When low-SES students perform better-than-expected on a standardized test: The role of teacher professional development

Christian Fischer[a], Barry Fishman[a], Abigail Jurist Levy[b], Arthur Eisenkraft[c], Chris Dede[d], Frances Lawrenz[e], Janna Fuccillo Kook[b], Kim Frumin[d], Ayana McCoy[c]

[a] *University of Michigan*
[b] *Education Development Center, Inc.*
[c] *University of Massachusetts at Boston*
[d] *Harvard University*
[e] *University of Minnesota*

For more information about this paper or project, contact:
Arthur Eisenkraft, arthur.eisenkraft@umb.edu

Acknowledgements

This work is supported by the National Science Foundation through the DRK-12 program, award number 1221861. The authors also thank the following people for their contributions to this work: Ted Gardella and Amy Wheelock of the College Board, Allison Scheff of the Massachusetts Board of Higher Education, and the thousands of AP teachers who helped shape and participated in this project. The views contained in this paper are those of the authors, and not their institutions, the College Board, or the National Science Foundation.

When low-SES students perform better-than-expected on a standardized test: The role of teacher professional development

Abstract: This paper describes a study using quasi-experimental design to examine teachers' preparations in low-income schools for a revised version of the AP Biology and AP Chemistry examinations, and explores variables associated with student scores on the AP science examinations that are better than would be predicted based on their PSAT scores. Considering the frequently-measured achievement gap on high-stakes examinations, identifying "what works" to raise student performance of at-risk students is an urgent area for research. The analyses indicate that (a) districts per-student funding allocations, (b) teachers' knowledge and experience, and (c) teachers' participation in professional development activities with a responsive agenda and effective support for teaching the redesigned AP science course are significantly associated with higher students' average performance on the AP science exams than would be predicted.

Keywords: Science education, high-stakes testing, professional development, school context

1 Introduction and problem statement

Following Nelson Mandela's wisdom that "education is the most powerful weapon which you can use to change the world," we share responsibility to offer equitable educational opportunities to all students, not only promoting individual success but also fostering sustainable societal development. Focusing on how to enhance learning and achievement for students who are economically disadvantaged is of special importance in the mission of striving for educational equity. Whereas some indications exist that achievement gaps due to ethnicity in the United States have been narrowed from the 1950s to now, the income achievement gap has been widened substantially. Integrating data from twelve nationally representative studies, Reardon (2011, 2013) found that whereas the African American - White achievement gap has decreased from the mid-1950s to the turn of the century of about .60 standard deviations whereas the income achievement gap for students in the top and bottom 10th percentile has increased from the mid-1940s to the turn of the century of about .50 standard deviations. In 2000, the income achievement gap exceeds the African American – White achievement gap by about .50 standard deviations (e.g., Reardon, 2011, 2013). The influence of socioeconomic status on achievement is also documented in large-scale international comparative studies. For instance, the 2012 *Program for International Student Assessment (PISA)* study indicates that 15% of the variation of U.S. student performance is contributed by students' socioeconomic background (OECD, 2013a, 2013b). The income achievement gap is very persistent from elementary school on through the end of and secondary education as demonstrated on various nationally representative samples (e.g., Coley, 2002; Duncan & Magnuson, 2011; Reardon, 2011, 2013).

Educational attainment is not only highly influential for impacting occupational opportunities, unemployment rates, life-time earnings, health benefits, and tax payments, among others (e.g., Autor, Katz, & Kearney, 2008; Baum, Ma, & Payea, 2013; Murnane, Willett, & Levy, 1995) but also in impacting each student's individual life trajectory. Therefore, it is important to evaluate how to best support teachers in economically disadvantaged schools. High-stakes examination might serve an important role for better preparing disadvantaged students for success because they might influence teachers' classroom instruction. At the high school level, College Board's Advanced Placement (AP) program in the sciences and other subject areas is seen as an opportunity for students to engage in rigorous college-level learning experiences. Research indicates that participation in AP courses and succeeding on AP examinations is associated with students' academic success in higher education, for instance leading to higher enrollment rates in four-year colleges (e.g., Chajewski, Mattern, & Shaw, 2011), higher general college graduation rates (e.g., Dougherty, Mellor, & Jian, 2006; Mattern, Marini, & Shaw, 2013), and higher college grade point averages (e.g., Hargrove, Godin, & Dodd, 2008; Patterson, Kobrin, & Packman, 2011; T. P. Scott, Tolson, & Lee, 2010).

Given the importance of narrowing the income achievement gap, it is important to identify "what works" to support teachers and students in economically disadvantaged schools. Therefore, this study explores the characteristics of the AP science teacher population and their school context in low-SES schools and examines the relationships of school, teacher, and teaching characteristics on students' AP science performance, controlling for student characteristics.

2 Theoretical framework

2.1 The AP program

The College Boards' AP examinations and their corresponding courses provide rigorous, college-level curricula for high school students in a broad variety of subjects. The summative nationwide high-stakes assessments issue grades on a 1-5 scale. The exams are scored using predetermined criterion-based rubrics. Students who receive a 3 or higher on an AP examination may be able to count their AP grade towards their college degree completion, depending on the policies of their institution of higher education.

Relations of students' AP participation on success in higher education. Prior research on the relationships of student participation in the AP program on students' academic success in college yield meaningful insights for higher education regarding college enrollment, successful college graduation, and college grade point average (GPA).

Demonstrating the importance of the AP program for college admission (e.g., Chajewski, Mattern, & Shaw, 2011; Geiser & Santelices, 2006; Schneider, 2009), Chajewski et al. (2011) conducted a quasi-experimental study with a nationwide sample of over 1.5 million students analyzing the role of participation in the AP program for enrollment in a 4-year college indicating that taking at least one AP examination had a substantially higher association with enrollment in a four-year college than taking no AP examinations. Additionally, student participation in AP classes and examinations is associated with a higher probability of successfully graduating college compared to students not participating in the AP program, as indicated through a study from Dougherty et al. (2006) following above 67,000 students in Texas from 8th grade to their potential college graduation and the study of Mattern et al. (2013) analyzing two national datasets with a combined total of above 790,000 students. Furthermore, students' participation in AP classes and examinations is associated with increased college GPAs, especially for the first year of college, compared to students not participating in the AP program (e.g., Hargrove, Godin, & Dodd, 2008; Scott, Tolson, & Lee, 2010). The analysis of Patterson, Kobrin, and Packman (2011) of a national sample of above 195,000 college students indicate that students' AP performance is positively associated with students' grades in corresponding discipline-specific first year college courses. Shaw, Marini, and Mattern (2013) analysis of influences of the AP program on students' first year college GPA using a nationwide sample of above 100,000 students indicated that the average AP score of all AP examinations for an individual student is positively associated with students' first year college GPA.

Given overall tendencies of AP programs to potentially positively influencing students' academic success in higher education, it is important to examine how teachers can make better use of the opportunities for success the AP program might have to offer for a vulnerable student population in the U.S., students who are economically disadvantaged.

Redesign of the AP science curriculum and examination. Focusing on the sciences, recommendations for changes in the AP program emerged from the National Research Council's (NRC) *Committee on Programs for Advanced Study of Mathematics and Science in American High Schools* (National Research Council, 2002) stating that

> [t]he primary goal of advanced study in any discipline should be for students to achieve a deep conceptual understanding of the discipline's content and unifying concepts. Well-designed programs help students develop skills of inquiry, analysis, and problem solving so that they become superior learners. (pp. 197-198)

In addition to this and other federal policy documents, research findings in educational research, the learning sciences, and other related fields promoted a shift away from science learning through algorithmic-centered instruction and rote memorization. Responding to these recommendations, the College Board redesigned their AP science curricula, increasing the emphasis on scientific practices, critical thinking, inquiry, and reasoning in order to deepen students' understanding of relevant science concepts (e.g., Magrogan, 2014; Pellegrino, 2013; Yaron, 2014). The redesigned AP Biology examination was first administered in May 2013, the redesigned AP Chemistry examination in May 2014, and the redesigned AP Physics examinations will be administered in 2015. Many of these changes are in line with nationwide science standards described in the *Framework for K-12 Science Education* (National Research Council, 2012) and the *Next Generation Science Standards* (NGSS) (NGSS Lead States, 2013).

The redesigned AP science curricula is guided by *big ideas* of the subject area that include several *enduring understandings* supported by *essential knowledge*, which students should learn through engaging in *scientific practices* described in *learning objectives* (The College Board, 2012, 2014a, 2014b). Every redesigned AP science curriculum follows the same structure culminating in equal science practices across disciplines.

The redesign of the AP science curricula has major implications for the redesign of the summative AP science examinations. Given the importance of assessments for shaping the manifestation of sustainable change in educational systems, the specific design of the AP science examinations might impact deeply students' preparation and learning experiences (e.g., Pellegrino, 2013). A major design change of the assessment is the reduction of question items focusing on factual knowledge or purely algorithmic procedures in favor of an increase of items focusing on deeper conceptual understanding and higher-order cognitive skills (e.g., Domyancich, 2014; Magrogan, 2014; Price & Kugel, 2014). This is accomplished for the AP science examination applying three different strategies; (a) transforming existing lower-order cognitive items to assess students' deep conceptual understanding, (b) redistributing the number of items based on content, and (c) changing the item format ratio of multiple choice to free- and open response items based on content (e.g., Domyancich, 2014; Magrogan, 2014).

2.2 Conceptual framework

The conceptual framework of this study modifies Opfer and Pedders' (2011) dynamic model of teacher learning and change describing the complex interaction between school context, the individual teacher including their teaching preparation and enactment, and teachers' PD participation influencing student learning and achievement, as illustrated in figure 3-1.

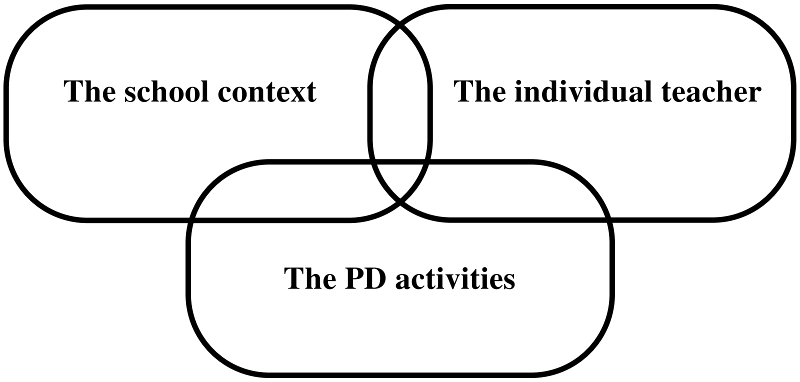

Figure 3-1: Dynamic model of teacher learning and change, after Opfer and Pedder (2011).

This study investigates differences in school context, individual teacher characteristics, and teachers' PD participation patterns across the AP science teacher population in economically disadvantaged schools. Furthermore, the associations with student performance on the AP science examinations are explored in order to identify contributing factors for successfully supporting at-risk students to perform well on the high-stakes assessments.

The school context. Despite the difficulties major curricular revision bring for all educational stakeholders, teachers and students in low-SES schools especially are facing additional challenges. The school context shapes both teacher preparing their students for success on standardized high-stakes assessments and how students interact with the assessments.

Access to resources. Low-SES schools might face challenges to provide their students with sufficient resources to succeed on high-stakes assessments. School funding inequalities are influential factors which hinder the narrowing of the income achievement gap. Synthesizing data from several nation-wide studies Biddle and Berliners' (2003) analysis indicates that the total per-student expenditures per district are substantially lower in districts with a higher percentage of students living in poverty. For instance, districts with total per student expenditures of less than $4,000 have an average percentage of students living in poverty of 22.6% compared to 7.8 % for districts with $12,000 to $12,999 total per student expenditures. Also, district funding is not necessarily equally distributed within the school district, raising concerns about the additional severity of underfunding for low performing schools (P. T. Hill, Guin, & Celio, 2003). Having lower school funding does not only lead to poorly equipped classrooms and science laboratories, higher student-teacher ratios, and other effects, but it has also implications for teacher recruitment processes (e.g., Biddle & Berliner, 2003; Elliott, 1998; P. T. Hill et al., 2003). It is challenging for low-SES schools to recruit and retain highly qualified and experienced teachers (Biddle & Berliner, 2003; Elliott, 1998; P. T. Hill et al., 2003). Low-SES schools receive substantially fewer applications for open positions than high-SES schools because teachers are more likely to apply to schools with a higher median family income, lower crime rates, located in areas with an increased availability of amenities (Boyd, Lankford, Loeb, Ronfeldt, & Wyckoff, 2011; P. T. Hill et al., 2003; Ingersoll, 1999). Additionally, qualified senior teachers in low-SES schools are more likely to switch schools to high-SES schools just as qualified senior teachers in high-SES schools are more likely to leave the teaching profession instead of teaching at economically disadvantaged schools (Boyd et al., 2011; P. T. Hill et al., 2003; Ingersoll, 2001). Furthermore, economically disadvantaged students are more likely to be taught by teachers teaching out of their field of expertise (e.g., Ingersoll, 1999).

Access to the AP program. The origin of the educational inequities for low-SES students to access to the AP program can be seen in the historical roots being designed in the 1950s as a program for the most gifted students in order to challenge them with rigorous content preparing 'the best and brightest' for leadership roles in science and politics (e.g., Schneider, 2009). Although overall access to the AP program has substantially increased over the last decades, low-SES students still have limited opportunities to attend AP programs for a variety of reasons. Both tracking systems and a lower number AP course offerings are hindering low-SES students to enroll in advanced coursework (e.g., Klopfenstein, 2004; Klugman, 2013; Schneider, 2009; Zarate & Pachon, 2006). Therefore, extensive efforts to increase access for low-SES students to AP programs have been undertaken (e.g., Conger, Long, & Iatarola, 2009; Lichten, 2010; The College Board, 2014c; Wyatt & Mattern, 2011). However, simply raising access to the AP examination for low-SES students does not increase the percentage of students passing AP examinations (e.g., Hallett & Venegas, 2011; Lichten, 2010), commonly viewed as a score of 3 or higher. Data presented in the *10th Annual AP Report to the Nation* (The College Board, 2014c) indicate an increase of the AP participation of low-SES students, defined by the College Board as students eligible to free- or reduced lunch programs, from 11.4 percent (N=58,489) in 2003 to 27.5 percent (N=275,864) in 2013. However, in 2013, only 21.7 percent of the low-SES students scored a passing grade on at least one AP examination compared to 75.3 percent of the non-low-SES students (The College Board, 2014c).

Sociocultural dimensions of high-stakes assessment. Suppose students in well-equipped low-SES schools have access to advanced coursework, being optimally taught by well-qualified teachers facilitating students' learning and encouraging them to take high-stakes assessments. Students' performance might still be mitigated because students' individual backgrounds might not coincide with dominant occurring cultural and societal *Discourses* (Gee, 2008) in science classrooms and students are suffering from *stereotypes threats* (Steele, 1997). Discourses are shaping students' internal processes including mental functioning, problem-solving abilities, among others, and influence students' engagement with the science instruction ultimately impacting students' test performance (e.g., Noble et al., 2012; Solano-Flores & Nelson-Barber, 2001). Examples include the use of specific language in science contexts and familiarity with different question formats being used in assessments (e.g., Noble et al., 2012). Especially, for English language learners (ELL) this poses additional challenges due to false assumptions of assessment design about the interaction ability of assessment systems with ELL (e.g., Abedi, Hofstetter, & Lord, 2004; Solano-Flores, 2008). Stereotype threats are induced through individuals' identification with groups, in which one is stereotyped (e.g., Steele, 1997). Additionally, stereotype threats are more heavily affecting persons who care about their performance in the tasks they are stereotyped in (e.g., Steele, 1997). Furthermore, stereotype threats are additive such that ethnic minority, low-SES, female, ELL students might experience multiple stereotype threats while taking the AP science examination.

The individual teachers. Teacher characteristics and the quality of instruction are widely regarded as important preconditions for students' success on the AP science examinations (e.g., Hallett & Venegas, 2011; Klopfenstein, 2004; Lichten, 2010). As Abell (2007) illustrates given the mediocre teaching quality of college courses held by well-respected scholars in their scientific fields, content knowledge is not a sufficient criteria for quality of instruction. Extending Shulman (1986) models of teacher knowledge of *Subject Matter Knowledge*, *Pedagogical Content Knowledge*, and *Curricular Knowledge*, Ball, Thames, and Phelps, (2008) developed the conceptual framework of *Content Knowledge for Teaching* emphasizing a multi-faceted approach on teachers' knowledge necessary for good instruction.

Although not sufficiently measuring teachers' knowledge and skills, researchers tend to use variables such as educational degree attainment, results of basic skill tests, or courses as an estimation method (H. C. Hill, Rowan, & Ball, 2005). For instance, Garet, Porter, Desimone, Birman, and Yoon (2001) used teachers' experience in years, whether teachers have an in-field certification, and a composite variable indicating whether teachers' PD participation enhanced teachers knowledge in the following fields: "(a) curriculum […]; (b) instructional methods […]; (c) approaches to assessment […]; (d) use of technology in instruction […]; (e) strategies for teaching diverse student populations […]; and (f) deepening knowledge of mathematics" (p. 929). Banilower, Heck, and Weiss (2007) used teachers' experience in years as a contextual variable and created composite variables including "attitudes toward standard-based teaching" (p. 380), "perceptions of pedagogical preparedness" (p. 380), " perceptions of science content preparedness" (p. 380), "use of traditional teaching practices" (p. 380), "use of investigative teaching practices" (p. 380), and "use […] designated instructional materials" (p. 381). Penuel, Fishman, Yamaguchi, and Gallagher (2007) used a composite variable called "knowledge of pedagogy" (p. 942) beside teachers' educational degree attainment and whether teachers hold a science-education certification. As Desimone (2009) indicates in the empirically grounded path model "for studying the effects of professional development on teachers and students" (p. 185) associations between teachers' knowledge and skills, instructional practice, and student achievement exist.

The PD activities. Systematically conducted empirical research studies in the last decades on best practices of PD (e.g., Banilower et al., 2007; Borko, 2004; Garet et al., 2001; Wilson & Berne, 1999) lead to a consensus of certain PD characteristics that constitute 'high quality' PD. Desimone (2009) describes these five "core features" (p. 184) as (a) content focus, (b) active learning, (c) coherence, (d) duration, and (e) collective participation (p. 183).

Content focus. Content focus refers to both pedagogically focused content knowledge and discipline specific knowledge. Consistent with the concept of "content knowledge for teaching" (Ball et al., 2008), the more knowledge dimensions PD activities include, the greater the gains towards teachers' knowledge and skills and changes in classroom practice (Desimone, Porter, Garet, Yoon, & Birman, 2002; Garet et al., 2001; Heller, Daehler, Wong, Shinohara, & Miratrix, 2012; Penuel et al., 2007; Roth et al., 2011). For instance, PD activities providing teachers with examples how to effectively support students' scientific inquiry processes as well as letting teachers work inquiry-based fosters changes in teaching practices (Penuel et al., 2007; Roth et al., 2011).

Active learning. PD activities incorporating active learning emphasize the importance of teachers self-constructing knowledge and active engagement in thinking processes. Methods fostering active learning include reviewing student work, observing expert teachers teaching, being observed at actual classroom teaching, and engaging teachers to lead discussions, among other features, opposed to non-engaging PD activities such as forcing teachers to solely listen to lectures or stubbornly enacting pre-prepared curricula. Research indicates that PD including opportunities for active learning supports teachers' knowledge building, the use of technology, and implementations of higher order instructional practices (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Birman, Desimone, Porter, & Garet, 2000; Desimone et al., 2002; Garet et al., 2001; Heller et al., 2012; Penuel et al., 2007).

Coherence. Coherence refers to the alignment of the PD activity to communication structures with colleagues, department chairs, and the school principal, as well as to alignment with existing curriculum frameworks, standards, and assessment systems such as the redesigned AP science curriculum and assessment. Additionally, coherence also refers to alignment with teachers' prior PD experiences, their instructional practices, and professional goals. However, teachers' perception of coherence is often mediated through contextual factors including teachers' prior knowledge and beliefs (e.g., Coburn, 2001; Cuban, Kirkpatrick, & Peck, 2001)

Generally, research indicates that PD activities with high coherence have associations with increases in teachers' knowledge and skills and changes of classroom practices (Garet et al., 2001; Penuel et al., 2007).

Duration. The duration of PD activities can quantitatively be measured through the length of total contact time as well as the time span in which PD activities take place. Research indicates that increases in duration are associated with increases of teachers' knowledge and skills (Banilower et al., 2007; Birman et al., 2000; Desimone et al., 2002; Garet et al., 2001; Heller et al., 2012; Penuel et al., 2007; Roth et al., 2011). An increased PD duration might provide the necessary time needed to process knowledge and skills (e.g., Desimone, 2009). Also, it increases opportunities to incorporate other features of "high quality" PD, as shown in the study of Garet et al. (2001) through positive relations to content focus and teaching strategies, active learning, and coherence.

Collective participation. Collective participation describes the different forms of participants' PD attendance. PD activities might be attended by groups of individual teachers, of department representatives of the same or different schools, whole grade level teachers, or the whole teaching body of a school. Research studies indicate that, the more teachers with shared background characteristics attend PD activities, the higher are gains on general teacher change (Penuel et al., 2007), active learning (Garet et al., 2001), coherence (Garet et al., 2001), and the use of technology (Desimone et al., 2002). For instance, teachers might use of technology more often because they feel lower barriers to draw from technological knowledge from colleagues (e.g., Desimone et al., 2002). These findings indicate that teachers engage in informal learning potentially due to more trustful and supportive relationships with peers, besides participation in formal PD activities, which might even have stronger associations with learning outcome

(Penuel et al., 2007). These associations might be sustainable because teachers with similar backgrounds who participate in PD might gain shared values and ideas and therefore, might enhance the coherence of implementation in a school setting, which could in turn, yield more sustainable changes in classroom practice (Birman et al., 2000; Desimone et al., 2002; Garet et al., 2001; Heller et al., 2012).

Although, prior research has established a foundation of characteristics of 'high quality' PD activities, the alignment of PD activities to specific curricula (Fishman et al., 2013) and the relations of teachers' patterns of PD participation towards increased teacher learning and student achievement still require systematic empirical explorations (H. C. Hill, Beisiegel, & Jacob, 2013). Therefore, aligned with Borko, Jacobs, and Koellners' (2010) suggestions that PD needs to attend to aspects of change in the school context, offering interactions with model instructional strategies and opportunities to build professional learning communities, this study extends Desimone's (2009) list of core PD features with a more nuanced and interactive approach better suiting the complex structure of educational systems. Building on Opfer and Pedders' (2011) framework the relations of teacher characteristics and teachers' PD participations supporting teaching the redesigned AP science courses within teachers' school context towards student achievement on the AP science examinations are examined.

3 Research questions

Given the mandated top-down implementation of the AP science curriculum reform and changes in the large-scale high-stakes AP science assessment, hundreds of thousands of students taking the redesigned AP science examinations and tens of thousands of AP science teachers are affected by this fundamental nationwide change. This is a unique opportunity to study how teachers are responding to a change of this scale. This research study focuses on one of the most vulnerable groups in the U.S. educational system, students in low-income schools. The study attempts to contribute to the overall research base by identifying impactful factors on at-risks students' performance on large-scale high-stakes assessments. Through the analysis, this study intends to identify areas of research that can guide recommendations for educational policy makers and practitioners on how to raise student achievement in low-income schools, ultimately striving to foster educational equity throughout the U.S. The analysis is framed by the following two research questions:

1) How do key characteristics of teachers and schools (e.g., PD participation, teachers' knowledge and experience, principal support) compare across the AP science teacher population in low-income schools?
2) What is the relationship between school, teacher, and teaching characteristics on students' AP performance leap controlling for student characteristics?

We define low-income schools as schools with at least 50% of the student population enrolled in free- or reduced lunch programs. Students' AP performance leap is defined as the difference between students' AP science actual scores and the scores that would be predicted by their PSAT examination scores.

4 Methodology

This study follows the methodological tradition of a large-scale quasi-experimental design study for generalized causal inference using survey research and measurement, focusing on quantifiable effects.

4.1 Data sources

The main data source in this study consists of survey responses from a web-based survey which is annually (2013 - 2015) sent to every AP science teacher (2013: Biology; 2014: Biology, Chemistry; 2015: Biology, Chemistry, Physics) teaching redesigned AP science courses during the school year in which the survey is distributed. This study only uses data from the 2014 AP Biology and 2014 AP Chemistry surveys. Additional data is provided by the College Board supplying a large data set including student, additional teacher, school, and district information. The data collection has a nested data structure with student, teacher, school, and district level data. Using unique identifiers, teacher level data is tied to the school level data but student data can only be linked to the school in which students took their AP science examinations, due to missing identifiers for student data to their corresponding teachers. School level data is tied to district level data.

All data analyses are conducted with the largest meaningful sample possible. For instance, data preparations include applying missing data approaches, creating composite variables, and using all available observations and variables. However, in combining the data stets across the levels, observations with ambiguous identifying information are dropped. These includes students taking the AP science examinations in schools with more than one AP science teacher in the corresponding discipline or teachers' who are associated with two or more different schools.

Student level data. Student level data is provided from the College Board. This data consists of self-reported demographic information (e.g., students' ethnical background, parents' education level, languages, GPA) for all students taking the AP science examinations. Additionally, the College Board reports sub- and final scores on the PSAT, SAT, and all AP examinations including the corresponding test dates. For this study, data on students taking the AP Biology and AP Chemistry examinations in May 2014 is used.

Teacher level data. Teacher level data is gathered through a web-based survey emailed to every AP science teacher in late May 2014. The surveys differ only marginally across subject areas. Besides changing the wording of the AP science subject (Chemistry to Biology/Physics), other changes are limited to the inclusion and exclusion of discipline-specific PD activities. For instance, the *AP Central Webcast: Exploring Atomic Structure Using Photoelectron Spectroscopy (PES) Data* provided by the College Board is AP Chemistry specific. Survey questions ask for information on *demographics* (e.g., ethnicity, gender, age), *teaching background* (e.g., teaching experience, university education), *PD participation* (indicating participation and involvement in traditional PD activities including face-to-face workshops, online courses, and online teacher communities as well as other teaching preparation materials), *general attitudes towards PD* (e.g., perceived effectiveness of PD, belonging to professional organizations, engagement as AP Exam Reader or AP Consultant), *characteristics of their AP science course* (e.g., length of instruction, number of students / sections / preps, AP enrollment and fees), *characteristics of their AP science instruction and school context* (e.g., teaching practices and challenges at the time of responding to the survey and retrospectively one year ago, teaching self-efficacy, school equipment, principal support), and their *levels of concern* (at the time of responding to the survey and retrospectively one year ago) regarding this mandated top-down curriculum reform.

School level data. School level data is provided from the College Board. The data include demographic information (e.g., name, school type, location) and school characteristics (e.g., enrollment, enrollment in free- and reduced lunch programs, ethnic make-up, offering of special education classes). Also**,** data on districts school funding allocations is provided from the College Board. The data includes per student expenditures, subdivided in total expenditures and expenditures for instructional material only.

4.2 Population and sample

Our overall population for the 2014 Biology and Chemistry data consists of all students taking the AP Biology examination ($N_{\text{Students, Biology}}$ = 203,304) and all students taking the AP Chemistry examination ($N_{\text{Students, Chemistry}}$ = 133,323) in May 2014. The web-based survey was sent to AP Biology teachers ($N_{\text{Teachers, Biology}}$ = 9,511) and AP Chemistry teachers ($N_{\text{Teachers, Chemistry}}$ = 7,098) who did not opt out of College Board's professional email communication. The survey link was opened by 2,646 AP Biology teachers (opening rate: 27.82 %) and 2,732 AP Chemistry teachers (opening rate: 38.49 %) and 2,408 AP Biology teachers (response rate: 25.32 %) and 2,493 AP Chemistry teachers (response rate: 35.12 %) responded to the survey. Response rates of greater than 25 % are considered good for web-based surveys with this population size (Shih & Fan, 2009). A response is counted if teachers answered at least the first two sections (PD participation and specific questions about the PD participation) of the survey; the amount of attrition was 172 teachers for AP Biology and 159 teachers for AP Chemistry.

All data preparation procedures are conducted using the dataset with 203,304 AP Biology students, 2,408 AP Biology teachers, 133,323 AP Chemistry students, and 2,493 AP Chemistry teachers. This sample is called "full sample." However, it is important to evaluate the patterns of non-respondents (students whose teachers didn't respond to the survey) in order to estimate the selection bias using the reduced data set. The non-response analysis for each discipline is conducted three-fold applying a case-wise deletion approach for missing data. On the student level, students' PSAT and AP science scores are compared. On the teacher/school level the percentage of students enrolled in free- or reduced lunch programs, as a measure for socioeconomic status, is compared. Table 4-1 summarizes this analysis.

Table 4-1: Summary of non-response analyses, rounded to the second decimal place.

		Observations		**Mean (SD)**
		[n]	[%]	
PSAT Scores (Biology)	Respondents	34,813	17.12	166.35 (27.09)
	Non-Respondents	129,188	63.54	164.92 (27.51)
	Missing	39,303	19.33	-
PSAT Scores (Chemistry)	Respondents	36,300	27.23	174.33 (26.59)
	Non-Respondents	74,811	56.11	173.28 (27.09)
	Missing	22,212	16.66	-
AP Biology Scores	Respondents	43,463	21.38	3.00 (1.04)
	Non-Respondents	159,841	78.62	2.90 (1.05)
	Missing	0	0	-
AP Chemistry Scores	Respondents	43,079	32.31	2.81 (1.24)
	Non-Respondents	90,244	67.69	2.63 (1.26)
	Missing	0	0	-
Free & Reduced-Price Lunch (Biology)	Respondents	41,862	20.59	24.69 % (22.91 %)
	Non-Respondents	154,507	76.00	26.94 % (24.13 %)
	Missing	6,945	3.42	-
Free & Reduced-Price Lunch (Chemistry)	Respondents	41,201	30.90	22.19 % (21.48 %)
	Non-Respondents	86,318	64.74	24.69 % (22.82 %)
	Missing	5801	4.35	-

Non-parametric Mann-Whitney tests are conducted to compare the mean values on each variable between respondents and non-respondents. The effect size of the differences is calculated using Cohen's *d*. Conducting these tests indicate that students' PSAT scores for teachers who didn't respond to the survey are significantly lower than students' PSAT scores for teachers who respond to the survey for both AP Biology, z= - 9.35, $p < .001$, d = -.052, and AP Chemistry, z = -5.60, $p < .001$, d=-.039. Additionally, students' AP scores for teachers who didn't respond to the survey are significantly lower than students' AP scores for teachers who respond to the survey for both AP Biology, z = -17.46, $p < .001$, d = -.095 and AP Chemistry, z = -24.71, $p < .001$, d = -.143. Furthermore, schools in which teachers didn't respond to the survey have significantly higher percentages of students enrolled in free- or reduced lunch programs than schools in which teachers responded to the survey for both AP Biology $z = 15.89$, $p < .001$, d=.094, and AP Chemistry, $z = 18.28$, $p < .001$, $d = .112$.

The sample population in low-income schools for this study consists of the combined student and teacher data set including 11,800 AP science students (AP Biology: 6,410 students; AP Chemistry: 5,390 students) and 638 AP science teachers (AP Biology: 318 teachers, AP Chemistry: 320 teachers). Teacher and corresponding student data is only included in this combined data set if teachers are teaching within the United States, teachers are the only AP science teacher in the discipline, teachers are only affiliated with one school, and teachers are teaching at least one student who is taking an AP science examination in the corresponding discipline.

The subsequent statistical analyses are based on the reduced data set with 11,800 AP science students and 638 AP science teachers. This data set will be called "low-income sample" in the subsequent analysis. With the exception of the creating composite variables, all subsequent statements refer to the low-income sample, unless otherwise indicated. For instance, 20% of missing student data should be interpreted as 20% of 11,800 yielding to 2,360 missing cases.

4.3 Analytical methods

This section describes the analytic methods used for this study. This includes data preparation strategies (e.g., missing data approaches, factor analysis, etc.) and the statistical methods used for exploring the research questions (e.g., analysis of variance, hierarchical linear modeling, etc.).

Data preparation. Before conducting statistical tests, the data sets were prepared for the analysis. Initially, the AP Biology and AP Chemistry data sets are treated in different files. After initial data preparations, missing data strategies are applied. Afterwards, composite variables are computed. Both the missing data approaches and the computation of composite variables are done with the "full sample," and not the "low-income sample" of 11,800 AP science students and 638 AP science teachers, to reduce sampling bias. However, to account for potential differences across subject areas, both missing data strategies and computing of composite variables are conducted on the discipline specific data sets. As a final step, the AP Biology and AP Chemistry data sets are merged to generate a combined AP science data set.

Initial data preparation strategies. Initial data preparation strategies include deleting cases from the teacher data set that do not contain meaningful information. For instance, all observation of teachers opening the email with the survey invitation but exiting the survey before completing the first two survey sections are dropped from the data set. Another data preparation strategy is to recode survey responses for "check all that apply" survey items. For instance, teachers are asked to check all PD activities they participated in from a pre-defined list of about 20 different PD options. If teachers checked off a box, the response is coded as "1." If teachers are not checking the box, the response is coded as "0" instead of missing. This is a limitation of the data because teachers who didn't answer these questions cannot be distinguished from teachers not participating in corresponding PD activities.

Missing data approaches. Missing data is imputed using a Markov Chain Monte Carlo (MCMC) multiple imputation method with 150 iterations and 40 imputations yielding to a power falloff of considerable less than 1% by comparison to a full information maximum likelihood (FIML) approach (e.g., Graham, 2009; Graham, Olchowski, & Gilreath, 2007). The imputed datasets are collapsed averaging the values for each variable across the imputed data sets to create a merged data set (Cheema, 2014), in order to receive unbiased standard errors of the parameter estimates (Rubin, 1978 in Cheema, 2014, p. 8). Calculations are based on the merged data set.

Missing data on variables of the teacher data set is imputed in two stages. This teacher data set includes teacher and teaching variables, as well as school level variables. Data from the teacher data set is imputed separately for the 2014 AP Biology and 2014 AP Chemistry data sets. Stage 1 uses all variables and all observations of the full data set in survey sections are answered by every teacher. Stage 2 includes variables resulting from responses of the "Specific PD Questions" survey sections for each of the predefined PD activities and all observations from the full data set. These survey sections are only displayed to teachers indicating that they participated in this PD activity and result in considerably lower total responses. Missing data on these sections is only imputed for variables on each individual section. The percentage of missing data is almost below 5% for all variables, tables A3 and A4 describe the missing data percentages for each variable included in the statistical models (cf. Appendix, Section 8.1).

For the imputing missing data on student and school variables, all available variables and observations provided by the College Board and all school level variables of the combined 2014 AP Biology and 2014 AP Chemistry data sets are used. The percentage of missing data is below 8% for almost all variables, as described in tables A1-A4 (cf. Appendix, Section 8.1).

Computing composite variables. On the teacher level, composite variables are computed using all "Stage 1" variables. The analysis is conducted on two equal-sized independent data sets using random sampling. On each of these independent data sets, exploratory and confirmatory factor analysis are conducted.

Exploratory factor analysis (EFA) is conducted to evaluate the contributions of survey items to a latent variable. The number of retained factors is based on the Guttman-Kaiser criterion (Guttman, 1954; Kaiser, 1961), stating that the eigenvalues of a factor need to be greater than one, and on the analysis of scree plots. Items are gradually excluded from a composite variable if the factor loading for an item on every factor is below a threshold of an absolute value of 0.25. This is a very conservative approach due to our large sample size. For instance, Stevens (2009) recommends a critical value below 0.1 for a p=.01 significance level with a sample size of 1,000. However, our threshold is still below conventional used thresholds of 0.3-0.5 (e.g., Grice, 2001). Extracting parameters is based on normalized oblimin oblique rotation methods, which are powerful in simple structures (Lorenzo-Seva, Kiers, & Berge, 2002). Oblique rotations use the conceptual assumption that the resulting factors are correlated to each other (e.g., Abdi, 2003). Although orthogonal rotation methods are more frequently used in the social sciences, some researchers argue that accounting for correlations among latent variable constructs might yield more generalizable results (e.g., Costello & Osborne, 2005).

Confirmatory factor analysis (CFA) using the maximum likelihood estimation method is conducted to compare the model fit of several different statistical models (e.g., Brown, 2006; Harrington, 2009). The compared models are based on the EFA and vary in two dimensions. First, the covariances between latent variables are either forced to zero or not. Second, non-significant paths are included or excluded for model estimations. Comparing the model fit is based on several goodness-of-fit statistics including the root mean squared error of approximation (RMSEA: the smaller, the better the model fit), Akaike's information criterion (AIC: the closer to zero, the better the model fit), Bayesian information criterion (BIC: the closer to zero, the better the model fit), the comparative fit index (CFI: the closer to 1, the better the model fit), the Tucker-Lewis index (TLI: the closer to 1, the better the model fit), and the standardized root mean squared residual (SRMR: the smaller, the better the model fit) (e.g., Brown, 2006; Harrington, 2009). Additionally, likelihood ratio tests are conducted to compare the model fit of nested models. Only the results of the best model are reported in the analysis.

After establishing a factor model based on EFA, CFA, and checks for conceptual reasonability the factor scores are computed using blockwise factor score regression approaches (Skrondal & Laake, 2001). The calculated Bartlett factor scores are maximizing validity both through high correlations to true factor values and unbiased estimates of factor score parameters (e.g., DiStefano, Zhu, & Mindrila, 2009) and can be seen as advantageous for subsequent structure equation modeling (Skrondal & Laake, 2001). Using this refined method produces standardized factor scores with means of zero and standard deviations approaching one (DiStefano et al., 2009). Afterwards, using the retained items for each factor score, Cronbach's α and the average interitem covariance are computed to estimate the internal consistency of the generated scales.

Exploring research questions. Using the prepared data set, the research questions (RQ) are explored using hypothesis testing to compare characteristics of the AP science teacher population in low-income schools (RQ1) and multi-level linear modeling exploring the relationships of teacher, school, and district characteristics on students' performance on the AP science examination, controlling for student characteristics (RQ2).

RQ1: Characteristics of the AP science teacher population. Key characteristics are compared among three groups of the AP science teachers, teachers teaching students performing on average (a) worse-than-expected, (b) as-expected, and (c) better-than-expected on the AP science examination than predicted by their PSAT scores. All variables analyzed exploring the first research question are also included in the final model exploring the second research question. Independently testing for differences on each characteristic across the three groups, parametric one-way analysis of variance (ANOVA) tests are conducted.

Assumptions for using ANOVA include that the observations are independent, the tested variable is normally distributed in each group, and that the variances are equally distributed (homoscedasticity) opposed to unequal distributed variances (heteroskedasticity) (e.g., Coolidge, 2000; Downie & Heath, 1983; Field, 2009). The observations are independent because teachers are uniquely distributed across all three groups. ANOVA are fairly stable against non-normal distributions, usually resulting in small effects of Type I errors (e.g., Coolidge, 2000; Downie & Heath, 1983; Field, 2009). Therefore, this assumption is tested through graphing plots of each variable instead of conducting formal statistical tests. If distributions do not substantially differ from a normal distribution, this assumption is relaxed. Homogeneity of variance is tested using Levene's test based on mean values if the data is normally-distributed, the Brown-Forsythe test based on the median if the data is heavily skewed (e.g., following a Chi-square distribution with four degrees of freedom), or the Brown-Forsythe test based on a trimmed mean if the data is heavily tailed (e.g., following a Cauchy distribution).

If no assumptions are violated, an ANOVA is conducted. Multiple-group comparisons are conducted using the Tukey-Kramer test because of the unequal group sizes (e.g., Kirk, 1998). If assumptions are violated, the non-parametric Kruskal-Wallis H test is conducted. Across group comparisons are computed through post-hoc Mann-Whitney tests with Bonferroni corrections. A preferred measure for effect sizes are eta-squared (e.g., Cohen, 1973; Levine & Hullett, 2002).

RQ 2: Relationship of student, teacher, and school characteristics on students' AP science performance. Exploring direct effects of teacher, and school characteristics on students' performance on the AP science examinations, controlling for student characteristics, uses two-level fixed-effects hierarchical linear modeling (e.g., Raudenbusch & Bryk, 2002; Raudenbusch, Bryk, Cheong, Congdon, & du Toit, 2004). Students are nested within teachers such that a multi-level modeling approach is necessary to analyze the relationships. Because of the missing teacher identifier data limitation, only one teacher per school is included in the sample. Therefore, a two-level modeling approach is sufficient for this data instead of introducing a third level of teachers nested within schools. In order to estimate the correlations of observations within a cluster the intraclass correlation coefficient (ICC) is computed.

Assumptions of hierarchical linear modeling include independence of observations, independence of error terms, no perfect multicollinearity between independent variables, homoscedasticity of residuals, normality of residuals, and that independent variables have a linear relationship with dependent variable, among others (e.g., Raudenbusch & Bryk, 2002). The observations are independent because the combination student-teacher are uniquely distributed in the data. Multicollinearity of the independent variables on both levels is tested calculating variance inflation factors. Testing homoscedasticity of residuals uses the same methods as for exploring the first research question.

4.4 Measures

This section describes dependent and independent variables used in the statistical analysis displaying basic descriptive information. Single indicator and composite variables are separately described on the student, teacher, and school level.

Dependent variable. The dependent variable used for the analysis is a continuous score comparing students' performance on the AP science examination with their performance on the PSAT examination. Students' PSAT performance is used as an academic achievement measure prior to students enrolling in AP science classes and taking the AP science examinations. PSAT scores for students in this study correlate strongly with AP science scores, $r = .672$, $p < .001$, which is consistent with prior research (e.g., Ewing, Camara, & Millsap, 2006; Lichten, 2010; Lichten & Wainer, 2000), such that students' PSAT scores can be viewed as a predictor of AP science performance.

For comparing students' PSAT and AP performance, the original PSAT score range is transformed into a 1-5 scale, using the percentiles of the AP science grade distributions for the 2014 AP Biology and 2014 AP Chemistry examinations. Next, the difference between AP scores and transformed PSAT scores is computed for each individual student. A positive difference indicates that a student is performing better on the AP examination than expected based on the PSAT performance, and vice-versa. Figure 4-1 depicts this variable for the low-income sample of students taking the 2014 AP Biology (*N* = 6,410, *M* = .069, *SD* = .727) and 2014 AP Chemistry (*N* = 5,390, *M* = .140, *SD* = .916) examinations used in the statistical analysis. The difference between AP science and PSAT scores is called "AP performance leap."

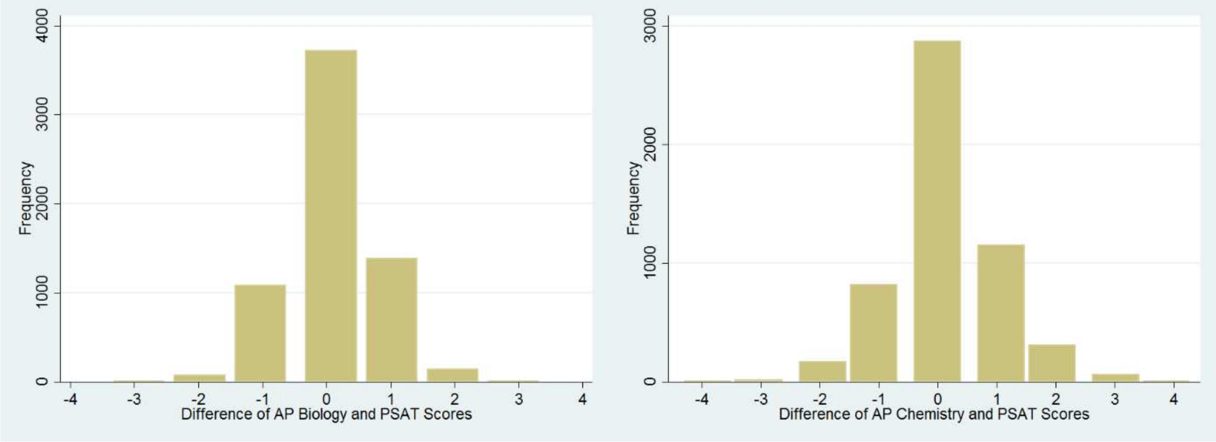

Figure 4-1: Distribution of students' AP performance leap. Left: AP Biology; Right: AP Chemistry.

To identify teachers whose students perform on average better than projected, a new continuous variable averages students' difference scores for all students taught by one teacher. The distribution of average difference scores for the combined 2014 AP Biology and AP Chemistry teachers in low-income schools (*N* = 638, *M* = .058, *SD* = .429) is illustrated in figure 4-2.

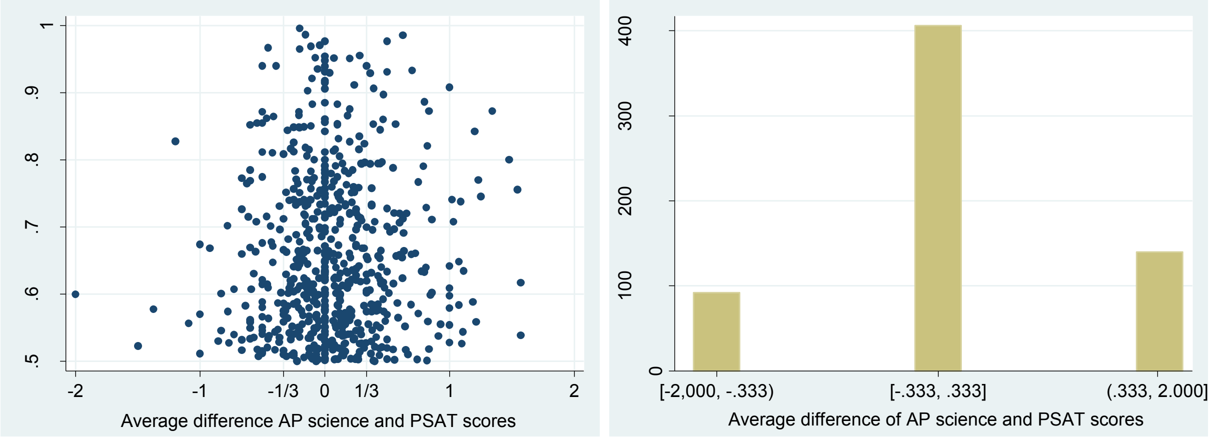

Figure 4-2: Distribution of students' average AP performance leap per teacher. Left: Scatterplot with students' average AP performance leap, Right: Frequency diagram with students' average performance leap below -1/3, from -1/3 to 1/3, and above 1/3.

Single indicator independent variables. Single indicator independent variables are included on the student and teacher/school/district level. The statistics of the variables refer to the low-income sample of combined AP Biology and AP Chemistry students, teachers, and schools included in the statistical analysis. The statistics includes imputed values.

Student level. Student level single indicator independent variables are included as covariances in order to reduce confounding effects. Included are a series of dummy variables regarding students' ethnic background and a variable indicating whether English is a students' best language. The coding of the student level single indicator variables is shown in table 4-2.

Table 4-2: Coding of student level single-indicator independent variables. The mean values of the ethnic make-up variables do not add up to 1 due to rounding effects. D: Dichotomous variable, O: Ordinal variable, C: Continuous variable.

Variable	Description	Values		Mean
S_White	Students' ethnicity: White (0: No, 1: Yes)	D	0,1	.373
S_Black	Students' ethnicity: Black or African American (0: No, 1: Yes)	D	0,1	.151
S_Native	Students' ethnicity: American Indian or Alaska Native (0: No, 1: Yes)	D	0,1	.010
S_Asian	Students' ethnicity: Asian, Asian American, or Pacific Islander (0: No, 1: Yes)	D	0,1	.183
S_Hispanic	Students' ethnicity: Mexican, Mexican American, Puerto Rican, or other Hispanic, Latino, or Latin American (0: No, 1: Yes)	D	0,1	.279
English language	Students' best language (0: Other than English, 1: English or English and another language)	D	0,1	.977

Teacher level. Single indicator variables on the teacher level include demographic information including age, gender, a series of dummy variable indicating teachers' ethnical background, teachers' focus of their prior degree program, and teaching relating practices. The variable indicating number of courses has a ceiling effect because a "31" represents "30 or more courses" instead of "31 courses." For more information, please refer to table 4-3.

Table 4-3: Coding of teacher level single-indicator independent variables. The mean values of the ethnical make-up variables do not add up to 1 due to rounding effects. D: Dichotomous variable, O: Ordinal variable, C: Continuous variable.

Variable	Description	Values		Mean (SD)
T_White	Teachers' ethnicity: White (0: No, 1: Yes)	D	0,1	.770
T_Black	Teachers' ethnicity: Black or African American (0: No, 1: Yes)	D	0,1	.080
T_Native	Teachers' ethnicity: American Indian or Alaska Native (0: No, 1: Yes)	D	0,1	.022
T_Asian	Teachers' ethnicity: Asian, Asian American, or Pacific Islander (0: No, 1: Yes)	D	0,1	.082
T_Hispanic	Teachers' ethnicity: Mexican, Mexican American, Puerto Rican, or other Hispanic, Latino, or Latin American (0: No, 1: Yes)	D	0,1	.071
Age	Teachers' age in years	C	22-67	43.26 (10.39)
Gender	Teachers' sex (0: Male, 1: Female)	D	0,1	0.651
Degree	Teachers' higher education degree attainment (1: Associate's degree, 2: Bachelor's degree, 3: Master's degree, 4: Certificate of advanced study, 5: Doctoral degree)	O	1-5	2.91 (.75)
Courses	Number of disciplinary courses taken in college and graduate school	C	0-31	13.41 (7.70)
Labs	Number of laboratory investigations students complete in school year	C	0-26	12.70 (5.50)

School level. On the school level, the variables included are an indicator of the existence of an enrollment criteria for AP science courses; the length of the school year (in days), an indicator of offering of special education classes, an indicator whether the school is a charter school, a series of dummy variables indicating the school neighborhood, and the percentage of students' enrolled in free- or reduced lunch programs. For more information, please refer to table 4-4.

Table 4-4: Coding of school level single-indicator independent variables. The mean values of the school neighborhood variables do not add up to 1 due to rounding effects. D: Dichotomous variable, C: Continuous variable.

Variable	Description	Values		Mean (SD)
Enrollment criteria	Enrollment criteria for AP course (0: No, 1: Yes)	D	0,1	.549
Length of school year	Length of the school year in days (calculated from start- and end date of the AP course)	C	1-351	275.78 (33.08)
Special education	School offers special education classes (0: No, 1: Yes)	D	0,1	.931
Charter	School is a charter school (0: No, 1: Yes)	D	0,1	.020
Rural	School neighborhood: Rural/non-metro (0: No, 1: Yes)	D	0,1	.141
Suburban	School neighborhood: Suburban (0: No, 1: Yes)	D	0,1	.315
Urban	School neighborhood: Urban (0: No, 1: Yes)	D	0,1	.163
Town	School neighborhood: Town (0: No, 1: Yes)	D	0,1	.381
Lunch program	Percentage of students enrolled in free- or reduced lunch program	C	0.5-1.0	.659 (.121)

District level. On the district level, two variables indicating districts' funding allocations are included in the statistical model. Both variables have some uncertainty because their initial coding is describing ranges instead of discrete values. In order to create a continuous variable, the mean value for each range is computed. Furthermore, these variables have ceiling effects because a "500.00" represents "$500.00 or more dollars" instead of "$500.00" and "13.00" represents "$13,000.00 or more dollars" instead of "$13,000.00." Descriptive information are described in table 4-5.

Table 4-5: Coding of district level single-indicator independent variables. C: Continuous variable.

Variable	Description	Values		Mean (SD)
District funding: Materials	Per students instructional materials expenditures in U.S. Dollar	C	72.50-500.00	220.83 (111.42)
District funding: All	Total per students expenditures in $1,000 U.S. Dollar	C	3.25-13.00	9.00 (2.34)

Composite independent variables. Composite independent variables are included on the student, teacher, and school level. Descriptive statistics for variables resulting from the exploratory and confirmatory factor analysis (cf. Section 4.3.1) refer to the combined factor scores of the 2014 AP Biology and AP Chemistry data for student and teachers in low-income schools. Composite variables resulting from the factor analysis are individually computed for both Biology and Chemistry data using the full-sample. Table A6-A12 (cf. Appendix, Section 8.2) show the rotated factor loadings and scoring coefficients for all survey items generating the composite variables. The variables described in this section are based on the low-income sample with the exception of Cronbach's α, which refers to the arithmetic average of the Cronbach's α for the AP Biology and AP Chemistry data using the full sample.

Student level. On the student level, the composite variables include parents' educational level and students' AP GPA. Both variables are not computed through the factor analysis approach described above. Parents' educational attainment is computed through the arithmetic average of mother's and father's education level for each student. Students' AP GPA is computed through the average of all AP examination scores for each student, with the exception of the corresponding AP science score. For instance, students with provided AP Biology teacher data exclude their AP Biology examination score from the AP GPA variable but potentially include their AP Chemistry examination score. Descriptive information are shown in table 4-6.

Table 4-6: Coding of student level composite independent variables. C: Continuous variable.

Variable	**Description**	**Values**		**Mean (SD)**
Parents education	Average parental educational attainment (1: Grade school, 2: Some high school, 3: High school diploma, 4: Business or trade school, 5: Some college, 6: Associate's degree, 7: Bachelor's degree, 8: Some graduate or professional school, 9: Graduate or professional degree)	O	1-9	4.73 (2.22)
AP GPA	Students' grade point average on all AP examinations, excluding disciplinary AP science examination score	C	1-5	2.24 (.946)

Teacher level. On the teacher level, composite variables include teacher- and teaching-centered composite variables as well as composite scores for the features of PD activities teachers participated in. All teacher level composite variables are continuous. Teacher-centered variables include teachers' knowledge and experience and teachers' PD inclination. Teaching-centered variables include the enactment of practices and the curriculum of the AP redesign and teachers' challenges with the AP redesign. Descriptive information are described in table 4-7.

Table 4-7: Coding of teacher- and teaching-centered composite independent variables. ^{C}: Continuous variable, ^{O}: Ordinal variable, $^{\times}$: 5-point Likert scale item, $^{(-)}$: negative scoring coefficient

Variable	**Description**	**Range**	**α**	**Mean (SD)**
Teachers' knowledge and experience	Composite variable: (a) years teaching high school science ^{C}, (b) years teaching AP Biology/Chemistry ^{C}, (c) number of professional organizations related to science teaching ^{C}, (d) number of conference attendances organized by these professional organizations within the past three years ^{C}, (e) years serving as AP Reader, (f) years serving as AP Consultant ^{C}, (g) time point being assigned to teach AP science this year$^{O, (-)}$	[-1.94, 3.77]	.55	-.330 (.857)
PD inclination	Composite variable: (a) importance PD to instructional performance$^{\times}$, (b) importance PD to student performance$^{\times}$, (c) self-teaching is as effective as formal PD participation$^{\times, (-)}$, (d) teaching performance not greatly affected by PD participation$^{\times, (-)}$, (e) enjoyment participation in face-to-face PD activities$^{\times}$	[-4.52, 1.81]	.81	.168 (1.055)
Challenges with the AP redesign	Composite variable: Challenges with (a) Biology/Chemistry content$^{\times}$, (b) organization of Biology/Chemistry content$^{\times}$, (c) labs$^{\times}$, (d) inquiry labs$^{\times}$, (e) format of questions/problems/exam$^{\times}$, (f) application of science practices to the content$^{\times}$, (g) development of a new syllabus$^{\times}$, (h) understanding the "exclusion statements$^{\times}$," (i) designing new student assessments$^{\times}$, (j) using the text appropriately$^{\times}$, (k) working with a new or different textbook$^{\times}$, (l) pacing of course$^{\times}$, (m) moving students to a conceptual understanding of Biology/Chemistry$^{\times}$	[-2.91, 3.17]	.87	.179 (1.068)
Enactment AP practices	Composite variable: (a) have students work on laboratory investigations$^{\times}$, (b) provide guidance on test questions which integrate content$^{\times}$, (c) provide guidance on test questions that are open/free response$^{\times}$, (d) have students report laboratory findings to other students$^{\times}$, (e) have students perform inquiry laboratory investigations$^{\times}$	[-4.88, 3.19]	.65	.051 (1.220)
Enactment of AP curriculum	Composite variable: (a) refer to the "Big Ideas" of Biology/Chemistry$^{\times}$, (b) use a science practice in your class outside of the classroom$^{\times}$, (c) refer how enduring understandings relate to the "Big Ideas$^{\times}$," (d) refer to the learning objective from the AP curriculum in class$^{\times}$, (e) refer to the curriculum framework$^{\times}$	[-2.53, 2.40]	.83	.250 (1.088)

Teachers' PD participation is divided into pre-selected conventional and unconventional PD activities. Conventional PD activities include formal face-to-face (F2F) PDs, online PDs, and participation in teacher online communities. Unconventional PD activities include teacher meetings, mentoring and coaching activities, and conference participation as well as teachers' use of materials. Teachers' PD participation is coded as a dichotomous variable coded as 0: did not participate, 1: did participate. Each conventional PD activity is described through up to six

5-point Likert scales highlighting the key PD characteristics. The characteristics include (a) degree of active learning experiences [AL], (b) responsiveness to the needs and interests of participants [RA], (c) focus on student work [SW], (d) focus of modeling teaching [MT], (e) intentional design to build relationships with other teachers [RT], and (f) effectiveness of support for teaching the AP redesign [EF]. Teachers not participating in a specific PD activity have by default missing values on these six characteristics. The duration of the PD activity is reported within the following three categories: 1 - low duration (≤ 8 hours), 2 - moderate duration (8-40 hours), 3 - long duration (> 40 hours). Existence of PD features, duration, and number of participating teachers for both conventional and unconventional PDs are described in table 4-8.

Table 4-8: Description of PDs and participation rates. *: Computed from individual teachers' response on how often the online community is visited and how long each session last; [†]: Provided by the College Board; [‡]: Biology only; [+]: Chemistry only.

Variable	Duration	AL	RA	SW	MT	RT	EF	N
F2F: AP Summer Institute[†]	2	X	X	X	X	X	X	360
F2F: AP Fall Workshop[†]	1	X	X	X	X	X	X	125
F2F: Transition to inquiry-based labs workshop[†]	1	X	X	X	X	X	X	35
F2F: Day with AP Reader[†;‡]	1	X	X	X	X	X	X	14
F2F: Laying the Foundation, by NMSI[‡]	2	X	X	X	X	X	X	16
F2F: BSCS Leadership Academy, by BSCS and NABT[‡]	2	X	X	X	X	X	X	3
F2F: District, regional, local college, or teacher-initiated meetings	Unconventional PD: F2F							123
F2F: Mentoring or coaching one-on-one or with other teachers	Unconventional PD: F2F							92
F2F: Conferences or conference sessions	Unconventional PD: F2F							55
Online: Transition to inquiry-based labs[†]	1	-	-	X	X	-	X	18
Online: Introduction to AP Biology/Chemistry[†]	1	-	-	X	X	-	X	19
Online: AP Central Webcast: Exploring atomic structure using photoelectron spectroscopy[†,+]	1	-	-	X	X	-	X	26
Online Community: AP Teacher Community[†]	*	-	X	X	X	X	X	299
Online Community: National Science Teachers' Association (NSTA) online community	*	-	X	X	X	X	X	49
Materials: AP course and exam description[†]	Unconventional PD: Materials							609
Materials: AP lab manual[†]	Unconventional PD: Materials							543
Materials: Textbook teacher guide and related materials	Unconventional PD: Materials							457
Other Materials: Instructional materials developed by colleagues	Unconventional PD: Materials							506
Other Materials: Articles from magazines or journals	Unconventional PD: Materials							315
Other Materials: Video resources	Unconventional PD: Materials							385

Creating composite variables of conventional PD activities for each PD feature follows the idea of exposure, summing up the scores on every PD features teachers participated in. First, Likert scale items are recoded using the following approach: (initial value) → (new value); 1 → 0, 2 → 1, 3 → 2, 4 → 3, 5 → 4. Using this recoding method, if teachers report the lowest rating on a PD features this is equal to PDs with teachers not participating in this PD or participating in a PD activity that doesn't support this PD feature. Second, accounting for the dosage of PD exposure, each score on every features item is multiplied with the score of the duration category. Third, the scores on each PD features are added across all PDs to generate composite variable. For unconventional PD activities, the composite variables describe the total number of unconventional PD activities teachers engage in, separated by face-to-face and materials. The resulting composite variables are continuous. Table 4-9 illustrates descriptive information of the low-income data set.

Table 4-9: Descriptive information on conventional PD activity independent composite variables.

Variable	Range	Mean	(SD)
Active learning	[0, 12]	2.18	(2.06)
Responsive agenda	[0, 16]	3.24	(2.65)
Focus on student work	[0, 16]	2.39	(2.55)
Modeling teaching	[0, 17]	2.62	(2.67)
Building relationships	[0, 17]	3.54	(2.84)
Effective support	[0, 18]	3.86	(3.12)
Unconventional PD: Face-to-face	[0,3]	.42	(.73)
Unconventional PD: Materials	[0,6]	4.41	(1.34)

School level. On the school level, composite variables include teachers' perceived principal support and teachers' AP workload. Both school level composite variables are continuous. Descriptive information are described in table 4-10.

Table 4-10: Coding of school level composite independent variables. ^{C}: Continuous variable, $^{\times}$: 5-point Likert scale item, $^{\Diamond}$: 4-point Likert scale item, $^{(-)}$: negative scoring coefficient

Variable	**Description**	**Range**	**α**	**Mean (SD)**
Principal support	Composite variable: (a) principal understands challenges for AP Science students$^{\times}$, (b) principal understands challenges for AP Science teachers$^{\times}$, (c) principal supports PD$^{\times}$, (d) lighter teaching load for AP Science teachers$^{\times}$, (e) fewer out-of-class responsibilities for AP Science teachers$^{\times}$, (f) AP Science is given additional funding exclusively for the course$^{\times}$, (g) availability of equipment to perform labs$^{\Diamond}$, (h) availability of expendable (consumable) supplies to perform labs$^{\Diamond}$	[-2.99, 2.29]	.73	-.212 (1.125)
AP workload	Composite variable: (a) number of students across all AP Biology/Chemistry section^{C}, (b) number of AP Biology/Chemistry sections^{C}, (c) number of preps per week$^{C,(-)}$	[-1.61, 4.97]	.65	-.301 (.865)

5 Findings

5.1 Key characteristics of the AP science teacher population

The first research question asks to identify distinctive features of the AP science teacher population in low-income schools, defined as schools with at least 50 % of students enrolled in free- or reduced lunch programs. Because this research is guided by the overall attempt to inform educational policy and practitioners to improve at-risk students' learning and achievement in low-income schools, characteristics on various levels are compared among three teacher subpopulations: teachers whose students perform on average more than 1/3 of an AP science score lower than their PSAT score predict (*lower-than-expected, N = 92 teachers)*, teachers whose average students' AP science performance is within a range of 1/3 below and above their predicted score (*as expected, N = 406 teachers*), and teachers whose students perform on average more than 1/3 of an AP science score higher than their PSAT score predicts (*better-than-expected, N = 140 teachers*).

Variables included in this multiple-group analysis are on the school level (days of the school year, districts funding allocations - total and instructional materials only, percentage of students enrolled in free- or reduced lunch programs, and composite variables on principal support as well as AP workload), the teacher level (age, educational degree attainment, number of disciplinary courses taken in college and graduate school, and composite variables on teachers' knowledge and experience as well as PD inclination), regarding teaching characteristics (the number of laboratory investigations and composite variables on teachers' challenges with the AP redesign, the enactment of practices of the AP redesign, and the enactment of the curriculum of the AP redesign), and teachers' PD participation (teachers' combined ratings on the active learning components of the PD, the responsiveness of agenda to their interests and needs, the focus on student work, the occurrence of modeled teaching, the opportunity to build relationships with other teachers, and the effectiveness of support towards the AP redesign , as well as the number of unconventional PD activities teachers participated in – face-to-face and material-based).

Following the analytical methods described, homoscedasticity is tested using Levene's test if the data is normally-distributed, Brown-Forsythe test based on the median if the data is heavily skewed, or the Brown-Forsythe test based on a trimmed mean if the data is heavily tailed. Normality is assessed through graphing plots of the variable. If a variable shows heteroskedasticity or substantially differs from a normal distribution Kruskal-Wallis H test is conducted with reported chi-square with ties statistics. Otherwise, an ANOVA test is conducted with reported *F*-statistics. Effect sizes are calculated using η^2. Table 5-1 describes omnibus between-groups effects of *worse-than expected, as expected, and better-than expected teacher* groupings. Therefore, the degrees of freedom (*df*) for the effects of the model is two for all reported tests.

The analysis indicates that significant differences exist on every level between the teacher populations teaching in low-income schools with respect to their students' average academic

Table 5-1: Omnibus group comparisons using ANOVA and Kruskal-Wallis H tests. Significant effects on the .05 level or below are bolded.

Variable	*Normality*	*p (Homo-scedasticity)*	*Test*	*F or χ^2*	*η^2*	*p*
School Characteristics						
Principal support	✓	.982	ANOVA	.22	.000	.799
AP workload	✗	**.122**	**Kruskal-Wallis**	**9.66**	**.015**	**.008**
Days of school year	✗	**.172**	**Kruskal-Wallis**	**7.62**	**.012**	**.022**
District funding: Materials	✗	.683	Kruskal-Wallis	3.58	.006	.167
District funding: All	✓	**.071**	**ANOVA**	**5.84**	**.018**	**.003**
Lunch program	✗	.778	Kruskal-Wallis	3.74	.006	.154
Teacher Characteristics						
Age	✓	.604	ANOVA	1.22	.004	.295
Degree	✓	**.221**	**ANOVA**	**5.03**	**.016**	**.007**
Courses	✗	**.560**	**Kruskal-Wallis**	**7.80**	**.012**	**.020**
Knowledge and experience	✗	**.075**	**Kruskal-Wallis**	**12.30**	**.019**	**.002**
PD inclination	✓	.787	ANOVA	.07	.000	.929
Teaching Characteristics						
Labs	✓	**.875**	**ANOVA**	**3.08**	**.001**	**.047**
Challenges with the AP redesign	✓	.909	ANOVA	1.74	.005	.176
Enactment AP practices	✓	.476	ANOVA	.64	.002	.528
Enactment of AP curriculum	✓	.476	ANOVA	.34	.001	.713
PD Activities						
Active learning	✗	.275	Kruskal-Wallis	.31	.000	.857
Responsive agenda	✗	**.448**	**Kruskal-Wallis**	**6.07**	**.009**	**.048**
Focus on student work	✗	.203	Kruskal-Wallis	1.52	.002	.467
Modeling teaching	✗	.251	Kruskal-Wallis	.42	.001	.812
Building relationships	✗	.739	Kruskal-Wallis	1.09	.002	.581
Effective support	✗	.251	Kruskal-Wallis	4.58	.007	.101
Unconventional PD: Face-to-face	✗	.473	Kruskal-Wallis	1.81	.002	.406
Unconventional PD: Materials	✓	.506	ANOVA	.38	.001	.687

performance. This is a first indication that both the school environment and the teacher with his or her inherent characteristics, teaching practices, and individual PD choices are associated with differences in student achievement.

On the school level, Kruskal-Wallis H tests indicate that there are small significant differences between teachers in the *lower-than-expected, as-expected,* and *better-than-expected* groups on teachers' self-reported AP workload, $\chi^2(2, 635) = 9.66$, $p < .01$, $\eta^2 = .015$, and the days of the school year, $\chi^2(2, 635) = 7.62$, $p < .05$, $\eta^2 = .012$. Using an ANOVA indicates that there is a small to medium significance difference regarding schools' overall district funding allocations, $F(2, 635) = 5.82$, $p < .01$, $\eta^2 = .018$.

On the teacher level, ANOVA tests indicate that there is a small significance difference across the three groups regarding teachers' educational degree attainment, $F(2, 635) = 5.03$, $p < .01$, $\eta^2 = .016$. Using Kruskal-Wallis H tests indicate that there are small to medium significant differences between teachers' disciplinary coursework in college and graduate school, $\chi^2(2, 635) = 7.80$, $p < .05$, $\eta^2 = .012$, and teachers' knowledge and experience across the three groups, $\chi^2(2, 635) = 12.30$, $p < .01$, $\eta^2 = .019$.

Regarding teaching characteristics, an ANOVA test indicates small significant differences across the groups in the number of laboratory investigations conducted in the AP science courses, $F(2, 635) = 3.08$, $p < .05$, $\eta^2 = .001$.

Exploring teachers' PD participation, Kruskal-Wallis H test indicate a small significant difference across the three groups regarding teachers' combined ratings of the responsiveness of the agenda of the PD to teachers' interests and needs within their formal PD participations, $\chi^2(2, 635) = 6.07$, $p < .05$, $\eta^2 = .009$. Additionally, teachers' ratings of their whole formal PD participation as effectively supporting their needs with respect to the redesigned AP science course are almost approaching significant difference across the three groups, as indicated by the Kruskal-Wallis H test, $\chi^2(2, 635) = 4.58$, $p = .101$, $\eta^2 = .007$.

In order to explore the differences and the directions of the effects within the three groups post-hoc Tukey-Kramer tests and the Mann-Whitney U tests with Bonferroni corrections are conducted. Only groups within with significant differences across the three groups in the omnibus tests are analyzed. Table 5-2 shows descriptive information on the variables and describes the results of the multi-group comparisons.

Table 5-2: Post-hoc multiple-group comparisons on independent variable showing omnibus significant differences across teacher groups: *worse-than-expected (0), as-expected (1), and better-than-expected (2).* ~p<.1, *p<.05, **p<.01, ***p<.001.

Variable	***Worse-than-expected***		***As-expected***		***Better-than-expected***		**0→1**	**1→2**	**0→2**
	Mean	(SD)	Mean	(SD)	Mean	(SD)	***p***	***p***	***p***
AP workload	-.524	(.653)	-.286	(.883)	-.201	(.913)	*	*n.s.*	**
Days of school year	267.68	(46.38)	276.39	(31.37)	279.34	(26.08)	*n.s.*	*n.s.*	*
District funding: All	8.40	(2.50)	8.97	(2.20)	9.46	(2.47)	*n.s.*	*n.s.*	*
Degree	2.86	(.69)	2.86	(.76)	3.09	(.73)	*n.s.*	*	*n.s.*
Courses	11.78	(7.79)	13.55	(7.48)	14.06	(8.17)	*	*n.s.*	~
Knowledge and experience	-.426	(.854)	-.392	(.803)	-.087	(.965)	*n.s.*	**	*
Labs	13.21	(5.32)	12.30	(5.57)	13.53	(5.30)	*n.s.*	*n.s.*	*n.s.*
Responsive agenda	2.98	(2.88)	3.16	(2.53)	3.66	(2.81)	*n.s.*	*n.s.*	~

Analyzing the differences in the school, teacher and teaching, and PD participation variables yield interesting insights. Unsurprisingly, the linear trends of all effects follow the intuitive logic suggesting that these variables might be impactful for explaining conditions for teachers that enable their students to perform on the AP science examinations on average better than the PSAT scores predict.

One the school level, post-hoc Whitney-Mann U tests with Bonferroni corrections reveal statistically significant higher scores on teachers' AP workload for teachers in the *better-than-expected* group (M = -.201, SD = .913), U = -3.13, p < .01, and the *as-expected* group (M = -.286, SD = .883), U = -2.46, p < .05, than teachers in the *lower-than-expected* group (M = -.524, SD = .653). Also, conducting post-hoc Whitney-Mann U tests with Bonferroni corrections indicate that the number of days in the school year is significantly higher for teachers in the *better-than-expected* group (M = 279.34, SD = 26.08) than for teachers in the *lower-than-expected* group (M = 267.68, SD = 46.38), U = -2.80, p < .05. Furthermore, Tukey-Kramer multiple-comparison tests indicate significantly higher total district level funding allocations to schools of teachers in the *better-than-expected* group (*M* = \$9,461, *SD* = \$2,472) than to schools of teachers in the *lower-than-expected* group (*M* = \$8,403, *SD* = \$2,503), *TK* = 4.81, *p* < .05.

On the teacher level, Tukey-Kramer multiple-comparison tests indicate significantly higher educational degree attainments for teachers in the *as-expected* group (*M* = 2.86, *SD* = .76) than for teachers in the *lower-than-expected* group (*M* = 2.86, *SD* = .69), *TK* = 4.38, *p* < .05. Conducting post-hoc Whitney-Mann U tests with Bonferroni corrections indicate that teachers attended significantly more disciplinary coursework in college and graduate in the *as-expected* group (*M* = 13.55, *SD* = 7.48) than teachers in the *lower-than-expected* group (*M* = 11.78, *SD* = 7.79), *U* = -2.70, *p* < .05. The comparison to teachers in the *better-than-expected* group (*M* = 14.06, *SD* = 8.17) is approaching significance, *U* = -2.90, *p* < .10. Additionally, teachers' knowledge and experience in the *better-than-expected* group (*M* = -.087, *SD* = .965) is significantly higher than for teachers in the *as-expected* group (*M* = -.302, *SD* = .803), *U* = -3.20, *p* < .01, and for teachers in the *lower-than-expected* group (*M* = -.426, *SD* = .854), *U* = -2.90, *p* < .05. Conducting Tukey-Kramer multiple-comparison tests indicate no significant differences of the number of laboratory investigation in teachers' AP science classes within the three groups.

Regarding teachers' PD participation, Whitney-Mann U tests with Bonferroni corrections indicate that teachers' combined rating of their whole PD experience as responsive to their interests and needs are approaching significance for higher teacher ratings in the *better-than-expected* group (*M* = 3.66, *SD* = 2.81) than in the *lower-than-expected* group (*M* = 2.98, *SD* = 2.88), *U* = -2.24, *p* < .10.

5.2 Relationships towards students' AP science performance

After exploring characteristics of the AP science teacher population, several features have been identified that might distinguish teachers teaching students performing on average lower than predicted compared to teachers teaching students performing on average higher than predicted. Those features derived from the exploration of the first research question are districts funding allocations, teachers' knowledge and experience, and teachers participating in PD activities which are responsive to their interests and needs, among others (cf. Section 5.1). The second research question seeks to explore the relationship between these teacher, teaching, teacher PD participation, and school characteristics on students' difference score between AP science and PSAT scores. The analysis includes the above characteristics, additional dichotomous single-indicator school characteristics (special education class offerings, charter school status, school neighborhood, and entry-criteria for AP course enrollment), and teachers ethnical background, controlling for student level covariates (GPA on all AP examinations, English language leaner status, ethnical background, and parents' educational level,).

Applying hierarchical linear modeling – with students on level one and teacher/school characteristics on level two – attempts to detect direct effects towards students' achievement measured by the difference between students' actual performance on the AP science exam and their projected score predicted by the PSAT exam. The proportion of the variance of students' difference between AP science and PSAT scores explained by the variance between schools is represented by the intraclass correlation coefficient (ICC). In this study, the ICC is .19 indicating that 19 % of the variance in students' performance measure is explained on the teacher/school level. The remaining 81% of the variance is at the student level. The small to medium ICC value indicates that students enrolled in one school demonstrate somewhat similar behavior due to their common exposure to teacher, teaching, and school characteristics Therefore, a multi-level modeling approach is more appropriate than an ordinary least square multiple nested regression approach. Our analysis strives to explain the 19% of the variance at the teacher/school level exploring the relations towards student performance in order to generate recommendation for educational policy makers and practitioners on how to improve at-risk students' achievement.

The data fulfills the assumptions of hierarchical linear modeling, as described in Appendix 8.3. Therefore, fixed-effect hierarchical linear models with robust standard errors applying a full maximum likelihood estimation method with 100 iterations are computed to explore direct effects on the difference between students' AP science and PSAT performance. Variables are gradually included to analyze the model progression on the percentage of explained school level variance. Table 5-3 describes the results of the hierarchical linear modeling with model 1 only including student level variables; model 2 adding school level variables; model 3 adding teacher and teaching variables; and model 4 additionally including variables on teachers' PD participation. Model 4 is referred to as the "full model."

Table 5-3: Fixed-effect hierarchical linear models. Model comparison tests are comparing gradual model progression from null model (not included) to the full model. White is the reference category for race/ethnicity series of dummy variables; Town is the reference category for the school neighborhood series of dummy variable. ~p<.1, *p<.05, **p<.01, ***p<.001.

	Model 1		***Model 2***		***Model 3***		**Model 4**	
AP performance leap	***β***	***(SE)***	***β***	***(SE)***	***β***	***(SE)***	***β***	***(SE)***
Level 1 (Student level)								
AP GPA	.181***	(.013)	.181***	(.013)	.178***	(.013)	.178***	(.013)
Parents' education level	-.035***	(.004)	-.034***	(.004)	-.035***	(.004)	-.035***	(.004)
English language learner	-.381***	(.061)	-.378***	(.061)	-.380***	(.061)	-.382***	(.061)
Black or African American	.154***	(.025)	.142***	(.026)	.147***	(.026)	.145***	(.026)
Asian or Asian American	.070*	(.028)	.064*	(.029)	.062*	(.028)	.062*	(.028)
Hispanic	.037	(.024)	.035	(.024)	.033	(.024)	.032	(.024)
American Indian or Alaska Native	.125	(.092)	.134	(.095)	.141	(.095)	.136	(.094)
Level 2 (Teacher and school level)								
Intercept	.170*	(.068)	-.345*	(.170)	-.243	(.188)	-.242	(.190)
School Characteristics								
District funding: All (in $1,000 increments)			.028***	(.007)	.027***	(.007)	.027***	(.007)
District funding: Materials (in $100 increments)			.016	(.014)	.016	(.013)	-.017	(.013)
Percentage of free- or reduced lunch students			-.054	(.123)	-.000	(.124)	.009	(.124)
Days of school year (in 10 day increments)			.010*	(.004)	.008*	(.004)	.007~	(.004)

Special education classes			.035	(.054)	-.002	(.057)	-.006	(.054)
Charter school			-.014	(.071)	-.038	(.078)	-.059	(.080)
Rural or non-metro			.007	(.045)	.026	(.045)	.030	(.046)
Suburban			.001	(.034)	.007	(.034)	.009	(.034)
Urban			-.077~	(.047)	-.072	(.045)	-.071	(.044)
Principal support			-.003	(.014)	-.008	(.013)	-.012	(.013)
AP workload			.015	(.016)	.008	(.017)	.004	(.016)
Criteria for AP enrollment			.096***	(.029)	.111***	(.030)	.118***	(.029)
Teacher Characteristics								
Age					-.003~	(.002)	-.002	(.002)
Gender					.032	(.033)	.030	(.035)
Degree					.024	(.020)	.019	(.020)
Courses					-.003	(.002)	-.003	(.002)
Knowledge and Experience					.072**	(.022)	.073***	(.022)
PD Inclination					-.016	(.014)	-.018	(.015)
Black or African American					-.007	(.049)	-.004	(.048)
Asian or Asian American					.081	(.069)	.084	(.069)
Hispanic					.149*	(.063)	.153*	(.064)
American Indian or Alaska Native					-.099	(.076)	-.076	(.075)
Teaching Characteristics								
Labs					.002	(.002)	.003	(.003)
Challenges with the AP redesign					-.020	(.013)	-.019	(.013)
Enactment AP practices					.011	(.014)	.010	(.014)
Enactment of AP curriculum					-.026~	(.015)	-.025~	(.015)
PD Activities								
Active learning							-.004	(.010)
Responsive agenda							.018~	(.011)
Focus on student work							-.003	(.011)
Modeling teaching							-.017	(.011)
Building relationships							-.013	(.011)
Effective support							.019~	(.011)
Unconventional PD: Face-to-face							.038~	(.020)
Unconventional PD: Materials							-.004	(.011)
Deviance	26,794.91		26,753.23		26,720.77		26,707.84	
Number of parameters	10		22		36		44	
Level 2 variance	.0998		.0965		.0841		.0818	
Explained variance (lvl 2)	21.13 %		28.35 %		33.50 %		35.36 %	
Δdf	7		12		14		9	
χ^2	566.15		41.68		32.46		12.93	
***p*-value**	<.001		<.001		.004		.114	

The analysis of the full model indicates three major findings that are consistent with the trend indicated through the explorations of the first research question: First, teacher characteristics help explaining variance in student performance on the AP science examinations in low-income schools beyond students predicted scores. Second, PD participation makes a direct contribution to student outcome. Third, the school context contributes to explaining variance in student outcome.

Analyzing the explained variance on the school/teacher level through the model progression indicates that adding both the school level ($\Delta\sigma^2$ = 7.22 %) variables, $\chi^2(12) = 41.68$, $p < .001$, and the teacher and teaching level ($\Delta\sigma^2$ = 5.15 %) variables, $\chi^2(14) = 432.46$, $p < .01$, significantly improve the model fit. Although adding teachers' PD activities to the model ($\Delta\sigma^2$ = 1.86 %) is only almost approaching significance, $\chi^2(9) = 12.93$, $p = .114$, the included variables itself indicate direct effects of PD participation on student achievement on the AP science examinations. Detecting direct effects for individual PD variables is rather surprising because the literature suggest that PD acts through mediating processes on student achievement. For instance, Desimone's (2009) logic model of PD suggests that PD participation influences teacher characteristics such as knowledge, skills, attitudes, and beliefs, which then improves both instruction and pedagogy resulting in instructional changes that increase student learning.

Analyzing the direct effects of the independent variables in the full model assumes that a variable varies with all others held constant. Since this study focuses on the teacher and school level, effects of student level variables, included as covariates in the statistical models, on students' achievement on the AP science examinations are not discussed.

On the school level, districts total funding allocations are significantly associated with an increase in student achievement, $b = .027$, $t(603) = 3.92$, $p < .001$, indicating that for every additional $1,000 per student funding students' performance significantly increases on average .027 of an AP score compared to students' PSAT score prediction. Interestingly, directly increasing school budgets for instructional materials doesn't yield significantly improved student performance, *b* = -.017, *t*(603) = -1.30, *p* = n.s. This responds to the challenge of low-income schools to generate sufficient funds to fully equip classrooms and science laboratories, to reduce teacher-student ratios, and to recruit and retain experienced and qualified teachers (e.g., Biddle & Berliner, 2003; Elliott, 1998; P. T. Hill et al., 2003). Additionally, increasing the days of the school year is approaching significance for being related to improved student achievement, with an average .007 AP score increase in comparison to students' PSAT scores for every ten days of the school year, *b* = .007, *t*(603) = 1.82, *p* < .10. Furthermore, enforcing an enrollment criteria for student enrollment in the AP courses is significantly associated with an average increase of .118 of an AP score on students' performance on the AP science exam above their predicted score, *b* = .118, *t*(603) = 3.99, *p* < .001. However, although increasing selectivity in AP course admission has an impact on student performance, presumably through enrolling more able students and more homogeneous course structure, this remains a questionable approach regarding educational equity. This selection process restricts some students to receive meaningful AP instruction through reduced access to AP courses and examinations – which is contrary to College Board's current efforts to increase AP participation for all students (e.g., Conger et al., 2009; Lichten, 2010; The College Board, 2014c; Wyatt & Mattern, 2011).

On the teacher level, increased teachers' knowledge and experience is significantly associated with student achievement on the AP science exam above students' AP score prediction, *b* = .073, *t*(603) = 3.34, *p* < .001. Roughly an increase of one standard deviation on the teachers' knowledge and experience composite variable corresponds with an average .073 of an AP score improvement compared to students' PSAT performance. This is an especially striking finding underlining the challenge for low-income schools to recruit and retain highly qualified teachers (cf. Section 3.2; (Biddle & Berliner, 2003; Elliott, 1998; P. T. Hill et al., 2003). Analyzing the effects of teachers' ethnic backgrounds leads to the conclusion that all but one ethnical background, Hispanic, does not significantly differ from White teachers. Teachers' self-identifying themselves as Hispanics are teaching students significantly scoring on average .153 on the AP exam higher than predicted, $b = .153$, $t(603) = 2.39$, $p < .05$. Future analysis should explore interaction effects of teachers' ethnic background and the ethnic make-up of teachers' student population in AP science courses or individual students' ethnic background.

Regarding teachers' AP classroom instruction, an increased self-reported enactment of the curriculum of the AP redesign is approaching significance with a negative association towards students' performance on the AP science exam compared to students' PSAT scores, $b = -.025$, $t(603) = -1.67$, $p < .10$. An increase of approximately one standard deviation of teachers' rating on curricular enactments of the AP redesign is associated with an average .025 AP score decrease compared to students' predicted score. Potential explanations include that teachers might hold different perceptions on what enactment of the redesigned AP science curriculum constitutes. If their perceptions don't coincide with the College Board's conceptions, teachers with high ratings on curricular enactment might think that are preparing their students for the AP science examinations, whereas if this assumption doesn't hold this results in underprepared students who score worse than expected on the AP science exams. Another explanation might be that students with high PSAT scores know how to "game the system." Being prepared differently for a revised high-stakes examination focusing on inquiry learning and science practices instead of algorithmic and rote learning components and a curriculum focusing on depth of learning instead of a broad coverage on subjects (e.g., Domyancich, 2014; Magrogan, 2014; Price & Kugel, 2014) students can hardly shift their learning and thinking towards new curricular structures resulting in dropped AP science performance below the projected scores by their PSAT examination.

Teachers' PD participation is approaching significance for direct effects towards improving students' performance on the AP science exams beyond their predictions by the PSAT scores. Teachers' combined ratings on the responsiveness of PDs agenda to their interests and needs are positively associated with students' AP performance compared to their PSAT performance, $b = .018$, $t(603) = 1.66$, $p < .10$. Every point increase in teachers' rating (on a 0-4 scale with 0: almost completely fixed agenda to 4: almost completely responsive agenda) of a single PD activity teachers participate in increase students' average AP science performance compared to their PSAT performance by .018 of an AP score. For instance, if a teacher participates in only two PD activities with perceived maximum ratings of agenda responsiveness, students' average AP science performance improves by .144 of an AP score compared to their predicted scores. Teachers' combined ratings on the perceived effectiveness of their combined PD experiences supporting them with teaching the revised AP science course are positively associated with the difference of students' actual and projected performance on the AP science exams, $b = .019$, $t(603) = 1.73$, $p < .10$. Every point increase in teachers' rating (on a 0-4 scale with 0: not effective supported to 4: extremely effective supported) of a single PD activity teachers participate in corresponds with an average .019 higher AP science score than students' PSAT score predict. For instance, if a teacher participates in only two PD activities with perceived maximum ratings of effective support, students' average AP science performance improves by .152 of an AP score compared to their predicted scores. Furthermore, for every participation in an additional unconventional face-to-face PD activity (district, regional, local college, or teacher-initiated meeting; mentoring or coaching; conference participation) is associated with an average increase of .038 of an AP score compared to students' projected performance, b = .038, t(603) = 1.86, p < .10. This is consistent with the trend of prior research on features of effective PD, highlighting the responsiveness of the agenda, providing effective support, and the duration/quantity of PD activities teachers engage in, among others features, as important design characteristics that are associated with fostering teacher learning and student achievement (e.g., Desimone et al., 2002; Fishman, Marx, Best, & Tal, 2003; Garet et al., 2001; Penuel et al., 2007).

6 Discussion

6.1 Conclusion and scholarly significance

This study analyzes the relationship of school, teaching, and teacher characteristics, as well as teachers' PD participation on student achievement on the redesigned AP science examination within the context of low-income schools, controlling for student characteristics. The driving motivation of this study is to elicit recommendations to inform educational policy makers and practitioners, guide decision-making processes attempting to narrow the income achievement gap, and contribute to the shared striving for educational equity.

The three main findings of this study are the following: First, teachers' knowledge and experience is related to students' performance on the AP science examinations in low-income schools beyond students predicted scores. Second, PD participation is directly related to student outcome. Third, contextual variables on the school level are directly related to student achievement. More explorations in order to give recommendations for educational policy makers and practitioners should be pursued in the following areas:

- Teachers' knowledge and experience indicate a positive significant association with students' success on the AP science examinations. Therefore, incentivizing experienced and skilled teachers to be recruited and retained within low-income schools could be further explored.
- Certain types of PD activities indicate direct associations with student achievement on the AP science examinations. Therefore, teacher participation in PD activities with agendas responsive to teachers' interests and needs as well as a perceived effective support towards teaching the revised AP science courses could be further explored.
- Participation in unconventional face-to-face PD activities indicate direct associations with student achievement on the AP science examinations. Supporting and encouraging teachers to participate in district, regional, local college, or teacher-initiated meetings; in mentoring or coaching activities with other teachers; and conferences or conference sessions could be further explored.
- Districts per-student total funding allocations indicate a positive significant association to students' performance on the AP science examinations. Therefore, increasing districts total expenditures per students in low-income schools could be further explored.

This study is an important contribution to the research base it is a large-scale study with good representation of the AP science teacher population in general. Additional, mandated top-down science curriculum and assessment change also constitutes an unique opportunity for research. Teachers are forced to adopt to the changing landscape of their AP science courses due to the nature of these large-scale redesigns instead of repeating their usual “teaching to the test.” Therefore, modifying Opfer and Pedders' (2011) *dynamic model of teaching learning and change*, this study gains insights on the associations of PD participation and other school, teacher, and teaching characteristics with student achievement. The approach of evaluating students’ predicted performance with their actual achievement for identifying “what works” for teachers to aid their students represents an advance on prior research. Conjectures from this study indicate that PD and proactive educational policies can make a difference in challenge of striving for educational equity and assisting at-risk students to succeed on their path through the U.S. education system.

6.2 Limitations and future work

This study has limitations both within the data sets and through the statistical methods applied. Limitations within the data sets posing threats to internal validity include that teachers’ survey data is self-reported, that some variables have ceiling effects through too low cutoff values, and some “check all that apply” survey items are coded as zero instead of missing. Also, generally combining the Biology and Chemistry data for the analysis without controlling for the discipline might introduce some bias. Threats to external validity include the absence of identifiers between the student and teacher level such that this study only analyzes schools with only one AP science teacher in the discipline. Also student identifiers are unique for AP Biology and AP Chemistry such that if the same student is taking AP science examinations in May 2014, this student will be treated as two separate cases which introduces some oversampling. This leads selection bias that reduces the overall generalizability of the implications. Another potential threat is that whereas the 2014 AP Chemistry data is looking at the first year of implementing the AP Chemistry redesign, the 2014 AP Biology data consists of data of the second year of the implementation of the AP Biology redesign.

Methodologically, the hierarchical linear modeling approach used assumes linear relationships between independent and dependent variables and detects direct effects. However, some relations might be better described through polynomial, exponential, or other relationships. Additionally, interaction and mediating effects might take place such that independent variables might have indirect, dynamic relationships towards student achievement. Therefore, future statistical approaches might add interaction effects to the hierarchical linear models and extend the statistical analyses to multi-level structure equation models and path analysis in order to explore mediating effects.

Generally, this study is part of a larger research project longitudinally exploring the large-scale changes in science education induced by the AP redesign. As part of this three-year project, longitudinal effects of the implementation of the AP science redesign could be explored. Combining these analytical approaches, the guiding vision of the research team ultimately aims for changes in the educational landscape that increase overall student learning and achievement and consequently – in Mandela's sense – change the world.

7 References

Abdi, H. (2003). Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA*, 792–795.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*(1), 1–28. http://doi.org/10.3102/00346543074001001

Abell, S. K. (2007). Research on science teacher knowledge. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of Research in Science Education* (pp. 1105–1149). Mahwah, NJ: Lawrence Erlbaum Associates.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*(6045), 1034–1037.

Autor, D. H., Katz, L. F., & Kearney, M. S. (2008). Trends in U.S. wage inequality: Re-assessing the revisionists. *The Review of Economics and Statistics*, *90*(2), 300–323.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes It special? *Journal of Teacher Education*, *59*(5), 389–407. http://doi.org/10.1177/0022487108324554

Banilower, E. R., Heck, D. J., & Weiss, I. R. (2007). Can professional development make the vision of the standards a reality? The impact of the National Science Foundation's local systemic change through teacher enhancement initiative. *Journal of Research in Science Teaching*, *44*(3), 375–395. http://doi.org/10.1002/tea.20145

Baum, S., Ma, J., & Payea, K. (2013). *Education pays 2013 - The benefits of higher education for individuals and society*. New York, NY: The College Board.

Biddle, B. J., & Berliner, D. C. (2003). *What research says about unequal funding for schools in America*. San Francisco, CA: WestEd.

Birman, B. F., Desimone, L., Porter, A. C., & Garet, M. S. (2000). Designing professional development that works. *Educational Leadership*, *57*(8), 28–33.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3–15.

Borko, H., Jacobs, J., & Koellner, K. (2010). Contemporary approaches to teacher professional development. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 548–556).

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2011). How neighborhoods influence NYC teachers careers. In G. J. Duncan & R. J. Murnane (Eds.), *Whither Opportunity* (pp. 377–396). New York, NY: Russell Sage Foundation.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

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

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*. http://doi.org/10.3102/0034654314532697

Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. *Educational Evaluation and Policy Analysis*, *23*(2), 145–170.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, *33*, 107–112.

Coley, R. J. (2002). *An uneven start: Indicators of inequality in school readiness. Policy information report.* Princeton, NJ: Educational Testing Service.

Conger, D., Long, M. C., & Iatarola, P. (2009). Explaining race, poverty, and gender disparities in advanced course-taking. *Journal of Policy Analysis and Management*, *28*(4), 555–576. http://doi.org/10.1002/pam.20455

Coolidge, F. L. (2000). *Statistics. A gentle introduction*. London, UK: Sage Publications Ltd.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1–9.

Cuban, L., Kirkpatrick, H., & Peck, C. (2001). High access and low use of technologies in high school classrooms: Explaining an apparent paradox. *American Educational Research Journal*, *38*(4), 813–834.

Desimone, L. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199. http://doi.org/10.3102/0013189X08331140

Desimone, L., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, *24*(2), 81–112.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1–11.

Domyancich, J. M. (2014). The development of multiple-choice items consistent with the AP Chemistry curriculum framework to more accurately assess deeper understanding. *Journal of Chemical Education*, *91*(9), 1347–1351. http://doi.org/10.1021/ed5000185

Dougherty, C., Mellor, L. T., & Jian, S. (2006). *The relationship between Advanced Placement and college graduation* (No. 2005 AP Study Series, Report 1). Austin, TX: The National Center for Educational Accountability.

Downie, N. M., & Heath, R. W. (1983). *Basic statistical methods* (5th ed.). New York, NY: Harper and Row Publishers, Inc.

Duncan, G. J., & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity* (pp. 47–69). New York, NY: Russell Sage Foundation.

Elliott, M. (1998). School finance and opportunities to learn: Does money well spent enhance students' achievement? *Sociology of Education*, *71*(3), 223. http://doi.org/10.2307/2673203

Ewing, M., Camara, W. J., & Millsap, R. E. (2006). *The relationship between PSAT/NMSQT scores and AP examination grades: A follow-up study*. New York: The College Board.

Field, A. P. (2009). *Discovering statistics using SPSS* (3rd ed.). London, UK: Sage Publications Ltd.

Fishman, B., Konstantopoulos, S., Kubitskey, B. W., Vath, R., Park, G., Johnson, H., & Edelson, D. C. (2013). Comparing the impact of online and face-to-face professional development in the context of curriculum implementation. *Journal of Teacher Education*, *64*(5), 426–438.

Fishman, B., Marx, R. W., Best, S., & Tal, R. T. (2003). Linking teacher and student learning to improve professional development in systemic reform. *Teaching and Teacher Education*, *19*(6), 643–658. http://doi.org/10.1016/S0742-051X(03)00059-3

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.

Geiser, S., & Santelices, V. (2006). The role of Advanced Placement and honors courses in college admissions. In P. Gandara, G. Orfield, & C. L. Horn (Eds.), *Expanding opportunity in higher education: Leveraging promise* (pp. 75–114). Albany, NY: SUNY Press.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(1), 549–576. http://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206–213. http://doi.org/10.1007/s11121-007-0070-9

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450. http://doi.org/10.1037//1082-989X.6.4.430

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149–161.

Hallett, R. E., & Venegas, K. M. (2011). Is increased access enough? Advanced Placement courses, quality, and success in low-income urban schools. *Journal for the Education of the Gifted*, *34*(3), 468–487. http://doi.org/10.1177/016235321103400305

Hargrove, L., Godin, D., & Dodd, B. (2008). *College outcomes comparisons by AP and non-AP high school experiences*. New York, NY: The College Board.

Harrington, D. (2009). *Confirmatory factor analysis*. New York, NY: Oxford University Press.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*(3), 333–362. http://doi.org/10.1002/tea.21004

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads, and challenges. *Educational Researcher*, *42*(9), 476–487. http://doi.org/10.3102/0013189X13512674

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Hill, P. T., Guin, K., & Celio, M. B. (2003). Minority children at risk. In P. E. Peterson (Ed.), *Our Schools and Our Future...Are We Still at Risk?* (pp. 111–139). Palo Alto, CA: Hoover Institution Press.

Ingersoll, R. M. (1999). The problem of underqualified teachers in American secondary schools. *Educational Researcher*, *28*(2), 26–37.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*(3), 499–534.

Kaiser, H. F. (1961). A note on Guttman's lower bound for the number of common factors. *The British Journal of Statistical Psychology*, *14*, 1–2.

Kirk, R. E. (1998). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Monterey, CA: Brooks/Cole Publishing.

Klopfenstein, K. (2004). Advanced Placement: Do minorities have equal opportunity? *Economics of Education Review*, *23*(2), 115–131.

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

Levine, T. R., & Hullett, C., R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*(4), 612–625.

Lichten, W. (2010). Whither Advanced Placement -- Now? In P. M. Sadler, G. Sonnert, R. H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program* (pp. 233–243). Cambridge, MA: Harvard Education Press.

Lichten, W., & Wainer, H. (2000). The aptitude–achievement function: An aid for allocating educational resources, with an Advanced Placement example. *Educational Psychology Review*, *12*(2), 201–228.

Lorenzo-Seva, U., Kiers, H. A., & Berge, J. M. (2002). Techniques for oblique factor rotation of two or more loading matrices to a mixture of simple structure and optimal agreement. *British Journal of Mathematical and Statistical Psychology*, *55*(2), 337–360.

Magrogan, S. (2014). Past, present, and future of AP Chemistry: A brief history of course and exam alignment efforts. *Journal of Chemical Education*, *91*(9), 1357–1361. http://doi.org/10.1021/ed500096f

Mattern, K. D., Marini, J. P., & Shaw, E. J. (2013). *Are AP students more likely to graduate from college on time?* (No. Research Report 2013-5). New York, NY: The College Board.

Murnane, R. J., Willett, J. B., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *The Review of Economics and Statistics*, *77*(2), 251. http://doi.org/10.2307/2109863

National Research Council. (2002). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools*. Washington, DC: National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: Achieve, Inc. On behalf of the twenty-six states and partners that collaborated on the NGSS.

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*(6), 778–803. http://doi.org/10.1002/tea.21026

OECD. (2013a). *Country note. Results from PISA 2012. United States*. Paris: OECD Publishing.

OECD. (2013b). *PISA 2012 results: What students know and can do - Student performance in mathematics, reading and science (Volume I)*. Paris: OECD Publishing.

Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, *81*(3), 376–407. http://doi.org/10.3102/0034654311413609

Patterson, B. F., Kobrin, J. L., & Packman, S. (2011). *Advanced Placement exam-taking and performance: Relationships with first-year subject area college grades*. New York, NY: The College Board.

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, *340*(6130), 320–323. http://doi.org/10.1126/science.1232065

Penuel, W. R., Fishman, B., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, *44*(4), 921–958.

Price, P. D., & Kugel, R. W. (2014). The new AP Chemistry exam: Its rationale, content, and scoring. *Journal of Chemical Education*, *91*(9), 1340–1346. http://doi.org/10.1021/ed500034t

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

Raudenbusch, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither Opportunity*, 91–116.

Reardon, S. F. (2013). The widening income achievement gap. *Educational Leadership*, *70*(8), 10–16.

Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. Z. (2011). Videobased lesson analysis: Effective science PD for teacher and student learning. *Journal of Research in Science Teaching*, *48*(2), 117–148.

Schneider, J. (2009). Privilege, equity, and the Advanced Placement program: Tug of war. *Journal of Curriculum Studies*, *41*(6), 813–831. http://doi.org/10.1080/00220270802713613

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

Shaw, E. J., Marini, J. P., & Mattern, K. D. (2013). Exploring the utility of Advanced Placement participation and performance in college admission decisions. *Educational and Psychological Measurement*, *73*(2), 229–253. http://doi.org/10.1177/0013164412454291

Shih, T.-H., & Fan, X. (2009). Comparing response rates in e-mail and paper surveys: A meta-analysis. *Educational Research Review*, *4*(1), 26–40. http://doi.org/10.1016/j.edurev.2008.01.003

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575.

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher*, *37*(4), 189–199. http://doi.org/10.3102/0013189X08319569

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553–573.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York, NY: Routledge.

The College Board. (2012). *AP Biology. Course and exam description*. New York, NY: The College Board.

The College Board. (2014a). *AP Chemistry. Course and exam description*. New York, NY: The College Board.

The College Board. (2014b). *AP Physics 1: Algebra-based and AP Physics 2: Algebra-based. Course and exam description*. New York, NY: The College Board.

The College Board. (2014c). *The 10th annual AP report to the nation*. New York, NY: The College Board.

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. *Review of Research in Education*, *24*(1), 173–209.

Wyatt, J. N., & Mattern, K. D. (2011). *Low-SES students and college outcomes: The role of AP fee reductions*. New York, NY: The College Board.

Yaron, D. J. (2014). Reflections on the curriculum framework underpinning the redesigned advanced placement chemistry course. *Journal of Chemical Education*, *91*(9), 1276–1279. http://doi.org/10.1021/ed500103e

Zarate, M. E., & Pachon, H. P. (2006). *Gaining or losing ground? Equity in offering Advanced Placement courses in California high schools 1997-2003*. Los Angeles, CA: Tomas Rivera Policy Institute.

8 Appendix

8.1 Missing data tables

This section describes missing data that is being imputed through the missing data approaches. For simplicity, only variables used in the statistical models for answering the research questions are described. However, the imputation processes use all variables available to the research team to increase precision. Table A1 describes the College Board provided student data and table A2 the College Board provided school data. Table A3 describes the missing data on the survey-based single-indicator variables and table A4 the missing data on the variables used for creating the composite variable based on the factor analysis.

Table A1: Missing data table on the un-imputed, full student data set combining both 2014 AP Biology and 2014 AP Chemistry data sets. N=313,649.

Variable	N	Missing	Missing [%]
PSAT scores	270,160	60,371	18.26%
AP science scores	330,531	0	0.00%
AP GPA	330,531	0	0.00%
Parents' education level	320,314	10,217	3.09%
English language learner	326,919	3,612	1.09%
Black or African American	313,649	16,882	5.11%
Asian or Asian American	313,649	16,882	5.11%
Hispanic	313,649	16,882	5.11%
American Indian or Alaska Native	313,649	16,882	5.11%
English language learner	313,649	16,882	5.11%

Table A2: Missing data table on the un-imputed, full College Board provided school data combining both 2014 AP Biology and 2014 AP Chemistry data sets. N=9,801.

Variable	N	Missing	Missing [%]
District funding: All	7,780	2,021	20.62%
District funding: Materials	7,780	2,021	20.62%
Percentage of free- and reduced lunch program	9,489	312	3.18%
Special education classes	9,594	207	2.11%
Charter school	7,956	1,845	18.82%
Rural or non-metro	9,087	714	7.28%
Suburban	9,087	714	7.28%
Urban	9,087	714	7.28%
Town	9,087	714	7.28%

Table A3: Missing data table of the survey-based single-indicator independent variables. $N_{Biology}$= 2,408, $N_{Chemistry}$= 2,493.

Variable	2014 AP Biology			2014 AP Chemistry		
	N	Missing	Missing [%]	N	Missing	Missing [%]
Criteria for AP enrollment	2377	31	1.30%	2454	39	1.56%
Age	2274	134	5.56%	2359	134	5.38%
Gender	2294	114	4.73%	2383	110	4.41%
Courses	2391	17	0.71%	2477	16	0.64%
White	2293	115	4.78%	2383	110	4.41%
Black or African American	2293	115	4.78%	2383	110	4.41%
Asian or Asian America	2293	115	4.78%	2383	110	4.41%
Hispanic	2293	115	4.78%	2383	110	4.41%
American Indian or Alaska Native	2293	115	4.78%	2383	110	4.41%
Labs	2351	57	2.37%	2412	81	3.25%

Table A4: Missing data table of the survey-based individual items used for the composite variables. $N_{Biology}$= 2,408, $N_{Chemistry}$= 2,493.

Variable	2014 AP Biology			2014 AP Chemistry		
	N	Missing	Missing [%]	N	Missing	Missing [%]
Principal support						
Principal understands students' challenges	2327	81	3.48%	2391	102	4.09%
Principal understands teachers' challenges	2323	85	3.66%	2388	105	4.21%
Principal supports PD	2320	88	3.79%	2382	111	4.45%
Lighter teaching load for AP science teachers	2326	82	3.53%	2389	104	4.17%
Fewer-out-of class responsibilities for AP science teachers	2327	81	3.48%	2390	103	4.13%
AP science is given funding exclusively for the course	2323	85	3.66%	2390	103	4.13%
Availability of equipment to perform labs	2331	77	3.30%	2389	104	4.17%
Availability of consumable supplies to perform labs	2331	77	3.30%	2386	107	4.29%
AP workload						
Number of students across AP Biology/Chemistry sections	2373	35	1.47%	2450	43	1.72%
Number of Biology/Chemistry sections	2381	27	1.13%	2457	36	1.44%
Number of preps each week	2385	23	0.96%	2455	38	1.52%
Teachers' knowledge and experience						
Years teaching high school science	2406	2	0.08%	2488	5	0.20%
Years teaching AP Biology/	2401	7	0.29%	2474	19	0.76%

Chemistry						
Number of professional science teaching organizations	2275	133	5.85%	2334	159	6.38%
Number of conference attendances	2337	71	3.04%	2426	67	2.69%
Years AP Reader	2370	38	1.60%	2432	61	2.45%
Years AP Consultant	2368	40	1.69%	2438	55	2.21%
Time point being assigned to teach AP science	2323	85	3.66%	2385	108	4.33%
PD inclination						
Importance PD to instructional performance	2364	44	1.86%	2453	40	1.60%
Importance PD to student performance	2354	54	2.29%	2441	52	2.09%
Self-teaching is as effective as formal PD participation	2364	44	1.86%	2455	38	1.52%
Teaching performance not greatly affected by PD participation	2354	54	2.29%	2453	40	1.60%
Enjoyment of participation in face-to-face PD activities	2364	44	1.86%	2452	41	1.64%
Challenges with the AP redesign						
Biology/Chemistry content	2327	81	3.48%	2369	124	4.97%
Organization of Biology/Chemistry content	2335	73	3.13%	2410	83	3.33%
Labs	2345	63	2.69%	2416	77	3.09%
Inquiry labs	2343	65	2.77%	2408	85	3.41%
Formats of questions/problems/exam	2343	65	2.77%	2419	74	2.97%
Application of science practices to the content	2344	64	2.73%	2419	74	2.97%
Development of new syllabus	2349	59	2.51%	2420	73	2.93%
Understanding the “exclusion statement”	2339	69	2.95%	2416	77	3.09%
Designing new student assessments	2341	67	2.86%	2418	75	3.01%
Using the text appropriately	2344	64	2.73%	2417	76	3.05%
Working with a new or different textbook	2263	145	6.41%	2248	245	9.83%
Pacing of the course	2347	61	2.60%	2414	79	3.17%
Moving students to a conceptual understanding	2346	62	2.64%	2414	79	3.17%
Enactment of AP redesign practices						
Students work on lab investigations	2344	64	2.73%	2413	80	3.21%
Guidance on content integration test questions	2348	60	2.56%	2416	77	3.09%
Guidance on open/free response	2347	61	2.60%	2418	75	3.01%

test questions						
Students reporting lab findings to other students	2343	65	2.77%	2418	75	3.01%
Students work on inquiry lab investigations (Chemistry only)				2412	81	3.25%
Enactment of AP redesign curriculum						
Refer to the "Big Ideas" of Biology/Chemistry	2348	60	2.56%	2418	75	3.01%
Use science practice outside of the classroom	2321	87	3.75%	2377	116	4.65%
Referring how "enduring understandings" relate to "Big Ideas"	2340	68	2.91%	2410	83	3.33%
Refer to learning objective from AP curriculum	2344	64	2.73%	2417	76	3.05%
Refer to curriculum framework	2319	89	3.84%	2414	79	3.17%

8.2 Composite variables

This section illustrates the descriptive information on the composite variables on the school, teacher, and teaching level based on the results of the exploratory and confirmatory factor analysis. The composite variables are separately created for both the 2014 AP Biology and 2014 AP Chemistry survey. The figures in the subsequent section include both a normal distribution and a fitted polynomial function. With the exception of the *PD inclination* (.10 bin size) composite variable (cf. figure A-4), figures A-1 to A-7 (except A-4) show the distribution

Table A5: Descriptive information on the composite variables. $N_{Biology}$= 2,408, $N_{Chemistry}$= 2,493.

Variable	2014 AP Biology			2014 AP Chemistry		
	Mean	SD	Variance	Mean	SD	Variance
Principal support	.00	1.09	1.19	.00	1.09	1.18
AP workload	.00	1.19	1.41	.00	1.22	1.50
Knowledge and experience	.00	1.16	1.36	.00	1.17	1.37
PD inclination	.00	1.09	1.19	.00	1.08	1.16
Challenges with the AP redesign	.00	1.07	1.14	.00	1.06	1.13
Enactment of AP practices	.00	1.22	1.50	.00	1.17	1.37
Enactment of AP curriculum	.00	1.10	1.22	.00	1.09	1.18

with a .25 bin size. Table A5 displays information on each composite variable and tables A6-A13 describe included variables, rotated factor loadings, uniqueness, and scoring coefficients.

Table A6: *Principal support* composite variable.

	2014 AP Biology			2014 AP Chemistry		
Included items:	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***
Principal understands students' challenges	.802	.357	.436	.810	.344	.443
Principal understands teachers' challenges	.817	.332	.478	.817	.333	.462
Principal supports PD	.561	.685	.159	.566	.679	.157
Lighter teaching load for AP science teachers	.326	.894	.071	.341	.884	.073
Fewer-out-of class responsibilities for AP science teachers	.287	.918	.061	.265	.930	.054
AP science is given funding exclusively for the course	.357	.872	.079	.349	.878	.075
Availability of equipment to perform labs	.444	.803	.107	.480	.770	.117
Availability of consumable supplies to perform labs	.466	.783	.116	.469	.780	.113

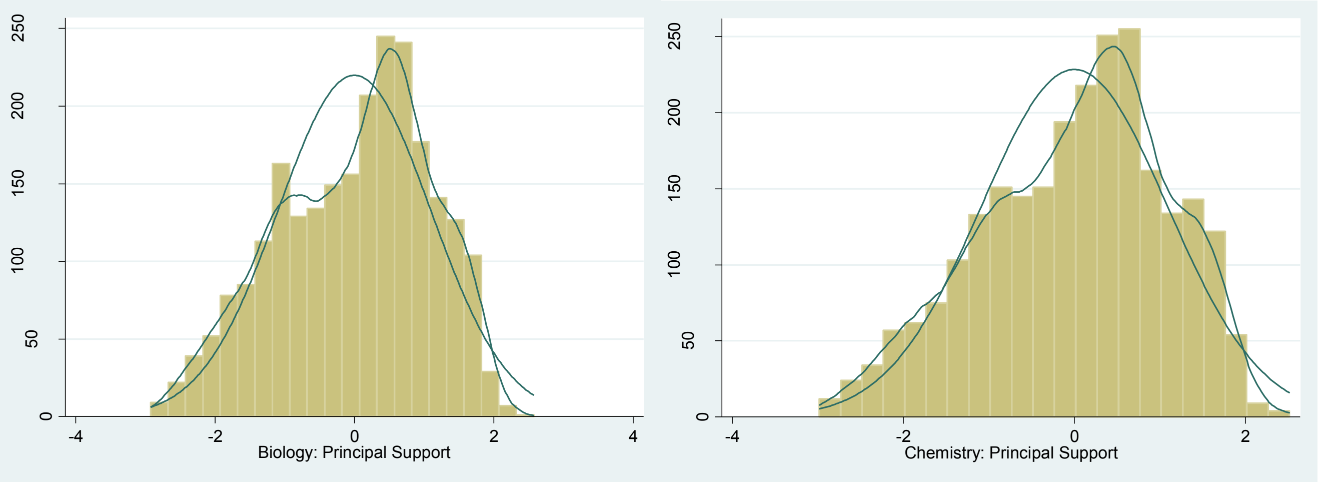

Figure A-1: Distribution of the *principal support* composite variable. (left) 2014 AP Biology data, (right) 2014 AP Chemistry data.

Table A7: *AP workload* composite variable.

	2014 AP Biology			2014 AP Chemistry		
Included items	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***
Number of students across AP Biology/ Chemistry sections	.781	.390	.671	.743	.447	.678
Number of Biology/Chemistry sections	.738	.455	.543	.716	.488	.599
Number of preps each week	-.429	.816	-.176	-.376	.859	-.179

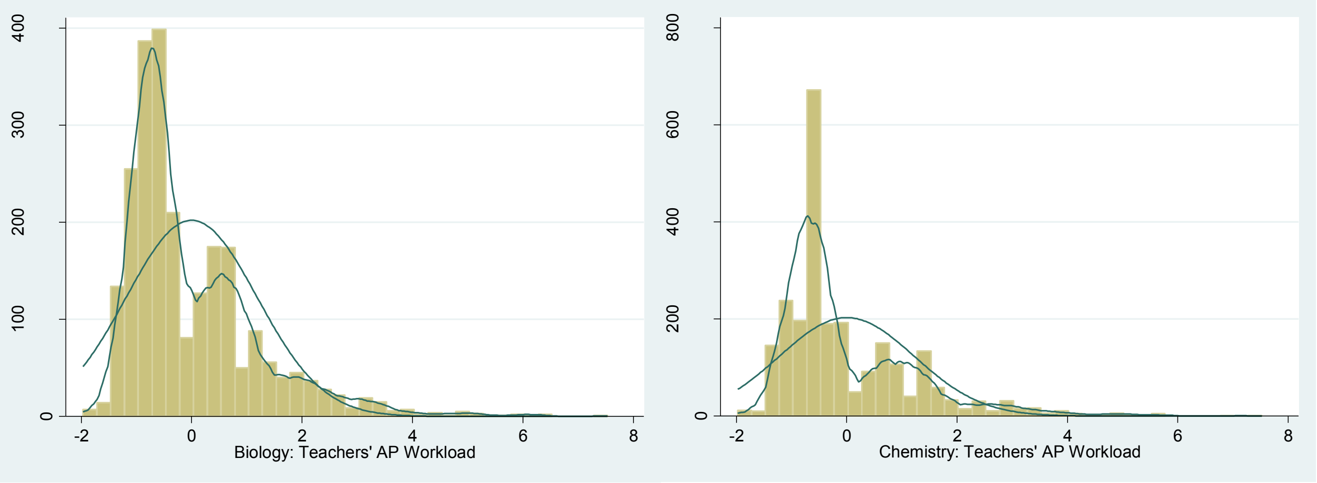

Figure A-2: Distribution of the *AP workload* composite variable. (left) 2014 AP Biology data, (right) 2014 AP Chemistry data.

Table A8: *Teachers' knowledge and experience* composite variable.

	2014 AP Biology			**2014 AP Chemistry**		
Included items	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***
Years teaching high school science	.687	.528	.417	.652	.574	.407
Years teaching AP Biology/Chemistry	.772	.404	.611	.742	.449	.592
Number of professional science teaching organizations	.340	.884	.123	.333	.889	.134
Number of conference attendances	.228	.948	.077	.293	.914	.115
Years AP Reader	.459	.790	.186	.480	.770	.223
Years AP Consultant	.369	.864	.137	.421	.823	.183
Time point being assigned to teach AP science	-.356	.873	-.131	-.288	.917	-.113

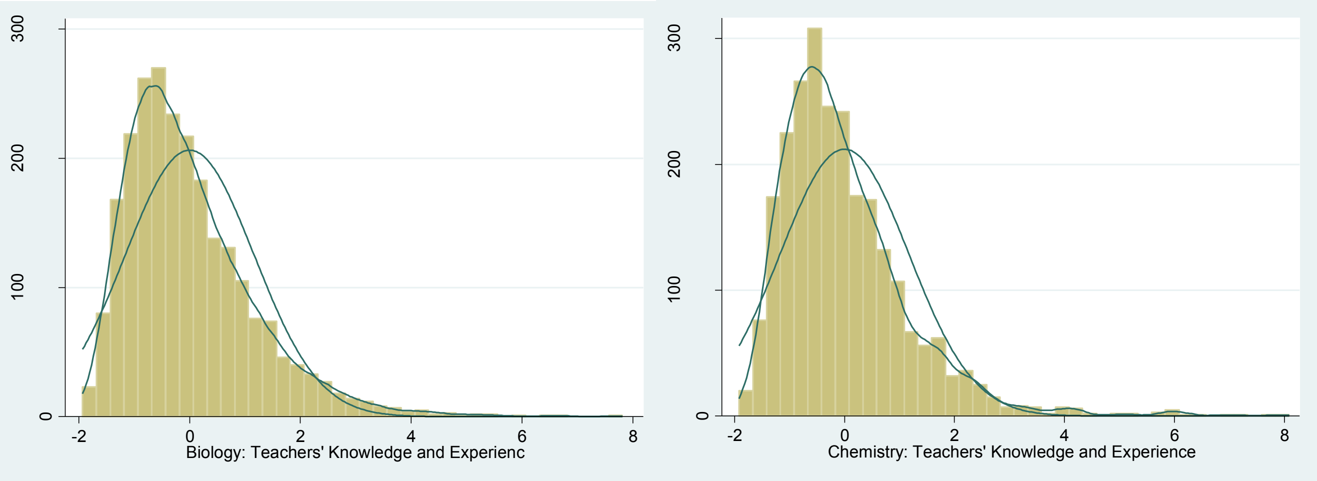

Figure A-3: Distribution of the *Teachers' knowledge and experience* composite variable. (left) 2014 AP Biology data, (right) 2014 AP Chemistry data.

Table A9: *Teachers' PD inclination* composite variable.

	2014 AP Biology			**2014 AP Chemistry**		
Included items	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***
Importance PD to instructional performance	.835	.302	.490	.859	.263	.467
Importance PD to student performance	.788	.379	.368	-.834	.304	-.392
Self-teaching is as effective as formal PD participation	-.522	.727	-.127	-.542	.707	-.109
Teaching performance not greatly affected by PD participation	-.696	.515	-.240	.717	.486	.211
Enjoyment of participation in face-to-face PD activities	.526	.724	.129	.549	.699	.112

Table A10: *Challenges with the AP redesign* composite variable.

	2014 AP Biology			**2014 AP Chemistry**		
Included items: ***Challenges with…***	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***
Biology/Chemistry content	.531	.719	.107	.621	.614	.136
Organization of Biology/Chemistry content	.604	.635	.138	.659	.565	.157
Labs	.594	.647	.134	.621	.615	.136
Inquiry labs	.593	.648	.133	.600	.640	.126
Formats of questions/problems/exam	.578	.666	.126	.571	.674	.114
Application of science practices to the content	.674	.546	.180	.666	.557	.161
Development of new syllabus	.541	.707	.111	.453	.795	.077
Understanding the "exclusion statement"	.625	.609	.149	.454	.794	.077
Designing new student assessments	.593	.648	.133	.627	.606	.139
Using the text appropriately	.601	.638	.137	.652	.574	.153
Working with a new or different textbook	.389	.849	.067	.421	.823	.069
Pacing of the course	.547	.701	.113	.645	.584	.149
Moving students to a conceptual understanding	.645	.584	.161	.650	.578	.151

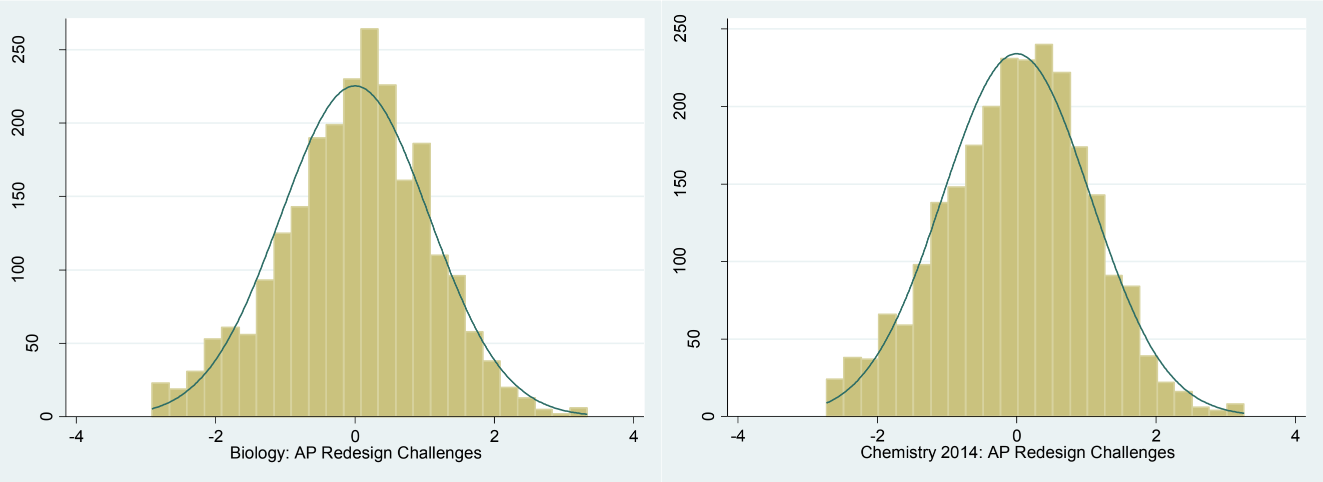

Figure A-5: Distribution of the *challenges with the AP Redesign* composite variable. (left) 2014 AP Biology data, (right) 2014 AP Chemistry data.

Table A11: *Enactment of AP redesign practices* composite variable.

	2014 AP Biology			**2014 AP Chemistry**		
Included items:	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***
Students work on lab investigations	.312	.903	.151	.417	.826	.196
Guidance on content integration test questions	.707	.501	.616	.693	.52	.518
Guidance on open/free response test questions	.699	.512	.596	.681	.536	.494
Students reporting lab findings to other students	.434	.812	.233	.430	.815	.205
Students work on inquiry lab investigations (Chemistry only)				.507	.743	.265

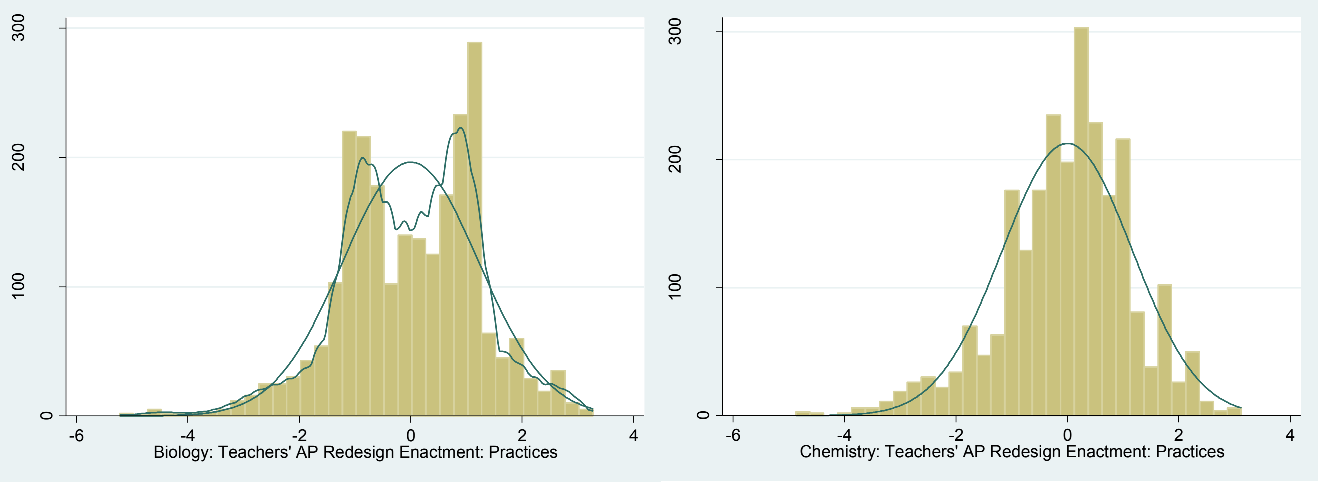

Figure A-6: Distribution of the *enactment of AP redesign practices* composite variable. (left) 2014 AP Biology data, (right) 2014 AP Chemistry data.

Table A12: *Enactment of AP redesign curriculum* composite variable.

	2014 AP Biology			**2014 AP Chemistry**		
Included items:	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***	***Rotated factor loading***	***Unique-ness***	***Scoring coefficient***
Refer to the "Big Ideas" of Biology/Chemistry	.725	.474	.304	.768	.410	.310
Use science practice outside of the classroom	.443	.804	.110	.375	.860	.072
Referring how "enduring understandings" relate to "Big Ideas"	.767	.412	.370	.781	.390	.331
Refer to learning objective from AP curriculum	.759	.424	.356	.785	.384	.338
Refer to curriculum framework	.686	.530	.258	.749	.440	.282

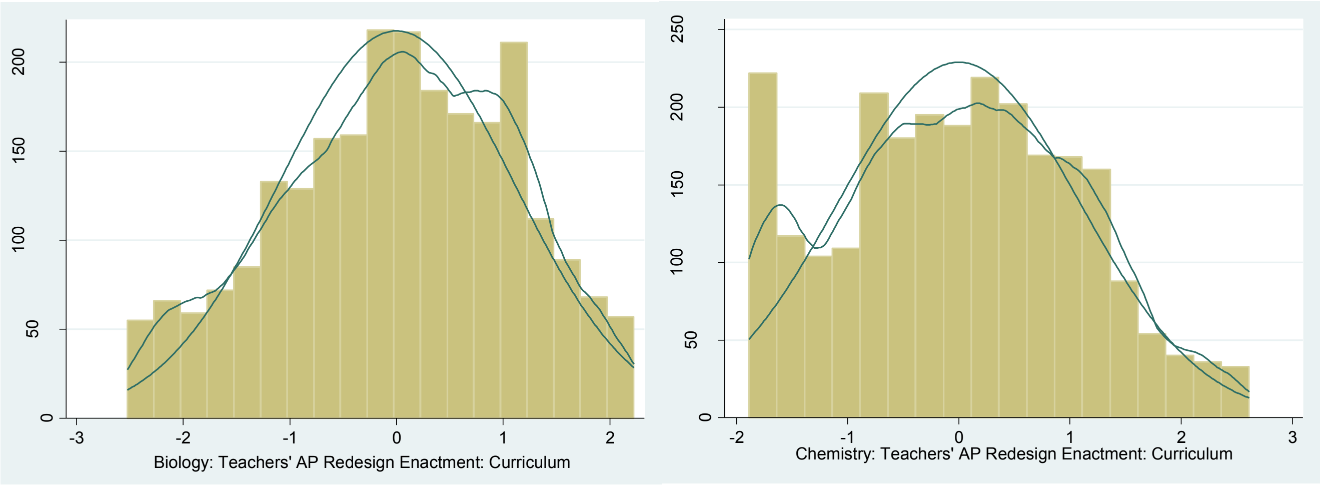

Figure A-7: Distribution of the *enactment of AP redesign curriculum* composite variable. (left) 2014 AP Biology data, (right) 2014 AP Chemistry data.

8.3 Exploring assumptions of hierarchical linear models

Assumptions for hierarchical linear models include independence of observations, homoscedasticity of the residuals, and the absence of perfect multicollinearity, among others. As described while exploring the first research question the quantitative variables included in the statistical model do not show heteroskedasticity. Therefore, we only need to test the multicollinearity assumption. Hence, variance inflation factors (VIF) for all independent variable for both levels are computed. Table A13 describes the results.

No VIF exceeds a value of ten indicating that no perfect multicollinearity exists between the independent variables included in the statistical model. Assuming a linear relationship between the independent variables and the dependent variable, hierarchical linear models can be conducted.

Table A13: Variance inflation factor for all independent variables included in the HLM analysis

Variable	VIF	Variable	VIF
Student level		**Teacher Characteristics**	
AP GPA	1.14	*Age*	1.49
Parents' education level	1.30	*Gender*	1.14
English language learner	1.04	*Degree*	1.19
Black or African American	1.28	*Courses*	1.16
Asian or Asian American	1.31	*Knowledge and Experience*	1.56
Hispanic	1.58	*PD Inclination*	1.13
American Indian or Alaska Native	1.03	*Black or African American*	1.19
School Characteristics		*Asian or Asian American*	1.16
District funding: All	1.38	*Hispanic*	1.11
District funding: Materials	1.28	*American Indian or Alaska Native*	1.07
Percentage of free- or reduced lunch students	1.19	**Teaching Characteristics**	
Days of school year	1.07	*Labs*	1.18
Special education classes	1.10	*Challenges with the AP redesign*	1.18
Charter school	1.14	*Enactment AP practices*	1.40
Rural or non-metro	1.43	*Enactment of AP curriculum*	1.42
Suburban	1.36	**PD Activities**	
Urban	1.42	*Active learning*	2.37
Principal support	1.13	*Responsive agenda*	5.06
AP workload	1.13	*Focus on student work*	4.09
Criteria for AP enrollment	1.11	*Modeling teaching*	4.66
		Building relationships	5.00
		Effective support	6.87
		Unconventional PD: Face- to-face	1.12
		Unconventional PD: Materials	1.19