorks and Lucas collaborated well to make sure they were very attentive to feedback from the participants—both in general across the different cities, as well as with some specifics to the locations—to make sure the professional development days were trying to meet the needs of the participants."

LER staff and PBLWorks coaches agreed they continually improved upon KIA professional development throughout the year, with coaches better able to meet teachers' developmental needs, by drawing from teacher feedback and evidence of their learning. Nonetheless, PBLWorks leadership shared they were working on a system to better adapt professional development using survey data collected from teachers after each Summer Institute day and Professional Development Session, as well as other informal assessments of teachers' learning. They planned to continue using unified driving questions and an "arc of learning," or trajectory of learning goals, for all teachers, but intended to better modulate learning for individual teachers and districts.

Key takeaways:

- Teachers came to KIA professional development with different levels of prior PBL understanding and experience. They also progressed through the KIA curriculum at varying rates. This variation necessitated customization of KIA professional development objectives, agendas, and activities to meet individual teacher needs.
- Customization required ongoing learning about each teacher's progress and needs, but PBLWorks activities were well-suited to customization.

Differing Perceptions of High-Quality PBL Instruction had Ramifications for Coaches and Teachers

As compared to coaches who helped develop the KIA curriculum and/or previously taught it, coaches who had been PBLWorks National Faculty prior to the KIA Efficacy Study had different perceptions of the definition of high-quality PBL curriculum and instruction. Most coaches with PBLWorks National Faculty experience believed the KIA curriculum was not well-aligned with PBLWorks' Gold Standard PBL Elements. One coach felt the Summer Institute's focus on the PBL Gold Standard Design Elements did not adequately continue into the Professional Development Sessions, which focused on PBL Gold Standard Teaching Practices. For this coach, this lack of focus on PBL elements meant teachers did not have adequate knowledge of PBL to make curricular adaptations. A PBLWorks interviewee shared his/her perception of this challenge:

> "The challenge for us is that PBLWorks has a vision and a stance on what high-quality project-based learning is. As of now, it's defined by what we call our Gold Standard. … Project design elements include sustained inquiry and authenticity and public product and things like that; there are eight of them. We did not have a hand in developing any of the curriculum, so there are some alignment issues with what we believe is quality PBL and with what the curriculum is. … If we were to evaluate the curriculum independent of all the work we're doing, we would have some concerns

about its level of 'PBL-ness,' if you will. That has surfaced around issues of authenticity and, I think, also around public product."

On the other hand, most coaches who had previously taught KIA defended the KIA approach to PBL, arguing PBLWorks doesn't have a "trademark or patent" on what PBL is, and that the KIA designers based the curriculum on learning theory and knowledge of AP standards. As one coach explained:

> "We worked through lots of learning theory and this is what we came up with. So it's not wrong—it's just a different way of doing it, and I'm sure we can make it better. … Let's talk about it, what's missing, what can we adapt. But I was like, 'Have you ever taught AP before?' There's a curriculum. You've got to get kids ready for a test."

Regardless of their perspective on the definition of high-quality PBL, seven of ten coaches, including both PBLWorks National Faculty and former KIA teachers, agreed that the areas of misalignment between KIA and the PBLWorks Gold Standard posed a challenge to teachers. Of particular challenge, PBLWorks' curriculum during the Summer Institute and Professional Development Sessions called for asking teachers to adapt KIA, an unfamiliar curriculum, to better align to PBLWorks Gold Standard PBL Elements, with which they also were not familiar.

However, several coaches believed teachers will be more successful creating appropriate adaptations in their second year:

> "What we noticed is that in your first year, it doesn't look a lot like PBL. It looks like you're just trying to follow the script of the curriculum, and that's just what we've seen for the past two years pretty much across the board. And then at the last two PD sessions, you saw teachers start to think about, 'Well, here's how I'd do it differently next year. Here's how I'd get the community more involved. Here's how I'd get the kids more independence in the project.' And then they started to really think in terms of PBL as opposed to just kind of going through the motions—and that's the real spirit of PBL."

One purpose of the now-underway KIA Maturation Study, following on the promise of the KIA Efficacy Study, is to learn about teachers' KIA practices in their second year and how their evolving proficiency impacts students' AP scores.

Key takeaways:

- Coaches' backgrounds, either as PBLWorks National Faculty as compared to playing a role in developing and/or teaching KIA, contributed to differences in their definitions of high-quality PBL and, accordingly, in their assessment of the extent to which KIA represents high-quality PBL.
- This disagreement was a source of tension between coaches, albeit not a paralyzing one. (Overall, as described below, the coaching community was strong and positive).
- According to coaches, areas of misalignment between KIA and the PBLWorks Gold Standard posed a challenge to teachers—particularly when they needed to adapt the unfamiliar KIA

curriculum to better align to PBLWorks Gold Standard PBL Elements, with which they also were not familiar. Teachers themselves, however, did not articulate this misalignment as a challenge.

The Right Amount of Professional Development "Structure" is a Tricky Balance

LER staff viewed the fundamental KIA professional development structure, including heavy use of the Buck Institute's PBL activities and rubrics, as a necessary foundation coaches could then adapt to the needs of teachers and districts. While LER staff felt teachers were more engaged with PBLWorks' "meatier" PBL activities and rubrics than "lighter" ones, they did not critique the heavy reliance on these pre-developed materials. From LER's perspective, without such a structure, variation between districts in the quality of professional development could have become a considerable challenge. LER staff described the necessity of structure for promoting effective use of time:

> "(The PBLWorks team), they should be able to say for every minute of the day why exactly they're using that minute and how it adds to the learning goals. And if they can't say that, then they need to go back and rethink that use of time. It's just kind of backwards planning for professional development sessions, and they know that. We talk about that all the time; it's good to see that they're onboard with that. That is one of the main areas of improvement moving forward."

Coaches did not universally share LER's belief about the importance of a foundational professional development structure. Eight of 10 coaches said KIA professional development was too structured; as a result, it undermined teachers' sense of professionalism and hindered opportunities for more informal or in-depth conversation about areas in which teachers desired greater support.

One coach described facilitation of such a large number of pre-defined PBLWorks activities and rubrics as, "Let's do this protocol and this protocol and this protocol." Another coach noted the reliance on structured activities and rubrics, which PBLWorks staff originally designed for fairly large audiences, as less effective in districts with smaller teacher groups. Other coaches described the PBLWorks activities and rubrics as "contrived" and "canned."

A few coaches felt the time set aside for teacher planning, which included clearly-defined expectations for how teachers should use their time and the product they should deliver at the end of the planning time, was too structured—again compromising teachers' sense of professionalism and preventing them from devoting time to their individual areas of greatest need.

A few coaches also felt the "Improvement Cycle" structure of virtual coaching was too structured, as most teachers really needed help "just digging into the curriculum and … understanding the scope and sequence. … You know, figuring out what was truly important and critical about each activity for the students." Though this perspective conflicted with the KIA professional development premises of teachers needing to develop their PBL instructional practices as well as grow in familiarity with KIA curriculum and materials, ultimately, PBLWorks and LER staff permitted virtual coaching flexibility. Coaches and teachers fit in conversations not directly addressed by the structured improvement cycle. As a coach explained,

> "For the teacher I was coaching, it would have been nice to have had a little bit less structure in the day during which we could attend to her pressing need—but we did it virtually instead and it worked out fine."

Key takeaways:

- There is a balance between too much structure and not enough. The latter can waste time and result in variations in professional development quality while the former can detract from teachers' sense of professionalism.
- LER and PBLWorks staff tended to believe in the need for more structure, yet coaches preferred less.

Virtual Coaching Structure Could Be Reconsidered

Despite the original intent of virtual coaching to support development of teaching practices through improvement cycles, teachers' participation was quite low. As described above, most teachers did not complete the virtual improvement cycle.

Early in the school year, PBLWorks and LER realized most teachers were not engaging with virtual coaching in the envisioned manner. They communicated to coaches the message that rather than strictly requiring teachers to engage in the improvement cycles, they could develop their coaching focus with each teacher on an individual basis. According to four coaches, this flexibility in the coach-teacher relationship and purpose was critical to its utility for teachers. They felt that the purpose of coaching was to "provide your teachers what they most need, (rather) than you just check the boxes of everything that you said you were going to do." That said, teachers' participation in planning meetings, in which coaches supported curriculum adaptation and lesson planning, was only slightly higher, with most participating in two of four sessions.

Several logistical challenges to virtual coaching may have contributed to low teacher participation. All 10 coaches reported that teachers often did not upload video recordings of their instruction, either because of technological challenges or, they speculated, due to teachers' sensitivity to sharing documentation of their practices. Without the videos, coaches did not have the evidence needed to conduct the improvement cycles as designed. Even when teachers submitted video, the submissions did not always provide coaches with the material needed to help. In some cases, the challenge was poor video quality. More often, however, the challenge stemmed from teachers' video segment selection. On occasion, teachers would only record the students and not provide video of instruction. As a reminder, virtual coaching had not been a standard PBLWorks practice prior to the KIA Efficacy Study, so it is possible these technical challenges can be resolved.

Nine coaches shared that it was challenging to schedule coaching sessions with teachers and that teachers rarely initiated contact. They attributed teachers' lack of participation to busy schedules and feeling they did not need help. Toward the end of the year, according to coaches, teachers were more focused on preparing students for the AP exam than on their teaching practices. Coaches also felt teachers were reluctant to schedule coaching improvement cycles immediately following professional development sessions, which drove PBLWorks and LER's decision to incorporate the pre-observation cycle into Professional Development Sessions.

Relevant to the minority of teachers who engaged in the full improvement cycles, six coaches felt the practice of "kick-starting" the cycles by conducting pre-observation during professional development sessions, initiated by PBLWorks midway through the year, encouraged teachers' participation. One coach explained how the in-person contact changed the nature of virtual coaching:

> "It was just so much more meaningful to have a virtual coaching cycle when we had collaboratively created some goals based on what we were learning in professional development. For me, that was the biggest game-changer and the thing that I enjoyed the most, because then it didn't make the conversations awkward or fake, it really made them fun."

Five coaches felt virtual coaching was more difficult as compared to face-to-face, because relationship-building online was problematic, and they lacked contextual information about the school environment. A coach explained why understanding the school context would have helped to support an APES teacher:

> "It's hard never going to their school and never really seeing what's going on, especially from the environmental science perspective. I would pull up Google Maps and do all kinds of different things just to try to find out what kind of resources they have around them that they could leverage—but that's really hard when you never see them, never go to their buildings."

Virtual coaching lessons learned:

- Challenges to video recording classroom instruction contributed to teachers' limited use of virtual coaching support. Some issues were logistical while others may have related to teacher's discomfort with recording.
- Participation in virtual coaching was lower than anticipated for both improvement cycles focused on instruction and planning meetings focused on curriculum.
- When virtual coaching took place, teachers and coaches addressed curriculum adaptations and lesson planning more frequently than instructional practices.

Coaching Talent is Critical to Success

Critical to the successful implementation of the KIA professional development program, according to five coaches and LER staff, was finding the right coaching talent. KIA coaches tended to have either more extensive experience in PBL coaching or in teaching the KIA curriculum. As could be expected, those with PBL coaching experience leaned toward emphasizing PBL teaching practices while coaches with prior KIA teaching experience tended to focus on helping teachers use the KIA curriculum. The difference was not too extreme, though. One coach, a former KIA teacher, confirmed being more comfortable supporting teachers with curricular issues. However, they also found teachers' requests were more focused on curriculum as opposed to instruction. But another coach—one who was a PBLWorks National Faculty member—said coaches who were former KIA teachers too readily used a "direct coaching approach," offering suggestions based on their own practice, rather than an "inquiry coaching approach" in support of teachers' development of novel solutions through exploration of their practice.

LER and PBLWorks leadership acknowledged the variation in coaching talent resulted in variability in professional development delivery, including the success of certain activities and in teachers' satisfaction with their coaches. Despite the variation, however, a PBLWorks staff person and three other interviewees identified coaches' efforts as critical to KIA professional development success. A leader shared glowing praise of coaches' responsiveness to teachers' needs:

> "When I think of our rock stars, the (coaches) just seem to be doing a great job. They're invested in the success of their teachers, and they hold themselves accountable to the success of their teachers. … They'll shoot a text to a teacher and say, 'Hey, is everything going okay? Let me know if you want to set up a call,' or, 'How did that whatever go that we worked on last week?' That type of communication goes a long way. That's what we're seeing. Personal investment and accountability for the success of teachers is a little bit of that secret sauce that we see good coaches do."

PBLWorks staff said coaches, through effort, could reach a "middle ground," or balance, between PBL coaching experience and KIA curricular expertise. In the case of coaches versed in PBL but no KIA experience, coaches needed to familiarize themselves with the KIA curriculum. For those with KIA experience but little in the way of PBL coaching, PBLWorks staff acknowledged the need to provide more personalized coach preparation and support. According to PBLWorks leadership, it is more difficult to develop PBL coaching skills as compared to gaining familiarity with the KIA curriculum.

Coaching talent lessons learned:

- Coaches had greater expertise in either KIA curriculum or in PBL coaching. Thus, they needed to expend time and effort ramping up their capacity in the area with which they were less familiar.
- Though few coaches had both KIA and PBL background prior to serving as KIA coaches, overall PBLWorks and teacher satisfaction with coaches' performance was high.

Preparation and Support for Coaches Needs Personalization

Three coaches—one with prior KIA teaching experience and two who are PBLWorks National Faculty—reported that PBLWorks-provided preparation and support was helpful. A coach familiar with KIA, though with little prior coaching experience, reflected the boot camp:

> "It was really interesting and helpful to meet Buck National Faculty, who either are full-time coaches or have coached or done PBL one-on-one through Buck. It was really interesting to hear their perspective, and to actually spend some time thinking about what it means to coach and what specific needs adult learners have versus students in my classroom."

Five coaches and LER staff, however, said there was need for improvements in coach preparation and support. In reference to customizing KIA professional development, LER staff noted, "variation is always an opportunity for improvement." Coaches emphasized the need for PBLWorks to better differentiate preparation and support based on coaches' areas of strengths and weaknesses. One coach

described coaches as "pulling on our own strengths," adding, that with more differentiated training, "all coaches could grow."

PBLWorks leadership recognized the need for additional and more personalized coach preparation and support. The program director referred to PBLWorks' performance rubric for coaching and said the organization plans to make better use of the rubric to evaluate and provide feedback to KIA coaches on their performances. PBLWorks' plans for improved training also include targeting "problems of practice" through small-group and one-on-one sessions with coaches that will draw from video recordings of coaches' interactions with teachers and facilitation of professional development sessions.

Key takeaways:

- PBLWorks' three-day KIA boot camp for coaches was most useful for one coach with prior KIA teaching experience and two who were PBLWorks National Faculty coaches.
- Coach preparation and support could have been more personalized, and PBLWorks plans to address this need.

Coaches' Sense of Ownership and Agency was Key to Their Success

A strength of the KIA professional development program, cited by six coaches and LER staff, was an effective fostering of coaches' sense of agency and harnessing their strengths. PBLWorks and LER took steps to develop coaches' sense of ownership and agency after the Summer Institute, when coaches shared they felt they were delivering a script (which is antithetical to PBL). For example, the Buck Institute employed two coaches to collaborate with the PBLWorks program manager on the ongoing design and revision of the professional development materials. In addition, each district's coaching team provided their "second pass" to daily agendas and instructional plans as a way to respond to individual teachers' ongoing feedback—gathered informally and formally through daily satisfaction surveys—and the greater needs of each district. Coaches met as a group each morning, during every lunch break, and at the close of every Summer Institute and professional development session day to make up-to-the-minute adaptations to meet the teachers' needs.

Though six coaches expressed satisfaction with their level of ownership for KIA professional development, four wanted more autonomy. One coach compared the challenge to that of a substitute teacher following the primary teacher's instructional plan. Another felt coaches did not have the necessary freedom to effectively meet the needs of teachers. LER staff also noted that meeting teachers' developmental needs required more flexibility across districts. Two coaches described the prepared outline for coaches as a "script," encumbering their sense of autonomy and professionalism. One of these coaches, however, felt a greater sense of ownership throughout the year as PBLWorks and LER incorporated more input from coaches into professional development plans.

There were challenges to having so many collaborators and multiple stages of revision. For example, following each district's professional development session (20 in total; four in each of five districts), LER and PBLWorks staff modified agendas and materials for the next session. According to the coaches, these constant modifications occasionally resulted in their receipt of materials too late for them to feel adequately prepared or to adapt the materials to the local context.

Key takeaways:

- The majority of coaches felt their level of autonomy was productive and satisfying, though a sizable minority would have liked more independence.
- Constant modifications to materials and plans resulted in late delivery to coaches, which sometimes challenged their ability to adequately prepare and/or adapt materials to local context and teacher needs.

The Coach Community Facilitated Success

In addition to PBLWorks staff's provision of preparation and support, coaches reported learning from one another in an informal manner. Among the six who said the coach community was an important facilitator for their work, they spoke of a positive working environment, being able to "lean on" other coaches, of swapping practices that worked well with their teachers, of colleagues "checking in," and of being able to "reach out" to other coaches. Coaches also described the value of learning from other coaches' practices through observing their colleagues during professional development sessions. As one coach explained, "I get so many tools for my own personal toolbox, just from being around other (coaches)." Another coach described the benefits of having a network of peers:

> "I have some strengths and because we're a team … each coach brings something to the table that is different than the other coaches. So I think, right now we do what our strength is, then let other people kind of help out, and then we learn from other people on the go."

Coaches found their "local" network of coaches—those working in the same district—to be most helpful, though they also learned from other districts' coaches during the summer boot camp. A coach described collaboration with the local and broader network of coaches:

> "I would say (the two coaches who co-facilitate with me) are the two that I'm the closest to, mainly due to working proximity, but in our professional development meetings that (PBLWorks manager) holds ever so often … everybody's so warm and supportive—it's like everyone's greatest cheerleader, I feel like. I think (the KIA coach boot camp) really helped foster that collaboration."

In the district we observed, the three coaches had a history of working and traveling together and seemed quite close, professionally and personally. They also had intense, substantive meetings each day about the agenda and adaptations to meet specific teacher needs.

One coach, however, shared that they would appreciate more formal opportunities to learn from veteran coaches outside of their district:

> "It was helpful to have other people with more experience to go to with questions—and those people were very willing to help, which was great. I think one thing that might have helped would've been being able to watch more experienced coaches' interaction before I began the year with my teachers, just to get more of a feel for what it might look like."

Though coaches rarely mentioned interpersonal and/or professional challenges within the coaching community, two individuals spoke of challenges stemming from different coaching styles and differences in ideas about how best to adapt professional development. Although these two coaches reported that working with some coaches required more "give and take," they still described the coach community as positive.

Key takeaways:

- The KIA coach community was a tightly-knit group, particularly within district teams, though also across districts.
- Coaches developed their coaching practices through observing other coaches.
- Interpersonal and professional challenges within the KIA coaching community were rare.

The Future of Knowledge in Action Professional Development

Coaches and LER staff believe teachers' PBL teaching ability will grow considerably during their second year of teaching with the KIA approach, when they will be more familiar with the curriculum. As part of a plan to support teachers beyond their first year, LER and PBLWorks leadership invited continuing teachers to join new cohorts during the 2017-18 Summer Institute and Professional Development Sessions. Also, LER and PBLWorks are considering how best to facilitate support between teacher cohorts, such as through "train the trainer" approaches.

Continuous Improvement

LER and PBLWorks are committed to continuously improving KIA professional development through collecting information about the experiences of teachers and coaches, and using it to inform adaptations. Staff from both organizations spoke of a need to gather data points beyond satisfaction surveys and observations of sessions, and to better respond to what they learn. Among their adaptations, LER and PBLWorks are narrowing the learning objectives for each Summer Institute day and Professional Development Session, in part so that they can more easily measure the extent to which teachers are meeting the objectives. Similarly, they are collecting more evidence of teacher learning by observing teachers and coaches participating in various activities and conversations, and documenting advice coaches provide to teachers and the products that teachers create. According to LER staff, as the year progressed, teachers' reflections on their progress became "more strategic and measurable and streamlined, as well as aligned with learning goals," and they believe the reflections also will help to inform KIA professional development improvements. LER and PBLWorks staff's goal is to continue using the knowledge gained about teachers' learning to be as responsive as possible to each instructor's needs.

Building Capacity for PBL

Essential to building teachers' PBL capacity, according to LER and PBLWorks leadership, is professional development that is sustained, "job-embedded" builds on teachers' strengths, and is timely and reactive to teachers' needs (see also Baines et. al, 2015). Elaborating on this belief, LER staff shared the following:

> "Professional development is not just something that facilitates learning of the curriculum. The community is not just this extra thing that revolves around the curriculum. They all work together. The curriculum is not the core of this experience. If anything, you could argue that a lot of the professional learning experiences and the conversations they have about project-based learning overall are the lynchpin for the entire experience, with the curriculum providing examples to hang their hat on something, where they feel like there's a concrete example of what this might look like in their classroom, so they can get started and not just end with those conversations. I think the framing of it as a program is really key to this and it's not driven only by curriculum."

Thus, moving forward, LER will continue to refer to KIA as a program with curriculum, professional development, and community—not just as curriculum on its own.

Also critically related to building capacity for PBL, PBLWorks leadership views KIA coaching as essential, yet acknowledges the potential financial challenge inherent to a low coach-teacher ratio:

> "If we removed that component from this whole program we've developed, we would lose the magic. You need a solid support system to enable teachers to be successful in this. If you just take the PD program and stick it in the binder and plop it on the desk at the district office, I think it's going to be challenging to have a highly-impactful implementation of that work—unless it is accompanied with a robust instructional coaching support system behind it."

With an interest in building capacity, LER and PBLWorks staff are creating a KIA "field guide," drawing from KIA professional development activities, rubrics, checklists, and other materials. LER emphasized the point that districts and other providers offering KIA professional development and using the field guide will need dedication to the idea of "continuous improvement" and a commitment to adapting materials based on teachers' interests and needs. LER staff also shared their hope that interest in KIA will grow organically as more schools, teachers, and students gain familiarity with KIA experiences and effects.

Appendix JJ: Professional Development Observations, 2016-17

KIA teachers participated in PBLWorks' KIA professional development program during the 2016-17 school year, as described in this appendix. Of note, PBLWorks has continued to develop and iterate upon its program since then.

In Districts A and E, USC research staff observed KIA Summer Institute and Professional Development Session activities through the 2016-17 school year. As noted in Appendix II, PBLWorks continued to develop and iterate upon their KIA professional development program after the 2016-17 school year. This Appendix material applies to KIA teachers' participation in the program during the 2016-17 school year.

Summer Institute Day 1: "Experiencing"

During the first day of the Summer Institute, teachers "experienced" project-based learning (PBL) firsthand. In the morning, teachers described the skills and dispositions of an "Ideal Graduate" as a way to connect their own learning about PBL instructional fundamentals and Knowledge in Action (KIA) curriculum to their goal of serving their student populations. Coaches from the Buck Institute for Education (BIE) then introduced teachers to their Gold Standard PBL Design Elements through "PBL Slice," a short, immersive project activity. So teachers could experience a KIA lesson from their students' perspective, they assumed the role of students while coaches assumed the teachers' roles. Teachers-as-students worked in teams of 3-4 on an abbreviated KIA project related to their subject area, then presented to the larger group.

The next activity was "Turn and Talk," during which teachers reflected on their PBL Slice project, discussing with a partner the skills and other knowledge the project required, the roadblocks they experienced, and the anticipated needs and misconceptions of students during the project. The objectives of the activity were for teachers to learn through experience: 1) operationalization of the KIA "engagement first" principle; and 2) what project work might be like for students. The immersive activity also provided teachers with an experience to draw upon in subsequent professional development and their future teaching.

After lunch, coaches used PowerPoint slides and a video about an elections project to explain how KIA, in which the entire curriculum is built around projects, differs from "dessert"—projects coming at the end of lessons, which tends to be more typical. After the presentation, teachers again participated in a "Turn and Talk," discussing what made the election project a "main course." Coaches then gave each teacher a cardstock diagram of the Gold Standard PBL Elements, and a reading on PBL, before leading them through the "Building Background Knowledge" activity. This activity helps the learner—here, the teachers—identify what they already know and build from that existing knowledge. Coaches and teachers then discussed how and when to use the Building Background Knowledge activity in their classrooms, and coaches listed teachers' ideas on a "strategies" chart, to which they added through the course of the Summer Institute week. As the week progressed, coaches papered the walls with charts documenting the results of various discussions.

Summer Institute Day 2: "Finding"

The focus of the second day of Summer Institute was "finding," or identifying, PBL elements within the KIA curriculum. Teachers engaged in a "Focused Reading" of the curriculum in which they identified the Gold Standard PBL Design Elements in the first project's tasks and lessons. Using annotated sticky notes, the teachers tagged examples of each Design Element. The teachers then posted these notes on chart paper labeled with the design elements. For example, a teacher posted on the "Reflection" chart, "Task 1, Lesson 7: Students complete a learning log review of the whole unit." With help from coaches, teachers also worked on identifying where skills and concepts "looped," or repeated, across the five curricular units. In the afternoon, coaches led teachers in a closer inspection of the first KIA unit, using the charts to identify where the curriculum and materials might be adapted to provide greater representation of the Design Elements.

Summer Institute Day 3: "Sharing"

On the third day, which focused on "sharing," coaches introduced teachers to BIE Gold Standard PBL Teaching Practices. They used the "Collective Wisdom" activity in which teachers shared their teaching expertise and ideas with their colleagues. In the afternoon, teachers revisited PBL Design Elements through "Bracketology," a competitive game in which participants matched up design elements against one another and discussed which was more important. The ultimate goal was for teachers to realize there is not a single "best" design element, but that they all work in concert. In one district, in response to teachers' expressed interest, coaches modeled a "Structured Academic Controversy" (SAC). During a SAC, a central component of the KIA curriculum, students explore a hot-button issue by first presenting contrasting positions, then working to approach a consensus. Some teachers participated with the coaches in the SAC while the other teachers observed the discussion in "fishbowl" fashion.

Summer Institute Day 4: "Refining"

On the fourth day, focused on "refining," teachers spent approximately three hours preparing to teach the first KIA unit to their students. During the morning, they used the lens of the Gold Standard Design Elements to adapt the first KIA project unit to their particular classroom setting. Then coaches and teachers, in groups, used a "Consultancy" rubric to provide feedback to each teacher's adaptations. First, a teacher shared his or her adaption, followed by colleagues and coaches asking clarifying and probing questions about the adaptation, and then they critiqued the teacher. In the final step, the teacher reflected on the feedback and planned next steps for his or her work. During the final afternoon worktime, teachers refined an adaption they had worked on during the Summer Institute week and uploaded it to the Sprocket online curriculum portal. Teachers also created a list of tasks they wanted to accomplish, then shared it with their groups so group members and coaches could hold them accountable to their plans.

In Table JJ1, we name each Summer Institute activity listed above, summarize the activity's learning objectives, and provide a brief description.

Table JJ1: Summer Institute Activities, organized chronologically as administered

Activity	Learning objectives	Description

Connecting Stories Icebreakers	• Build community • For coaches to learn the questions teachers have about KIA	Teachers share personal stories, teaching stories, and KIA questions in three rounds with the requirement that each story connects to the person who shared prior.
Ideal Graduate	• Connect what teachers are learning about PBL and KIA to their goals for their students.	Coaches write, on a poster, the teachers' collective descriptions of the dispositions and skills an ideal graduate should possess.
PBL Slice	• Experience a project like one teachers will teach to their students • Identify PBL principles	Teachers, acting as students, work in groups on an abbreviated KIA project related to their subject area, then present to the larger group.
Turn and Talk	• Reflect on their learning • Make connections to classroom practices • Share ideas • Brainstorming	Teachers turn to colleagues for focused discussion following an activity.
Building Background Knowledge	• Build background knowledge on Gold Standard PBL Design Elements	In groups, the teachers document, on a poster with three concentric circles, their background knowledge on a topic (inner circle), what they know after watching a video (middle circle), what they learned from reading (outer circle), and evidence of each PBL Design Element found in video or reading (outer circle divided into segments).
Focused Reading	• Identify the Gold Standard PBL Design Elements in the tasks and lessons of the first project	The teachers document, on a sticky note, evidence for each Design Element and where it appears in the unit. The teachers then post these notes on chart paper labeled with the design elements.
Collective Wisdom	• Share best practices from the teachers' "toolbox," using the lens of PBL Gold Standard Teaching Practices.	Teachers share their teaching expertise and ideas with colleagues in multiple "rounds." In the first round, teachers write questions about PBL teaching on posters labeled with PBL Gold Standard Teaching Practices. In the second round, teachers suggest best practices related to the question. A debrief with the whole group follows the two rounds.
Bracketology	• Encourage an informed discussion of PBL Gold Standard Design Elements	Teachers "match up" design elements against one another as part of a competitive game and discuss which is more important.

Consultancy	• Provide coach and colleague feedback on teachers' curricular adaptations	Teachers work with colleagues and coaches in groups to follow six rounds: (1) problem of practice; (2) clarifying questions; (3) reviewing the adaptation; (4) probing questions; (5) discussion; (6) reflection.
Need to Know	• Capture and address teachers' questions	Teachers document their questions throughout the Summer Institute and post them on a chart. Coaches address teachers' questions directly or through an instructional activity.
Meta Moments	• Reflect on learning • Make connections to classroom practices	Teachers record reflections on Sprocket at various points of the Summer Institute.
Calendaring	• Plan the first KIA unit	With the support of their coaches, teachers make a calendar for tasks, lessons, and assessments in the KIA curriculum.
Focused Work Time	• Plan KIA curriculum and instruction	Teachers plan for implementing the first unit and create two Gold Standard PBL adaptations of the KIA curriculum.

Professional Development Session 1

The focus of the first professional development session was building a classroom culture that supports high-quality student work. Coaches introduced teachers to the "Success Analysis," in which teachers reflected on successful implementation of a Unit 1 project, lesson, or activity, and shared their successes with peers. Through the Success Analysis activity, teachers identified several strategies for encouraging high-quality student work for use in the next project. Coaches also introduced teachers to a "Product Analysis" rubric to guide their planning and adapting of the second unit. Coaches and teachers engaged in the short "Chalk Talk" activity several times throughout the day, in which coaches asked, "How can we build a classroom culture that ensures high-quality student work?" and teachers responded and commented on their peers' thoughts. The goals of Chalk Talk were to activate teachers' prior knowledge about building a culture of high-quality student work while encouraging them to track new strategies and reflect on their learning.

During the morning, teachers spent an hour working independently to build upon what they had learned through the Success Analysis activity, as well as the Product Analysis rubric, to plan for the second unit. In the afternoon, teachers created "Y charts" for documenting what they imagined the last day of the second unit project would "look like," "sound like," and "feel like," with a focus on high-quality student work. Teachers shared goals such as "use of academic language," "productive discourse," "celebration and acknowledgment of hard work," and "application of curricular content." Teachers then spent another 30 minutes continuing the planning they had begun in the morning. At the end of the day, during the "Charrette" activity, teachers presented their planning either to a partner or several group members (depending on time) and asked for specific feedback through the use of a "framing question" (e.g., "What can I make better about…?" or "How can I improve…?").

Professional Development Session 2

The second professional development session continued to focus on building classroom culture so as to support high-quality student work. The day began with a "Block Party," during which coaches provided teachers with quotes from the short article *Beautiful Work,* by Ron Berger. Each teacher selected a quote with personal meaning to them, shared with a partner the quote and what it meant to them, exchanged quotes, and continued this exchange with two more partners. Teachers later read the whole article and participated in a "Say Something" discussion about the author's perspective on high-quality student work. Teachers highlighted meaningful passages for sharing with the group. After each individual shared, the group engaged in a discussion of the article.

The teachers next used the rubric "Looking at Student Work" to guide examination of examples of their own and other teachers' student work, focusing on students' areas of strength and weaknesses. Reflecting with colleagues, teachers identified concrete next steps for improving their teaching practices. With guidance from their coaches, each teacher drafted a personal "Theory of Action" that outlined concrete tasks intended to improve teaching practices and encourage students to produce high-quality work. Each teacher's theory of action guided their Improvement Cycles, as described below in this chapter's Coaching section. Teachers also used their Theory of Action to guide focused worktime, choosing from a menu of tasks for helping them improve the quality of student work in their classroom (e.g., lesson planning, sentence stems, graphic organizers). Then, during a "Turn and Talk" activity, teachers provided feedback for each other on the products developed during worktime.

In the afternoon, coaches guided teachers' use of the "Collective Wisdom" rubric to share best practices organized within the following domains: critique and revision, persistence, rigorous thinking, pride in doing high-quality work, and peer accountability. Teachers wrote strategies on posters labeled with each of the domains, visited each poster to read what other teachers shared, and chose one practice they would implement in their classroom.

Teachers then shifted to look ahead to the third project unit. Following a "Project Evaluation" rubric, the teachers reviewed the unit's overview documents, labeling representation of the Design Elements as well as key knowledge, understanding and skills developed, assessment opportunities, and connections to the driving question. They discussed their findings with colleagues, sharing the project's strengths and weaknesses, including ideas for enhancing the project. Teachers then identified planning next steps and, during an afternoon hour of focused worktime, began preparing for the next project. During a closing "Meta-Moment," teachers reflected on their commitments to improving the culture of high-quality student work in their classrooms and how their coach could support their efforts.

Professional Development Session 3

The third session focused on how to promote high-quality student work by "scaffolding," or providing specific learning supports. During the "Would You Rather…?" icebreaker activity, teachers thought about their own learning needs and preferences, then considered how their teaching approaches suited the learning styles of students in their classrooms. Teachers then engaged in a "Socratic Seminar," in which teachers and coaches sat in a circle and discussed a blog post by Sarah Field, *Scaffolding content and process in PBL*. Following the discussion, teachers generated "clarifying," "conceptual," and

"provocative" questions for the group discussion—following a strategy they then could use to help their students organize questions. Coaches closed the seminar by debriefing major points and describing implications for teachers' classroom practices.

In the next activity, coaches asked teachers to identify a "Student Dilemma" for an anonymous student who would benefit from specific learning supports (e.g., organizing question types into categories). The teachers then created a learner profile, called SING (Strengths, Interests, Needs, and Goals), for the student. "Turn and Talk" time provided teachers with a structured way to reflect on their SINGs with colleagues and discuss ways they might personalize learning supports based on each student's profile. Teachers then used a 30-minute morning worktime to edit their Theory of Action and produce a learning support strategy to help students in the upcoming third unit. Coaches next led groups' use of the "Consultancy" rubric to provide teachers with feedback on their work. In a second, 20-minute morning worktime, teachers polished their Theory of Action based on their groups' feedback. Before lunch, coaches led teachers in a brief discussion of when and how to remove learning supports.

During the afternoon's "Thinking Hats" activity, teachers considered the upcoming third project from various lenses, or "hats" (optimism, creativity, possible pitfalls, student experiences and feelings, creativity, and next steps). Coaches asked teachers to wear each hat as they read the next unit, taking notes as they read. Teachers then shared their thoughts related to each hat, after which they put on the "white hat" to brainstorm about adaptations necessary to meet their classroom context and student needs, and next steps for planning and practice. In the last step of the Thinking Hats activity, teachers engaged in a longer, open discussion of their ideas.

During a one-hour afternoon work session, teachers either finalized their student learning supports or worked on making the adaptations brainstormed during the Thinking Hats activity. Coaches led teachers in the last activity of the day, the "*I like, I wonder, I have* Gallery Walk," during which each teacher created a poster describing what they worked on in the day and how it was helpful. The other teachers then wrote out on sticky notes the phrases "I like…," "I wonder…," and "I have…," and applied the notes to the posters. To close the activity, teachers shared "shoutouts" to recognize others' accomplishments, and "*A-ha!*s," ideas they learned from others. The day closed with coaches leading teachers to reflect on daily activities they might use in their classroom.

Professional Development Session 4

The fourth and final professional development session—which took place 4-6 weeks before the AP examinations—focused on how to use inquiry-based instructional and assessment approaches to support student learning. Coaches created four posters, each featuring year-long professional development learning objectives, respectively labeled "Gold Standard PBL Design Elements," "Gold Standard PBL Teaching Practices," "KIA curriculum compared to your previous APES/APGOV course," and "I used to think but now I…" Coaches then led teachers in a "carousel" reflection of their learning throughout the year, in which teachers circled the room and wrote what they had learned over the course of the year about each objective and what they tried, or still hoped to try, in the classroom.

Coaches next facilitated a "Flip-It" activity. In the first part of this activity, teachers voiced their fears about the coming weeks as they prepared their students for the AP exam, then "flipped" their fears into questions. In the second part of the activity, teachers discussed actionable plans to address the "flipped" questions (and underlying fears).

Between the first and second stage of "Flip-It," coaches led teachers in an "In2Out" activity. The activity centered around three texts: a blog about what it means to be an inquiry teacher (Murdoch, 2018), a diagram of the step-by-step inquiry process (Barseghian, 2013), and a chapter summary about "creating a culture of questioning." Teachers first internally responded to the prompts, "When was a time you were engaged in the inquiry process as a learner? What was that experience like for you?" Then they responded to colleagues in groups of two or three to the prompts, "Describe a time in class you taught like an inquiry teacher. What did you notice about your students?" Teachers also discussed, as a group, their responses to the prompt, "How can we have our students more involved in the inquiry process to encourage student ownership of learning?"

During a 45-minute morning worktime, teachers planned an upcoming inquiry lesson for addressing AP exam preparation needs without a "drill-and-kill" approach. Coaches met with teachers one-on-one and provided feedback on their planning.

During the afternoon, in response to teachers' requests for AP exam support, the fourth professional development session focused on how teachers can assess students' preparedness for the exam using Webb's Depth of Knowledge (DOK) framework. The DOK framework, modeled after Bloom's Taxonomy, helps teachers categorize tasks based on the cognitive processes required to complete the task. Coaches led teachers' participation in the group "Graffiti" activity, during which they brainstormed about assessments falling under each categorical level of the DOK and documented them on chart paper. Each group then shared their top assessment ideas. In a "Turn and Talk" activity, teachers reflected with a colleague on why leveled assessment, or DOK, is important in a PBL unit.

Coaches then transitioned teachers to a "Free Response Question (FRQ) Breakdown" activity. Teachers categorized sample FRQ questions (from FRQs released over the past five years) into the appropriate levels of DOK, then reflected on the levels of DOK represented in the questions. Teachers next examined the fifth project's Understanding by Design framework, classified the tasks into DOK categories, and reflected on how the project represented the DOK levels. During the one-hour worktime in the afternoon, teachers planned for the project, making adaptations as needed so tasks required higher levels of DOK. As a final activity, teachers returned to the Flip-It charts and filled in ideas from the afternoon or their own practices that would be helpful for the group.

Each of the professional development session activity names, objectives, and descriptions is summarized in Table JJ2.

Table JJ2: Professional Development Session activities, organized chronologically

Activity	Learning Objectives	Description
Success Analysis	• Share successes related to project implementation as a way to	Teachers reflect on a successful project, lesson, or activity, and share with peers their success. The presenter first describes the success then answers

	understand the conditions promoting success	questions from peers, followed by the presenter and group reflecting on the success.
Product Analysis	• Plan and adapt the upcoming project through discussion about conditions that allow for high-quality student work	In groups, teachers examine examples of student work provided by coaches. They identify the content and skills evident in the product and evaluate to what extent the example represents high-quality student work. They follow by discussing their analysis with colleagues.
Y charts	• Reflect on a classroom culture that promotes high-quality student work	Teachers imagine and record on a "Y chart" what a project will "look like," "sound like," and "feel like" on the last day of the unit, with a focus on high-quality student work.
Charrette	• Solicit feedback from coaches and colleagues on teachers' work in progress	Teachers present their curricular adaptations to colleagues and coaches, then ask for feedback on specific areas of needs through the use of a "framing question" (e.g., "What can I make better about…?" or "How can I improve…?").
Chalk Talk	• Build knowledge about the features of high-quality work	Teachers write a response to the prompt, "How can we build a classroom culture that ensures high quality student work?" and comment on others' thoughts.
Block Party	• Preview a text by making personal connections to quotes *prior* to reading	Teachers read selected quotes from *Beautiful Work,* by Ron Berger, choose one with personal meaning to them, then share the quote and why they selected it.
Say Something	• Generate ideas about classroom structures, strategies, and activities that support a culture of high-quality work	Teachers read the article *Beautiful Work* and discuss the author's perspective on high-quality student work. The teachers highlight meaningful passages, then speak about them. After everyone shares, the group engages in a discussion of the article.
Looking at Student Work	• Identify strengths and weaknesses in student work samples from a project or assignment in order to gain insights about what students know, understand, and can do, and where they need support • Develop concrete next steps to support students' learning	Teachers use the "Looking at Student Work" rubric to guide their examination of examples of their own and other teachers' student work. Their focus is on students' areas of strength and weaknesses. Teachers, with the support of colleagues and coaches, then identify concrete next steps for improving their teaching practices.
Collective Wisdom	• Leverage collective wisdom about the structures, practices, strategies, and tools that support building a	Teachers write strategies on posters labeled with the high-quality student work domains (critique and revision, persistence, rigorous thinking, pride in

	culture of high-quality student work	doing high-quality work, and peer accountability), visit each poster to see what other teachers shared, and choose one practice to implement in their classroom.
Project Evaluation	• Engage in collaborative evaluation of KIA projects • Discuss potential curricular adaptations with colleagues • Calibrate understanding of Gold Standard Design Elements in KIA projects	Using the Project Evaluation rubric, teachers individually assess the project's representation of Gold Standard PBL Design Elements. Teachers then discuss with colleagues their findings, share the project's strengths and weaknesses, and propose ideas for enhancing the project. Teachers finish by planning, based on the evaluation, concrete next steps.
"Would You Rather…?" Icebreaker	• Reflect on differentiation and student support strategies as part of KIA instruction	As part of this icebreaker exercise, teachers consider their own learning needs and preferences, then reflect on the learning styles of students in their classrooms.
Socratic Seminar	• Learn about methods of supporting student learning in PBL	Teachers and coaches discuss a blog post by Sarah Field, *Scaffolding Content and Process in PBL*. Teachers develop "clarifying," "conceptual," and "provocative" questions for the group discussion, then discuss how to translate their learning about support strategies into their teaching practices.
Student Dilemma	• Focus on specific student needs	Teachers identify one student who would benefit from support strategies, write about the student and his or her needs, and share their "student dilemma" with colleagues and coaches.
SING	• Develop a learner profile of the student identified through the student dilemma activity so to better understand how tailor instruction	Teachers develop a learner profile (Strengths, Interests, Needs, and Goals), discuss with colleagues and coaches what they realized about the student through engaging in the activity, and brainstorm next steps to meet the student's needs.
Thinking Hats	• Evaluate the next project through a variety of lenses, or "hats," as to identify where to adapt the project	Teachers read the project Understanding by Design framework in multiple rounds trying on different "hats" (e.g., optimism, possible pitfalls, student experience) as a way read the project from various viewpoints. After taking notes, they discuss their critiques and propose possible adaptations.
I like, I wonder, I have Gallery Walk	• Celebrate and receive feedback on work accomplished	Teachers create a poster describing what they worked on that day and how it was helpful. Then they view each other's posters and record their thoughts (i.e., "I like…," "I wonder…," and "I

		have…") on sticky notes that they apply to the poster.
Carousel	• Reflect on PBL learning to date (throughout the Summer Institute and professional development sessions)	Teachers circle the room and write on posters labeled with various objectives (e.g., Design Elements, teaching practices, KIA curriculum) what they learned over the course of the year and what they tried or still hope to try in the classroom.
Flip-It	• Turn teachers' fears into actionable hopes	In the first part of the activity, teachers write on sticky notes about fears they have for the coming weeks as they prepare their students for the AP exam, then they "flip" their fears into questions. In the second part of the activity, teachers return to their "flipped" questions and discuss actionable plans to address the questions and underlying fears.
In2Out	• Build background on inquiry teaching and learning	Three texts are used for this activity: a blog about what it means to be an inquiry teacher (Murdoch, 2018), a diagram of the step-by-step inquiry process (Barseghian, 2013), and a chapter summary about "creating a culture of learning." Teachers respond to multiple prompts related to inquiry teaching and learning as an individual (writing reflective notes) with small groups and as part of a class discussion.
Webb's Depth of Knowledge (DOK)	• Learn about assessments requiring multiple "Depths of Knowledge"	Coaches introduce teachers to and assess their prior exposure to Webb's DOK.
Graffiti Walk	• Familiarize teachers with Webb's DOK as part of a collaborative activity • Generate a list of assessments teachers might use for lesson planning	Coaches lead teachers in a group "Graffiti" activity where teachers brainstorm assessments falling under each categorical level of the DOK. Teachers document the assessments on chart paper, rotating around the room to a chart labeled with each level of DOK. The groups then share their top assessment ideas.
FRQ Breakdown	• Backwards-plan for the AP exam • Adapt the KIA curriculum for better connecting to the levels of DOK in the AP exam	Teachers categorize each released FRQ question (from the past five years) into the appropriate levels of DOK and reflect on what levels of DOK are present in the questions. They then examine the Understanding by Design for the fifth KIA project, classify all the tasks into DOK categories, and reflect on representation of DOK levels in the project.

Focused Work Time	• Plan and adapt the KIA curriculum	Teachers use focused work time to plan and adapt the curriculum. Coaches ask teachers to produce a variety of "deliverables" including the following: PD 1: Teaching strategy; adaptation of a task/product PD 2: "Looking at Student Work" Theory of Action; curricular adaptation PD 3: Learner profile; scaffold or curricular adaptation that would benefit a student or the class PD 4: Inquiry lesson; questions and assessment for inquiry lesson

Appendix References

Altonji, J. G. & Mansfield, R. K. (2018). Estimating Group Effects Using Averages of Observables to Control for Sorting on Unobservables: School and Neighborhood Effects. American Economic Review 108 (10) 2902-46.

Baines, A., De Barger, A. H., De Vivo, K., Warner, N. Brinkman, J., Santos, S. (2015). What is Rigorous Project-Based Learning? Marin County, CA: Lucas Education Research.

Bandura, A. (1977): Self-Efficacy: Toward a Unifying Theory of Behavioral Change. Psychological Review, 84(2), 191.

Becker, S. and Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal*, *2*(4), 358-377.

Brand, J. P. L., S. Van Buuren, C. G. M. Groothuis-Oudshoorn, and E. S. Gelsema. 2003. "A Toolkit in SAS for the Evaluation of Multiple Imputation Methods." *Statistica Neerlandica* 57 (1):36–45.

Bransford, J., Vye, N., Stevens, R. Kuhl, P., Schwartz, D., Bell, P., … Sabelli, N. (2006). Learning Theories and Education: Toward a Decade of Synergy. In P. Alexander & P. Winne (Eds). Handbook of Educational Psychology (2nd Edition), Mahwah, NJ: Erlbaum.

Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. The Journal of the Learning Sciences, 2(2), 141-178.

Carpenter, J.R. & Kenward, M.G. (2013). Multiple Imputation and its Application. John Wiley & Sons, Ltd.

College Board. (2018). AP Archived Data. Retrieved from

https://research.collegeboard.org/programs/ap/data/archived.

Cook, T. D., Shadish, W. R. and Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of policy analysis and management*, *27*(4), 724-750.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and Passion for Long-Term Goals. Journal of Personality and Social Psychology, 92(6), 1087

Dweck, C. S. (2000). Self-Theories: Their Role in Motivation, Personality, and Development. Philadelphia, PA: Psychology Press.

Ferguson, C. J. (2009). A effect size primer: A guide for clinicians and researchers. Professional Psychology: Research and Practice 40 (5) 532-538

Flanagan, C. A., Syvertsen, A. K., & Stout, M. D. (2007). Civic Measurement Models: Tapping Adolescents' Civic Engagement. CIRCLE Working Paper 55. Center for Information and Research on Civic Learning and Engagement (CIRCLE).

Fogleman, J., McNeill, K. L., & Krajcik, J. (2011). Examining the Effect of Teachers' Adaptations of a Middle School Science Inquiry-Oriented Curriculum Unit on Student Learning. Journal of Research in Science Teaching, 48(2), 149-169.

Galston, W. A. (2001). Political Knowledge, Political Engagement, and Civic Education. Annual Review of Political Science, Vol. 4: 217-234.

Gould, J., Jamieson, K. H., Levine, P., McConnell, T., & Smith, D. B. (2011). Guardian of Democracy: The Civic Mission of Schools. Philadelphia, PA: Lenore Annenberg Institute for Civics of the Annenberg Public Policy Center and the Campaign for the Civic Mission of Schools.

Gu, X. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal of Computational and Graphical Statistics 2*, 405–420.

Iida, M., Shrout, P. E., Laurenceau, J.-P., & Bolger, N. (2012). Using diary methods in psychological research. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbooks in psychology®. APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics* (p. 277–305). American Psychological Association.

Kolluri, S. (2018). Advanced Placement: The Dual Challenge of Access and Effectiveness. *Review of Educational Research, 88* (5).

McDonald, R. P. (1999). Test Theory: A Unified Treatment. Oxford, UK: Lawrence Erlbaum Associates.

Northouse, P. G. (2013). Leadership: Theory and Practice (6th ed.). Thousand Oak CAs: SAGE, 6.

Oswald, F. (2004), Developing a Biodata Measure and Situational Judgment Inventory as Predictors of College Student Performance. Journal of Applied Psychology, 89, 187-207.

Parker, W., Lo, J., Yeo, A. J., Valencia, S., Nguyen, D., Abbott, R., … Vye, M. (2013). Beyond Breadth-Speed-Test: Toward Deeper Knowing and Engagement in an Advanced Placement Course. American Educational Research Journal, 50(6), 1424-1459.

Parker, W., Mosborg, S., Bransford, J., Vye, N., Wilkerson, J., & Abbott, R. (2011). Rethinking Advanced High School Coursework: Tackling the Depth/Breadth Tension in the AP U.S. Government and Politics course. Journal of Curriculum Studies, 43(4), 533-559.

Polman, J. (2015). What's Authentic, More or Less? Retrieved from http://composeourworld.org/blog/2015/10/17/whats-authentic-more-or-less

Reardon, S.F., Kalogrides, D., & Ho, A. (2017). Linking U.S. School District Test Score Distributions to a Common Scale (CEPA Working Paper No.16-09). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-09.

Roschelle J., Teasley S.D. (1995) The Construction of Shared Knowledge in Collaborative Problem Solving. In: O'Malley C. (eds) Computer Supported Collaborative Learning. NATO ASI Series (Series F: Computer and Systems Sciences), vol. 128. Springer, Berlin, Heidelberg, 70.

Rosenbaum, P., and Rubin, D. (1985): Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score, *The American Statistician, 39*, 33–38.

Saavedra, A. (2012). From Dry to Dynamic Civic Education Curricula. In D. Campbell, F. Hess, & M. Levinson (Eds.), Making Civics Count: Citizenship Education for a New Generation. Cambridge, MA: Harvard Education Press.

Sadler, P. M., Sonnert, G., Tai, R. H., & Klopfenstein, K. (2010). AP: a Critical Examination of the Advanced Placement Program. Cambridge, MA: Harvard Education Press.

Saldaña, J. (2016). The coding manual for qualitative researchers. Thousand Oaks, CA: Sage.

Schwartz, D., & Bransford, J. (1998). A Time for Telling. Cognition and Instruction. 16(4), 475-522.

Thissen, D. (2007). Linking Assessments Based on Aggregate Reporting: Background and Issues. In N.J. Dorans, M. Pommerich, & P.W. Holland (Eds.) Linking and aligning scores and scales(Pp. 287-312). New York, NY: Springer.

Tyack, D., Cuban, L. (1995). Tinkering Toward Utopia: A Century of Public School Reform. Cambridge, MA: Harvard University Press.

Valant, J., & Newark, D. A. (2017). My Kids, Your Kids, Our Kids: What Parents and the Public Want from Schools. Teachers College Record, 119(12), n12.

Wang, R., Lagakos, S., Ware, J., Hunter, D., & Drazen, J. (2007). Statistics in Medicine: Reporting of Subgroup Analyses in Clinical Trials. New England Journal of Medicine. 357:2189-2194.

Westheimer, J. & Kahne, J. (2004). What Kind of Citizen? The Politics of Educating for Democracy. American Educational Research Journal. Vol 41(2), Sum 2004, pp. 237-269.

Wong, V., Valentine, J.C., & Miller-Bains, K. (2017) Empirical Performance of Covariates in Education Observational Studies, *Journal of Research on Educational Effectiveness*. 207-236.