Effects of the California High School Exit Exam on Student Persistence, Achievement, and Graduation

Sean F. Reardon

Allison Atteberry

Nicole Arshan

Stanford University

Michal Kurlaender

University of California, Davis

EMBARGOED UNTIL APRIL 21, 2009

VERSION: APRIL 19, 2009

Direct correspondence to Sean F. Reardon (sean.reardon@stanford.edu). The research reported here was supported by a grant from the James Irvine Foundation and by additional support from the Hewlett Foundation through a grant to the Institute for Research on Educational Policy and Practice at Stanford University. We are indebted to the staff of the four school districts (in particular, James Gulek, Dave Calhoun, Robert Maass, and Peter Bell) for sharing their data and expertise with us; without their generosity and commitment to quality research this work would not have been possible. We also benefited from excellent research assistance of Noli Brazil and Demetra Kalogrides. Joshua Aronson, Geoffrey Cohen, and David Plank provided us with very useful comments. All that said, the analysis and opinions expressed here are entirely our own and do not necessarily represent the views of the participating districts or the Irvine or Hewlett Foundations.

Abstract

High school exit exams have become a popular policy tool in the current movement in U.S. public schooling toward more explicit standards of instruction and accountability. In this paper, we investigate the effects of a high school exit exam requirement on students' achievement, persistence in high school, and graduation rates, using data from three cohorts of students—one of whom was not subject to the exit exam requirement and two of which were—from four large school districts in California. We find that the exit exam requirement had no positive effect on student achievement, small negative or zero effects on students' persistence in high school, and large negative effects on graduation rates. We estimate that graduation rates declined by 3.6 to 4.5 percentage points as a result of the exit exam policy. Moreover, we find that these negative effects were concentrated among low-achieving students, minority students, and female students. We investigate several hypotheses to explain these differential effects by race and gender and find that the data are consistent with a stereotype threat explanation: on a high-stakes test (such as an exit exam), minority students and girls perform less well relative to white and male students than we would predict on the basis of their prior and contemporaneous performance on other (low-stakes) exams. As a result, low-achieving minority students and girls fail the exit exam at substantially higher rates than otherwise similar white and male students, leading to lower graduation rates under the exit exam requirement. These findings call into question both the effectiveness and the fairness of high school exit exam policies. For states with high-stakes exit exam requirements in place, we recommend they evaluate and implement strategies to reduce the inequitable impacts of exit exam requirements, particularly strategies to eliminate the impacts of stereotype threat.

Introduction

The increasing use of state-mandated public high school exit exams—tests each student must pass before he or she is awarded a high school diploma—is one manifestation of the current movement in U.S. public schooling toward more explicit standards of instruction and accountability. Unlike some aspects of accountability systems, the accountability consequences of failing an exit exam fall primarily on students, as opposed to schools or districts. The number of states requiring students to pass an exam to graduate has increased from 18 in 2002 to 22 in 2007, with an additional four states intending to implement exit exams by 2015. Soon, over 70 percent of U.S. students will soon be subject to such exam requirements (see, e.g., Center on Education Policy, 2004, 2005; Dee & Jacob, 2006; Warren, Jenkins, & Kulick, 2006). The effects of exit exam policies, however, remain somewhat unclear, despite a number of recent studies. Competing notions of how such exams might influence student and school behaviors lead to divergent predictions of how students will be affected. Some argue, for example, that a high school exit exam requirement will create incentives both for schools to provide better instruction to struggling students, as well as for these students to work harder to learn more before graduation. On the other hand, others have argued that creating additional barriers to graduation discourages students—particularly academically and socially disadvantaged students—from persisting in school and hence leads to increased dropout rates and greater inequality (for discussion, see Dee & Jacob, 2006; Reardon & Galindo, 2002; Warren et al., 2006).

In this paper, we use longitudinal data from four large California districts to estimate the effects of an exit exam graduation requirement on subsequent student achievement, persistence in high school, and graduation. In particular, we examine the effects of the requirement on the outcomes of students with low prior achievement levels, as these are the students most likely to be affected by exit exam policies. We begin with a discussion of the mechanisms through which exit exams might influence student outcomes. We then review relevant research on these topics with an emphasis on distinguishing between two strands of research: research on the effects of exit exam requirements *per se*, and research on the effects of failing (relative to passing) an exit exam, given the existence of the requirement. This paper falls under the first type of research, which is particularly useful for providing guidance to policymakers (in a companion paper, we address the second question; see Reardon et al., 2008). In the third section of the paper, we describe the California High School Exit Exam policy and its history in order to provide a context for the analyses that follow. Section IV describes our analytic strategy; Section V describes the data, sample, and measures used in our analyses. In Section VI we describe the results of our analyses. Section VII explores the mechanisms for some of our findings in more detail. Finally, we discuss the implications of our results in Section VIII.

I. A General Model of Exit Exam Policy Effects

Adapting the general model describing the effects of exit exam policies described by Reardon et al (2008), we conceive of student outcomes (persistence, achievement, graduation; denoted by Y_i here) as a function (f) of student covariates ($\mathbf{X_i}$) and cognitive skill prior to the time when the exit exam is first administered (T_i^{pre}), plus some effect g of the existence of an exit exam policy (P_i). This effect may itself be a function of student characteristics and cognitive skill:

$$Y_i = f\left(T_i^{pre}, \mathbf{X_i}\right) + \left[g\left(T_i^{pre}, \mathbf{X_i}\right)\right] \cdot P_i + \nu_i \quad (1)$$

A more general model might allow f and g to depend on school or district characteristics as well, though we omit these here for parsimony. Reardon et al (2008) note that exit exam requirements may affect student outcomes through two mechanisms. First, the requirements may alter curriculum, instructional practices, and organizational features of schooling in ways that affect the outcomes of some or all students, regardless of their performance on the exit exam itself (i.e., exit exam policies may lead to a greater focus on low-level skill instruction, or to more extensive tracking within schools, each of which may affect the achievement and persistence of students regardless of whether they pass or fail the exam initially). Second, the requirements may have different effects on students who fail the test initially than on those who pass. Students' initial performance on the test may serve as a signal to both students and schools of students' likelihood of passing the test eventually; this signal may induce increased or decreased motivation on the part of the student, and/or may affect school instruction and curriculum for students who failed (e.g., schools may place students who fail the exams at their initial sitting in remedial classes or tutoring programs). These two sets of mechanisms imply that we might separate g in Model (1) into two components, yielding this model:

$$Y_i = f\left(T_i^{pre}, \mathbf{X_i}\right) + \left[h\left(T_i^{pre}, \mathbf{X_i}\right) + \delta\left(T_i^{pre}, \hat{T}_i^{init}, \mathbf{X_i}\right) \cdot \left(\hat{T}_i^{init} < z\right)\right] \cdot P_i + \nu_i \qquad (2)$$

Here h is the function describing the average effect of the policy as a function of prior skill and student covariates for students who pass the exam at their first attempt; z is the required passing score on the exit exam; δ is the function describing the effect of failing the test at the first attempt. Note that δ may depend not only on true prior skill and student covariates, but also on the observed score on the initial exit exam (which is an error-prone measure of true skill, denoted here as $\hat{T}_i^{init}$).

Several recent papers have estimated δ for students with test scores near the passing score using data from Texas (Martorell, 2005), New Jersey (Ou, 2009), Massachusetts (Papay, Murnane, & Willett, 2008), and California (Reardon et al., 2008). Each of these papers uses a regression discontinuity estimator to identify the effect of failing the exam. These papers generally find little average effect of failing an exit exam in 10^{th} grade on subsequent student persistence or achievement for students near the margin of passing the exam (though not all of the papers test for such effects). Several of the papers do find evidence that failing a mathematics exit exam in 10^{th}

grade leads to lower graduation rates for low income and/or minority students, however. These effects appear confined to low-income urban students in Massachusetts (Papay et al., 2008), to low-income and minority students in New Jersey (Ou, 2009), and to (the disproportionately minority) students who also fail the English Language Arts exam in California (Reardon et al., 2008). Papay et al and Reardon et al find that the effects appear to operate through the denial of the diploma rather than through inducing students to drop out. Ou's results, however, indicate that failing an exit exam may in fact lead to higher early dropout rates; he finds that failing an exit exam leads to lower persistence in high school, particularly among low income and minority students.

The regression discontinuity estimator used in the papers described above provides a strong causal warrant for estimating the effect of *failing* the exit exam (for students near the passing margin), given that the policy is in place, but it does not provide evidence of the effects of the policy itself. Unlike the papers described above, this paper aims to provide estimates of the effect of the *exit exam requirement* on student outcomes.

II. Prior Research on the Effects of Exit Exam Requirements

Prior research on the effects of exit exam requirements has primarily focused on estimating the extent to which such policies affect high school dropout rates. Several studies using a nationally representative sample of students from a single cohort graduating high school in the early 1990's find negative effects of failing the exam on at risk-populations (Bishop & Mane, 2001; Dee & Jacob, 2006; Jacob, 2001), although a similar study by Warren and Edwards finds no such effects (Warren & Edwards, 2005). Recent studies using multiple cohorts of data from multiple states, however, suggest that more difficult exit exams increase the dropout rate by about 2 percent, with effects concentrated in states or districts with high percentages of poor students of color (Dee & Jacob, 2006; Warren et al., 2006). We describe key features of these studies below.

Bishop and Mane (2001) use data from the National Educational Longitudinal Study of 1988 (NELS-88), which follows a nationally-representative sample of 8th graders in 1988 through 1994. The authors fit regression models to estimate the association between the presence of a state exit exam requirement and high school completion or GED receipt, controlling for a number of observed student, family, school and state characteristics. Their findings indicate that students in states with exit exam requirements took longer to receive a diploma and were more likely to receive a GED in lieu of a diploma than were observationally similar students in states without such requirements. Moreover, although they found no relationship between exit exam requirements and high school completion rates for high-achieving students, low-achieving students were about 7 percentage points less likely to earn a GED or diploma in states with exit exams than in states without them. Conversely, however, they find that students in states with exit exams were about 2 percentage points more likely to attend college in 1993/1994.

Jacob (2001) uses the same data and similar methods as Bishop and Mane, but estimates the effects of exit exams on 12th grade academic achievement as well as on graduation. Moreover, Jacob argues that these low-level exams should only impact low-achieving students and should raise their achievement either internally (through motivation) or externally (through school support). Jacob finds no effect of the exit exam requirements on achievement, but finds that low-performing 8th grade students were 9 percentage points more likely to drop out in states with an exit exam than in states without an exit exam.

A third paper (Warren & Edwards, 2005) using the NELS data improves on the methodology of the Bishop & Mane and Jacob papers. Warren and Edwards first point out that exit exams were not implemented at random; they were clustered in the Southeast, therefore disproportionately impacting poor, urban and minority students. The authors improve upon earlier NELS-88 work in several ways: (1) they treat earning a high school diploma, passing the GED, and dropping out as three distinct outcomes; (2) unlike Jacob (2001), they use the NELS 1994 (not 1992) survey to determine whether students obtained secondary school credentials, allowing more time for students to complete their degree; and (3) they do not rely on the NELS-88 data to determine which students were subject to an exit exam (following Jacob, 2001). Using these more careful methods, Warren and Edwards find no relationship between the presence of an exam and student's graduation outcomes, even when looking specifically at disadvantaged or low achieving students.

The NELS-88 data therefore provide ambiguous evidence regarding the effects of exit exams, but these studies have two important limitations. First, the NELS-88 data may be out of date, given that exit exams are becoming both more prevalent and increasingly difficult. Warren, Jenkins and Kulick (2006) suggest that many states have moved to more rigorous exit exams since that time and generally conclude that "the consequences of state HSEEs have changed in important ways since the NELS-88 cohort moved through the secondary school system" (p. 146).[1] Second, the NELS data include students sampled from only a single cohort of students, and so rely entirely on between-state variation in exit exam policies (in 1992, when the NELS cohort was scheduled to graduate) to identify the effect of exit exams. Because there may be important unmeasured state-level factors affecting graduation rates (including school quality and curricula; labor markets, unemployment rates, and returns to education), estimates from such models may be biased in unknown directions. A better strategy would compare similar students from different cohorts (some of whom were subject to an exit exam requirement and others who were not) within the same state or district. Such comparisons potentially eliminate much or all of the bias that may be present in cross-sectional analyses, under the assumption that other state- or district-level factors affecting graduation rates do not change sharply at the same time as the introduction of the exit exam requirement. Two more recent studies using state and district level data correct for these

[1] The 2000 PUMS data from Dee & Jacobs, 2006, included students aged 18 between 1980 and 1998, so it only provided four additional years of data.

shortcomings and indicate that exit exams may increase states' dropout rates, with effects concentrated in states and districts with more disadvantaged students (Dee & Jacob, 2006; Warren et al., 2006).

Dee and Jacob (2006) use data from the 2000 Census to compare high school completion rates and labor market outcomes for students from multiple birth cohorts (those who were aged 18 between 1980 and 1998) within the same state who experienced different exit exam policies. They find an association between exit exams and higher dropout rates, with easier exams leading to a 0.5 percentage point reduction in high school completion or GED receipt rates, and harder exams leading to a 0.7 percentage point reduction in high school completion or GED receipt rates. The effects are roughly two to three times larger for black students. Because they use Census data, however, they cannot control for students' level of academic skill or test for differences in the magnitude of the effects for students of different skill levels.

In a second analysis in the same paper, Dee and Jacob (2006) use district-level panel data from Minnesota to estimate the effects of the Minnesota exit exam on dropout rates and their timing. This data is more recent and more detailed with regard to timing than either the NELS-88 or 2000 PUMS data, though it relies on aggregated data (grade-by-district counts) rather than individual data. Although they do not find that the Minnesota exit exam affected graduation rates overall, they find that dropout rates increased in urban, high poverty, and high minority (for Minnesota) school districts in response to the exit exam policy. In urban districts, for example, they found that the exit exam reduced graduation rates by 2 to3 percentage points as a result of the policy.

Another recent paper using data from multiple cohorts of students in multiple states also finds that exit exam requirements lower graduation rates. Warren, Jenkins and Kulick (2006) divide high school exit exams into two classifications, "minimum competency" exams and "more difficult" exams, and estimate the effect of these exams on state-level high school completion rates. The authors use state-level panel data from the graduating classes of 1975 to 2002. They estimate the effects of the exit exam policies on graduation rates and GED-taking rates using a state-by-year fixed effects model with a number of state economic and education policy covariates that may vary both across states and over time. They find that more difficult exit exams lower graduation rates by 2.1 percentage points and lower GED-taking rates by 0.1 percentage points. In a final set of models, the authors find that exit exams have a larger negative impact on graduation rates in states with larger proportions of poor students and in states with larger proportions of minority students.

Several conclusions can be drawn from the studies described above. First, there is persuasive evidence from the studies using panel data that exit exams reduce graduation rates by roughly 1 to 2 percentage points on average. Second, many of the studies find that the impact of exit exam requirements is larger in states or districts with more poor and minority students or is larger for low-achieving students (Dee & Jacob, 2006; Warren et al., 2006). It is not clear, however, whether the differential racial and socioeconomic effects in the panel studies result from different levels of academic achievement by race and socioeconomic status, or whether these patterns would persist even if the studies had been able to control for individual academic achievement. Moreover, the existing research provides little evidence on the effects of exit exams on academic achievement, which is a key aim of such policies.

In this paper, we improve on prior research in several ways. We use longitudinal, student-level data from multiple cohorts of students in the same districts. This allows us to compare achievement, persistence, and graduation rates of students who were and were not subject to an exit exam requirement. Moreover, we investigate whether there are patterns of differential effects by prior achievement, race/ethnicity, gender, poverty, and ELL status, rather than relying on aggregate data and ecological correlations as in the papers above (Dee & Jacob, 2006; Warren et al., 2006). Finally, we are able to investigate, to some extent, several competing hypotheses explaining the differential effects we do observe.

III. The California High School Exit Exam

The California State Legislature passed Senate Bill SB2X in March 1999, requiring California local school districts to administer a high school exit exam (the California High School Exit Exam, known as the CAHSEE) and provide supplemental instruction to those students who do not demonstrate sufficient progress toward passing the exam. The stated rationale for adopting the exit exam law was that, in order to "significantly improve pupil achievement in high school and to ensure that pupils who graduate from high school can demonstrate grade level competency in reading, writing, and mathematics, the state must set higher standards for high school graduation" (SB2X, Section 1(b)).[2]

As implemented, the CAHSEE is a two-part exam of mathematics and English language arts (ELA) skills. The math section assesses students' mastery of the California math content standards for 6th and 7th grade and their Algebra I skills using a multiple-choice format. The ELA section is aligned with state content standards through grade ten and utilizes a multiple-choice format along with one essay. Both tests are administered in English, regardless of a student's primary language. Both parts are scored on a scale from 275 to 450, and students must score at least 350 points on each part to pass the exam and earn a high school diploma.

The test is first administered to students in the spring of 10th grade, and students have at least five subsequent opportunities to retake the sections they have not yet passed (twice in 11th grade and up to three times in 12th grade, and as many times as necessary following the end of the

[2] Hhttp://www.leginfo.ca.gov/pub/99-00/bill/sen/sb_0001-0050/sbx1_2_bill_19990329_chaptered.html

12th grade school year). Testing dates are centrally scheduled by individual districts and the exam is administered over the course of two days (one day for each portion). The test is untimed, though students typically complete each section in three to four hours. Districts notify students and their parents of their CAHSEE performance about seven weeks after the exam is administered. Because students are told their exact score, not simply whether they passed or failed, students who fail have some sense of how close they came to scoring the requisite 350 they need to meet the CAHSEE requirement.

Some important aspects of the CAHSEE requirement have changed since the law was first passed. The original legislation identified the Graduating Class of 2004 (students entering high school in the fall of 2000) as the first graduating class subject to the CAHSEE graduation requirement. In most districts, students scheduled to graduate in 2004 were given the opportunity to take the CAHSEE exam in spring of 9th grade (2001), though these students were not required to take the exam until 10th grade. Only the Class of 2004 was offered the 9th grade administration; after spring 2001, the state mandated that students take the test for the first time in the spring of 10th grade. This meant that students in the Class of 2004 took the CAHSEE as early as spring 2001 (their 9th grade year) but students in the Class of 2005 did not take the CAHSEE for the first time until spring 2003 (their 10th grade year). Figure 1 describes the timing of the CAHSEE administrations for modal students in each cohort.[3]

Figure 1 here

In July of 2003, after the completion of the spring 2002–03 administrations of the CAHSEE (taken by 10th graders in the high school class of 2005 and by juniors in the class of 2004 who had

[3] In addition to the change in timing between the first two cohorts to take the CAHSEE, the State Board of Education voted in July 2003 to rescale the Math section in order to "reduce cognitive demands for mathematics questions while still assessing the same standards" (Wise et al., 2003, p. 90) (i.e., they made the math test easier to pass). The decision was made in response to recommendations from an independent study report released at the end of the 2002-03 school year. At the same time, the Board also shortened the ELA testing to a single day. In this paper, we convert 2003 CAHSEE scores to the same metric as the later exams using the equating formulae reported in Wise et al (2004, pp. 19-20).

not yet passed the exam), the State Board of Education voted to defer the CAHSEE requirement for two years.[4] As a result, students in the Classes of 2004 and 2005 took the CAHSEE exam at least once under the belief that passing would be required to graduate, however they were ultimately not subject to the policy (see Figure 1 above). For those students who were in 10th grade in spring 2004 or later, however, the CAHSEE requirement was consistently in place starting in their 10th grade year and was enforced for graduation in 2006.[5]

IV. Analytic Strategy

Our basic strategy relies on comparing a cohort of students who were not subject to the exit exam requirement to students in two cohorts who were subject to the requirement. More specifically, we compare the average outcomes of students who were in 10th grade in spring 2003 (and who therefore took the CAHSEE in 10th grade thinking it would be a requirement for their graduation, only to find out it would not be) with those of observationally similar students (students with similar 9th and 10th grade standardized test scores) who were in 10th grade in 2004

[4] In June 2000, an independent evaluation of the CAHSEE exam's development, implementation, and effects on students recommended the implementation of the CAHSEE requirement be delayed, stating that:

> "The…reason for considering a delay is that schools will need more time to prepare students to meet the standards assessed by the HSEE [sic]…the key legal issue in prior challenges to high stakes test is whether students have been provided adequate instruction in the material covered by the test. Current plans call for students and schools to be fully notified about the exam and its requirement this fall, as the first affected class (the Class of 2004) enters 9th grade. This will be too late to allow very significant changes in the 9th grade curriculum for these students…" (Wise et al., 2000, p. 70)

[5] A further minor complication in the CAHSEE timeline arose for the Class of 2006—the first class to ultimately be subject to the CAHSEE exam requirement. In February of their senior year (2006), a lawsuit (*Valenzuela v. O'Connell*) was filed on behalf of high school students who had not yet passed the CAHSEE exam. The plaintiffs alleged that students had been deprived of a fundamental right to public education and equal opportunity to pass the CAHSEE given the unequal distribution of resources across the states' schools. Indeed, for twelve days of their final semester, students in the Class of 2006 were relieved by an Alameda Superior Court Judge from their requirement to pass the CAHSEE. This decision was quickly overturned, however, by the California Supreme Court.

One worries that the debate surrounding the legality of the CAHSEE in the spring of 2006 may have led to some ambiguity for students about whether the CAHSEE would be enforced. However, seniors in the Class of 2006 had already completed their final administration of the CAHSEE *before* the twelve days when the CAHSEE requirement was temporarily suspended. For the students who entered their final semester of high school having met every graduation requirement except the CAHSEE, perhaps they saw some hope in the looming court case *Valenzuela v. O'Connell*, however there was never any formal indication from the California Department of Education that the CAHSEE requirement would be waived. The state Superintendent of Schools, Jack O'Connell, had issued several statements reaffirming his commitment to enforcing the CAHSEE exam for the Class of 2006.

or 2005 (and who therefore took the CAHSEE in 10th grade under the—accurate—belief that it would be required for their graduation). These cohorts operated under the same belief—that the CAHSEE would be required for graduation—through the end of 10th grade, but differed in their experience after 10th grade. While this design does not have as strong a causal warrant as the regression discontinuity designs used to estimate the effects of failing the exam for those at the margin of passing (because differences in outcomes between cohorts cannot unambiguously be attributed to the exit exam policy), it provides estimates that are much more generalizable and policy-relevant, because it indicates the average effects of the policy itself. Our basic model specification is this:

$$Y_i = \alpha_0\left(CST10_i^q\right) + \alpha_1\left(P_i \cdot CST10_i^q\right) + \mathbf{X_i\beta} + \gamma(P_i) + \epsilon_i \qquad (3)$$

where P_i is an indicator variable taking on the value of 1 for students subject to the exit exam requirement (those in 10th grade in spring 2004 or 2005, whom we will refer to as the 2004 and 2005 cohorts) and a value of 0 for those not subject to the policy (those in 10th grade in 2003, whom we will refer to as the 2003 cohort). Additionally, $CST10_i^q$ is student i's score on the 10th grade English Language Arts (ELA) California Standards Test (CST) (the state tests taken by all students and used for the state's school accountability purposes), expressed in percentiles of the statewide CST score distribution (more detail on this below) and centered at the midpoint of quartile q (e.g., centered at .125 in the bottom quartile, at .375 in the second quartile, and so on); and $\mathbf{X_i}$ is a vector of student-level covariates (including 9th grade ELA CST scores, a set of race/ethnicity indicator variables, gender, free-/reduced-price lunch eligibility status, and ELL-status). In some specifications, we add a vector of district fixed effects. The parameter of interest here is γ, which is interpreted as the average effect of the exit exam requirement for students with 10th grade ELA CST scores at the midpoint of quartile q.[6]

One concern with Model (3) above is that γ may reflect the effect of other factors that changed at the same time as the CAHSEE policy requirement and that also affect student outcomes. If, for example, districts changed their own graduation requirements between the 2003 and 2004 cohort, or labor market conditions changed, we might observe a change in average outcomes even if the CAHSEE policy had no effect. Likewise, if the measurement of the outcome variable Y changed between the cohorts (for example, if Y measures academic achievement and the scale of the test metric shifts over between cohorts), then γ may be reflect this change as well as any real effect of the CAHSEE requirement. To address this concern, we examine two additional outcome patterns. First, we expect that the CAHSEE will have its largest effects (either positive of negative) on low-achieving students, as these are the students for whom the CAHSEE requirement represents the largest barrier to graduation (see, e.g., Jacob, 2001). Any discouragement or motivational effects might be expected to be largest for low-achieving students. We examine the pattern of results to determine if the estimated effects are largest for students in the bottom quartile of the 10th grade achievement distribution and thus consistent with a CAHSEE explanation.

Second, we expect the CAHSEE requirement to have larger effects for students who fail the CAHSEE in 10th grade than for those who pass it. For those who pass the CAHSEE in 10th grade, the exit exam requirement poses no additional hurdle to graduation, and so is likely to have little or no effect on subsequent outcomes. The CAHSEE requirement may have some effect even on students who pass the test, of course, if it alters the curriculum that all students experience or if, for example, effects on failing students had spillover or peer effects on those who did not fail. For those students who fail the CAHSEE, we expect larger effects in comparison to the effects on those who initially

[6] We fit models that are quadratic in $CST10_i^q$ as well; we find, however, that inclusion of the quadratic term and its interaction with P_i never improves the model fit, so we report only the more parsimonious models. In the linear model here, γ indicates not only the effect of exit exam policy at the midpoint of quartile q, but can also be interpreted as the average effect of the exit exam policy for students in quartile q (this results from the fact that the density of the $CST10^q$ distribution is uniform).

pass. If we can observe whether students would fail the exit exam even if they were in a cohort not subject to it, we can use a difference-in-differences model to test whether the changes in outcomes between cohorts are larger for those who would fail than for those who would pass the exam in 10th grade if administered it. Let $F10_i$ be an indicator variable that takes on a value of 1 if student i would fail the CAHSEE in spring of 10th grade and a value of 0 if not. Then we would fit the model:

$$Y_i = \alpha_0\left(CST10_i^q\right) + \alpha_1\left(P_i \cdot CST10_i^q\right) + \mathbf{X_i}\boldsymbol{\beta} + \gamma(P_i) + \eta(F10_i) + \delta(P_i \cdot F10_i) + \epsilon_i \qquad (4)$$

Normally we would not be able to observe $F10_i$ for students who were not subject to the exit exam requirement because such students would have no reason to take the exit exam and we would have no way of knowing if they would have failed it in 10th grade. As we noted above, however, California 10th graders in 2003 took the CAHSEE exam under the belief that passing it was a graduation requirement, only to have the requirement lifted several months later. Thus, we have 10th grade CAHSEE scores (and passing status) for all three of our cohorts of students, allowing us to fit Model (4).[7]

In Model (4), γ indicates the difference in average outcomes between the cohorts subject to and not subject to the CAHSEE requirement among those who would pass the CAHSEE test in 10th grade if administered it. Under the relatively strong assumption that nothing but the CAHSEE requirement changed between the 2003 and 2004/05 10th grade cohorts, this can be interpreted as the effect of the CAHSEE policy on those who would pass the test in 10th grade. In the absence of that assumption, we cannot be sure whether γ represents an effect of the exit exam policy on the outcomes of those who pass the exam in 10th grade or some other correlated change in outcomes. Additionally, η indicates the average difference in outcomes in the absence of the CAHSEE policy between those who would fail and pass the test had they been administered it. Note that we cannot interpret this coefficient as the effect of failing the CAHSEE because failure on the 10th grade test is

[7] As noted above, we convert the 2003 10th grade CAHSEE scores into the metric used in 2004 and 2005, and then reclassify students who took that test in 10th grade in 2003 as passing or failing depending on whether their converted scores were above 349.

likely to be correlated with unmeasured student characteristics that negatively affect outcomes (such as mathematics skills or motivation) and that are not perfectly captured by the controls in the model. As a result, we would expect η to be negative even if failing the CAHSEE exam in 10th grade has no direct effect on subsequent student outcomes. Finally, δ indicates the difference between the effect of the policy on those who would fail and on those who would pass the test in 10th grade. Under the assumption that there was nothing but the CAHSEE policy that changed from 2003 to 2004/05 that would have affected failers' outcomes differently than passers' outcomes (controlling for the covariates in the model), we can interpret δ as the average effect of the policy on students who fail the test in 10th grade and who have 10th grade ELA CST scores at the midpoint of quartile q.

From Model (4) we can obtain upper and lower bound estimates of the average effect of the policy on students in quartile q: if we believe that the CAHSEE policy can only have an effect on those who fail the CAHSEE in 10th grade, then the average effect of the policy will be equal to $\pi_{F10}^{q} \cdot \delta$, where π_{F10}^{q} is the proportion of students in quartile q who would fail the CAHSEE in 10th grade if they were subject to the requirement. This is a lower bound estimate since it assumes that none of the difference in outcomes represented by γ is caused by the policy. In contrast, if we assume that the difference γ is entirely caused by the CAHSEE requirement, then we obtain a plausible upper bound for the effect of the policy, $\gamma + \pi_{F10}^{q} \cdot \delta$.[8]

In order to investigate whether the policy has a disproportionate effect on students of different race/ethnicity, gender, poverty status, ELL status or on students in different districts, we fit versions of Models (3) and (4) with interaction terms between the CAHSEE requirement and student demographic covariates (race/ethnicity, gender, free/reduced-price lunch eligibility, and ELL status).

[8] This upper bound will be approximately equal to the γ obtained from Model (3) above. It will differ slightly because we estimate Models (3) and (4) on slightly different samples, since Model (4) requires students have non-missing 10th grade CAHSEE scores.

Comparing test scores across cohorts

One complicating factor in our analysis is that our design relies on our ability to compare students in one cohort to those who have the same level of 10th grade academic skills but who are in a different cohort. Ideally, we would like to have administered identical standardized tests to the cohorts before and after the CAHSEE requirement came into effect in order to ensure we have comparable and independent measures of students' academic skill. Several factors complicate this in our data, however.

First, although students in California take statewide, grade-specific standardized tests (the CST tests) in English Language Arts (grades 2-11) and math (grades 2-7), students do not all take the same math test throughout high school; rather, they take math tests that correspond to the level of math course they are enrolled in each year. Two problems arise as a result of this: first, the test scores on different math tests are not designed to be comparable (e.g., a score of 300 on the algebra 1 test is not comparable to a score of 300 on the geometry test. Second, patterns of 9th, 10th, and 11th grade math course enrollment may have also been affected by the implementation of the CAHSEE requirement; this might alter the average math CST scores.

Since all California students take grade-specific ELA CST tests through 11th grade, the selection concerns regarding the math tests are not present with regard to the ELA scores. Thus, we rely on the 9th, 10th, and 11th grade ELA CST scores in our primary analyses because they are the only test scores we have that are comparable across our three cohorts and that allow us to rank students by percentile within cohorts.[9] That said, there is some reason to think that the metrics of the CST tests are not exactly comparable across cohorts. The equating of the metrics is designed to make scores directly comparable at the proficiency threshold; this does not guarantee comparability across the full range of scores. Figure 2 illustrates the (smoothed) cumulative

[9] Despite this, in some analyses below we will rely on the potentially biased math CST scores as well as ELA scores as control variables.

density functions of the 9th, 10th, and 11th grade ELA CST statewide score distributions for the three cohorts of students we use in our analyses. Note, for example, that the 2003 cohort's 9th-grade ELA CST test score distribution is far to the left of the 2004 and 2005 cohort distributions. If the test metric is identical across cohorts, this indicates that students in the 2004 and 2005 cohorts performed substantially better on their 9th grade ELA CST than did the 2003 cohort. On the other hand, if the true ELA skill distribution did not change across cohorts, the difference in the distributions implies that the test metrics are not identical across cohorts—the 2003 cohort took a test that was "harder" (meaning a given numerical score corresponded to a higher level of skill on the 2003 test than on the 2004 and 2005 tests).

Figure 2 about here

The 10th and 11th grade tests have more complicated patterns of change across cohorts than does the 9th grade test (see lower two panels of Figure 2). In both cases, the 2003 cohort's test cumulative density function is steeper than that of the 2004 and 2005 cohorts, implying less variance in test scores in the 2003 than in the later cohorts. This could result from some process that improved the skills of high-achieving students much more than of low-achieving students; or it could result from differences in the scaling of the tests across cohorts (or some combination of the two).

Because we cannot be sure how much of the difference in test score distributions across cohorts results from differences in true skill and how much results from differences in the test metrics, we rely on two different versions of the test scores for our analyses. First, we assume the test metrics are identical across cohorts and that the underlying true skill distributions varied. In this case, we use the reported 9th and 10th grade ELA CST scores as control variables in our analyses. For interpretability, we convert the scores to the corresponding percentiles of the distribution of the 2004 cohort.[10] Second, we assume instead that the underlying true skill distribution is identical in each cohort but the test metric was varies, and convert the test scores to the corresponding percentiles of their own cohort's distribution.[11] Because we are agnostic regarding which of these assumptions is most appropriate, we focus our attention to those results that are robust to our choice of which set of scores we use.

A final concern regarding the test scores is that the 10th grade ELA CST scores are measured with error. As a result, the coefficients on the test scores will be biased in model 1. However, in general, the measurement error in the test scores will not bias the coefficient of interest (the change in outcomes between the 2003 and 2004/05 cohorts), because test scores are not correlated with cohort.

V. Data and Measures

We use longitudinal student-level data from four large California districts—Fresno, Long Beach, San Diego, and San Francisco Unified School Districts—to investigate the effects of the CAHSEE exam. These are four of the eight largest school districts in California, collectively enrolling over 110,000 high school students (about 6 percent of high school students in the state) annually. For our primary analyses, we use data from three cohorts of students—roughly defined as the cohorts scheduled to graduate in 2005, 2006, and 2007—one of which (2005 graduates) was not subject to the CAHSEE graduation requirement, and two of which were (the 2006 and 2007 graduating class).

To be precise, we define a student's cohort as the year in which he or she was in 10th grade

[10] That is, for each grade, we replace a CST test score with its corresponding percentile in the 2004 cohort's statewide score distribution. For example, in 2004, a 10th grade ELA CST score of 323 was the median of the statewide test score distribution. We therefore assign a score of .50 to any student who scored a 323 on the 10th grade CST, regardless of whether s/he was in the 2003, 2004, or 2005 cohort.

[11] That is, a score of 323 on the 10th grade CST would correspond to a score of .50 in 2003, a score of .50 in 2004, and a score of .48 in 2005.

and scheduled to take the CAHSEE for the *first* time. For almost all students, this is in the spring of their 10th grade year, though for a very small number of students (0.6 percent of our total sample) it occurred in the spring of their second 9th grade year (because in some cases repeat 9th graders were considered 10th graders for the purposes of CAHSEE testing). The stipulation of *first-time* 10th graders ensures that students who repeat 10th grade are not assigned to more than one cohort. Thus, our analyses include students who were scheduled to take the CAHSEE for the first time in spring 2003, 2004, and 2005. As described above, students in each of these cohorts were originally required to pass the CAHSEE in order to graduate. We exclude from our analyses students classified as special education students (roughly 10 percent of students), because these students were not subject to the CAHSEE requirement in the years we examine.

<u>Measures</u>

We estimate the effect of the CAHSEE requirement on four outcomes—academic achievement, persistence through 11th grade, persistence through 12th grade, and graduation. We measure academic achievement using the spring 11th grade English Language Arts (ELA) California Standards Test (CST) score. We use the actual score on the test (rather than converting it to percentiles, as we do for the control variables) for ease of interpretability.

Although we cannot directly determine whether students have dropped out of high school—because students who leave a given district prior to graduation may be dropouts or may have left and enrolled elsewhere—we can identify whether students are present in the district one and two years after first taking the CAHSEE (in the spring of 11th and 12th grade respectively). We construct a binary variable indicating whether students are present in the district in a given spring semester using data on their GPA, their CST score, and their CAHSEE score. Students with any evidence that they were enrolled and attended school in a given term—specifically, a non-zero GPA, a non-missing CST score, or a non-missing CAHSEE score (for the 2004 and 2005 10th grade cohorts)—are coded as present in the district in that term, and not present otherwise. For students who leave the district and then return (present in some later term), we retroactively code them as present for all terms prior to the final one in which they are observed to be present. In addition, students who received a diploma from the district in an earlier semester are coded as present in order that they not be counted among leavers/dropouts in our persistence models (i.e., our "present" indicator is coded 0 for anyone who left the district for good prior to receiving a diploma, and is coded 1 for those who have graduated or are still enrolled at a given semester).

We use the indicator of presence in spring of the scheduled 12th grade year as an indicator of persistence in schooling. Of course, some students may not be present in the district because they have transferred to another district. Nonetheless, if we observe that the CAHSEE requirement affects the probability that a student is present in the district in 12th grade, we may reasonably assume that this is because the CAHSEE requirement affects persistence/dropout rates. It is unlikely that the CAHSEE requirement affects the probability of transferring to another district within the state, because students will be subject to the CAHSEE requirement in any district within the state. It is, however, possible that the CAHSEE requirement may affect the probability of transferring into private schools (where the CAHSEE is not required) or even out of the state, but such effects would likely be very small. Thus, we argue that any effects of the CAHSEE requirement on persistence are likely due to dropout effects.

Finally, we estimate the effect of the CAHSEE requirement on the probability of graduating from the district using a binary indicator of graduation status provided by the districts. Table 1 describes demographic characteristics of our sample and of the corresponding state student population.

Table 1 here

VI. Results

First we examine non-parametric plots describing the average values of each of the four outcomes, for each cohort, by 10th grade ELA CST percentile. Figure 3 shows average rates of persistence to 11th grade, rates of persistence to 12th grade, 11th grade ELA CST scores, and graduation rates for each cohort as a function of the cohort-specific 10th grade ELA score. Figure 4 shows the same, but uses the 10th-grade ELA CST percentiles that correspond to the distribution in 2004. Evident in both sets of figures are small differences between students in the 2003 and 2004/05 10th grade cohorts in persistence to 11th grade and 12th grade for students in the lowest quartile. There is a mixed pattern of differences in 11th ELA CST scores—11th grade CST scores are slightly lower for students with low 10th grade ELA CST scores in the classes of 2006 and 2007 than for similar students in the class of 2005; but are higher in the classes of 2006 and 2007 for students with high 10th grade ELA CST scores. It is unclear whether this pattern is a result of real changes in the academic skill distribution of students or is a result of slight differences in the test metrics in different years. Finally, note that there is a pronounced difference in graduation rates between the cohorts for students in the bottom third of the 10th grade test score distribution. In general, the patterns are not dependent on which version of the 10th grade CST score metric we rely on.

Figures 3 and 4 here

Figures 3 and 4 suggest the CAHSEE has modest negative effects on persistence in schooling and achievement and has large negative effects on graduation rates. In the next step of the analysis, we estimate the magnitude and standard errors of these effects in a regression framework. Tables 2A and 2B report estimates from a set of models of the type described in Model (3) above. We fit models separately for each quartile, centering the 10th grade CST score at the middle of the quartile, so that the estimates apply to the average student in that quartile. Table 2A shows estimates from models that condition on cohort-specific test score percentiles; Table 2B shows estimates from the same set of models, but conditioning on class of 2006-normed CST percentiles rather than the cohort-specific version. The estimates are very similar across both test metrics; we focus on the cohort-specific results here and throughout the rest of the paper, because these tend to yield slightly smaller (closer to zero) estimated differences, so this is the more conservative set of estimates. In addition, we focus only on estimates that are significant in both sets of models.

Tables 2A and 2B here

Each panel of Tables 2A and 2B includes three versions of the general Model (3) described above. Model 3a simply reports unadjusted differences between cohorts; Model 3b reports estimates adjusted for 9th and 10th grade ELA CST scores, as well as10th grade scores interacted with cohort, race, gender, free/reduced-price lunch eligibility, and ELL status. Model 3c adds district-fixed effects to Model 3b. As we would expect, the inclusion of the covariates and the fixed effects changes the estimates very little, because there is little correlation between student covariates and what cohort s/he is in. The inclusion of the covariates and the district fixed effects does, however, improve the precision of the models. We rely on model 3c here and in our subsequent analyses.

The estimates in Tables 2A and 2B correspond well with those in Figures 3 and 4. Among students in the lowest quartile, there are modest differences in persistence to 11th grade (2 percentage points lower in 2004/05) and persistence to 12th grade (5 percentage points lower in 2004/05). However, there are large differences (15 percentage points lower in 2004/05) in graduation rates. Together, these finding suggest that the CAHSEE induces 5% of bottom quartile students to leave school before the spring of 12th grade, and results in another 10% of bottom quartile students not receiving diplomas, despite the fact that they persist to 12th grade. In the second quartile, effects on persistence (2 percentage points in 11th and 12th grade, respectively) and graduation (3 percentage points) are much smaller. There are no differences in persistence and graduation for students above the state median 10th grade score.

The effects of the CAHSEE requirement on achievement are less clear in Tables 2A and 2B. Differences in average 11th grade CST scores are negative and significantly different from zero in the bottom two quartiles regardless of which test metric we use (Table 2A or 2B). In the top two quartiles, however, the pattern of differences depends very much on the test metric. In fact achievement is much higher in the top quartile for the cohorts subject to the CAHSEE requirement when we use the cohort-specific percentile metric (Table 2A), a pattern that is hard to attribute to the CAHSEE. Such a pattern would result if the test metric of the 2005 and 2006 11th grade CST tests (those taken by the 2004 and 2005 10th grade cohorts) were such that it were easier to get a very high score in those years than in the prior year. The bottom panel of Figure 1 is consistent with this pattern—the cumulative density of 2003 cohort 11th grade CST scores is steeper than the 2004 and 2005 cumulative densities, indicating fatter tails in the later cohorts' test score distributions. It is not clear whether this is a result of real changes in the distribution of true skill (induced by the CAHSEE policy or some other changes) or an artifact of changes in test scaling, though we suspect that latter is at least part of the issue. Our next set of analyses sheds some light on this issue.

Figures 3 and 4 and Tables 2A and 2B indicate sharp differences in outcomes between the 2003 and later cohorts, particularly for students with low prior achievement levels and particularly in graduation rates. But are these differences due to the CAHSEE requirement? The fact that these differences are concentrated among the low-achieving students is certainly consistent with what we would expect if the differences were driven by the CAHSEE requirement. Moreover, the trend—sharp differences between 2003 and 2004 but no significant differences between 2004 and 2005—suggests that whatever caused the differences between 2003 and the later cohorts was due to some factor(s) that changed sharply between the 2003 and 2004 cohorts rather than a gradual change over several years.

The difference-in-difference estimates from Model (4) provide a stronger test of whether we can attribute the changes in outcomes between the cohorts to the CAHSEE requirement. In Table 3, we report key estimated parameters from Model (4). The key parameters of interest here are γ, the estimated between-cohort (pre- and post-CAHSEE) difference in average outcomes among students who pass the CAHSEE in 10^{th} grade (row 1 of each panel) and δ, the estimated difference in between-cohort differences in average outcomes between students who pass and students who fail the CAHSEE in 10^{th} grade (row 3 of each panel). If the estimated δ in row three is significantly different from zero, this indicates that the outcome changed more between students who failed the CAHSEE than for similar students who passed the CASHEE. Note that the coefficient on the interaction term should not be interpreted as the *effect of failing* the CAHSEE in 10^{th} grade, given that the CAHSEE is required for graduation. Rather it should be interpreted as the *effect of the CAHSEE requirement* on the type of students who fail the CAHSEE in 10^{th} grade.

Table 3 indicates that the persistence and achievement differences between cohorts that are evident in Table 2A result primarily because persistence and achievement declined, on average, for all students in the bottom part of the achievement distribution after the CAHSEE requirement was in place. The top row of each panel in Table 3 shows that persistence rates declined, on average, by about 5 percentage points in the bottom quartile and by 2 percentage points in the second quartile between the 2003 and 2004/05 cohorts for students who passed the CAHSEE in 10^{th} grade. Likewise, achievement declined by about 5 points on the 11^{th} grade ELA CST test between the 2003 and 2004/05 cohorts for students who passed the CAHSEE in 10^{th} grade. The third row shows that there is no significant difference in the change in persistence and achievement patterns between students who failed the 10^{th} grade CAHSEE and those who passed (the interaction terms in row 3 are generally not significantly different from 0). This implies that the differences in persistence and achievement between the pre- and post-CAHSEE cohorts either a) are not caused by the CAHSEE requirement or b) result from the CAHSEE requirement negatively affecting both students who pass the CAHSEE and those who fail.

Table 3 here

The bottom panel of Table 3 shows that the CAHSEE requirement has large negative effects on graduation rates for students who fail the CAHSEE in 10^{th} grade, relative to differences in graduation rates for students who pass in 10^{th} grade. Among bottom quartile students who pass the CAHSEE in 10^{th} grade, graduation rates are roughly 6 percentage points lower for those who were ultimately subject to the CAHSEE graduation requirement (cohorts 2004/05). Because persistence rates for those who passed the CAHSEE were roughly 5 percentage points lower among those subject to the requirement, most of the difference in graduation rates is driven by differences in persistence. On the other hand, among students who *fail* the initial 10^{th} grade CAHSEE, graduation rates are 20% percentage points lower $(-.063 + -.138 = -.201)$ when the CAHSEE policy was in effect. The CAHSEE requirement also negatively affects graduation rates among students who fail the test in 10^{th} grade and who are in the second and third quartiles of the distribution, though the effects are smaller for these students (likely because they have a greater likelihood of subsequently passing the test).

It is important to note that interpretation of the coefficients in row 1 of each panel hinge on whether one thinks it plausible that the CAHSEE may have affected persistence, achievement, and graduation patterns for students who passed the CAHSEE. For example, the CAHSEE could affect outcomes even for students who pass if it induces changes in the curriculum and instruction patterns in the classrooms of low-achieving students. Because roughly 80% of bottom quartile students fail one or both portions of the CAHSEE in 10^{th} grade, bottom quartile students who pass the CAHSEE may find themselves in classrooms in 11^{th} or 12^{th} grade with a large number of peers who have not passed the CAHSEE. If the presence of these students affects the instructional practices and curriculum in their classrooms—likely by focusing instruction more on remediation and basic skills, at least in the cohorts for whom the CAHSEE was required—then this may have spillover effects on all students in the classroom, even among those who have already passed the CAHSEE (for recent evidence of such phenomena, see Lavy, Paserman, & Schlosser, 2008). If being in such classrooms leads to lower levels of academic achievement and increases the likelihood of leaving school early, then it is conceivable that the CAHSEE requirement affects persistence, achievement, and graduation patterns even for students who pass the test in the 10th grade. On the other hand, it is also quite conceivable that these patterns reflect some secular trend in persistence and achievement patterns or differences in the scaling of the 11th grade CST test over time, and so do not result from the CAHSEE requirement. The data we have cannot distinguish between these two possibilities, but it is clear from Table 3 that there is no evidence of a positive effect of the CAHSEE requirement on those students we would most expect it to help—low-skill students who fail the CAHSEE in 10th grade. If the test motivated students to learn more, we would expect the coefficient on the interaction term to be positive in the achievement models, which it never is.[12]

As we describe above, the estimates in Table 3 allow us to compute estimated bounds of the effects of the CAHSEE requirement on student outcomes. The upper- and lower-bound estimates differ in whether we attribute all or none of the difference in outcomes among students who pass the CAHSEE in 10th grade to the CAHSEE policy or to other factors. Table 4 reports estimates of these upper- and lower-bounds for each of the outcomes and quartiles.

Table 4 here

[12] In analyses not shown here, we estimate models using the 11th grade Math CST score as an outcome. As noted above, the math scores are not ideal, since they depend on which math class a student is enrolled in. If the CAHSEE requirement causes students who fail the CAHSEE in 10th grade to be assigned to lower-level math courses than they might have been in the absence of the CAHSEE policy, then these students would take easier (that is, based on lower-level math skills) math CST exams in 11th grade than they would have in the absence of the policy. This would lead to an upward bias in the estimated effects of the CAHSEE policy on observed 11th grade math scores. Our estimates from the difference-in difference models using math scores show a pattern very similar to that of the ELA CST scores shown in Table 3: the coefficients on the cohort variable (row 1) are negative and significantly different than zero in the bottom two cohorts; the coefficients on the interaction term (row 3) are negative and not significantly different than zero. Given that these estimates are upper bounds (because they may include some upward bias), they indicate the CAHSEE policy had no positive effect on math scores for low-achieving students who failed the test.

Table 4 yields a lower-bound (in absolute value) estimate of the effect of the CAHSEE requirement on graduation of -11 percentage points in the bottom quartile and -3 percentage points in the second quartile. The upper bound estimates from these models are -17 and -5 percentage points, respectively. The lower bound estimates are statistically significant (see relevant entries in Table 3), so even our most conservative estimates indicate rather large and statistically significant effects of the CAHSEE requirement on graduation rates. Because the interaction terms for persistence and achievement are not statistically different than zero, our lower bound estimates of the effects of the CAHSEE requirement on persistence and achievement are essentially zero.

Differential Effects of the CAHSEE Policy by Student Demographics

The preceding suggests there are no or modest negative effects of the CAHSEE policy on student persistence and achievement and large negative effects of the policy on graduation rates for students in the bottom quartile of the 10^{th} grade test score distribution. We next investigate whether these effects differ by race/ethnicity, gender, poverty, and ELL-status. We estimate versions of model 3c in Table 2A that include interactions between a demographic category and the CAHSEE policy variable (we fit separate models to test for interactions by race, by gender, by free/reduced-price lunch eligibility, and by ELL status). We fit these models for each of the 4 outcomes and using both versions of the CST percentile metric. We show only one set of estimates here, however—those for graduation, using the cohort-specific 10^{th}-grade CST percentile—because for other outcomes we find no significant pattern of differences by race/ethnicity, sex, poverty, or ELL status.

Table 5 here

Table 5 shows large differences in the effect of the CAHSEE policy on graduation rates by race and by gender for students in the bottom quartile. For white students in the bottom quartile, the imposition of the CAHSEE requirement is associated with no change in graduation rates; for black, Hispanic, and Asian students in the bottom quartile, however, the CAHSEE leads to large decreases in graduation rates (-19, -15, and -17 percentage points for Black, Hispanic, and Asian students, respectively). Changes in graduation rates between cohorts are not significantly different by race in the higher quartiles (save the third quartile, where Asian graduation rates increase after the start of the CAHSEE policy).

There are also large differences in the change in graduation rates by gender for students in the bottom quartile. Graduation rates for girls in the bottom quartile decline by 19 percentage points, compared to a decline of 12 percentage points for boys, a difference that is statistically significant. In the second quartile, graduation effects also differ by gender: graduate rates decline by 6 percentage points for girls and are unchanged for boys. There are no gender differences in changes in graduation rates after the start of the CAHSEE policy for students above the second quartile.

There are no significant differences in CAHSEE effects by free/reduced-price lunch eligibility or ELL status in any quartile. Although we do not show the results here, there are also no significant differences by subgroup for the other outcomes (achievement and persistence in the 11th and 12th grade years).[13] Moreover, there are no sizeable or statistically significant differences among the four districts in the patterns of racial and gender effects (results not shown).

In sum, the results of our models of differential effects suggest large race and gender differences in the effects of the CAHSEE requirement on graduation for students in the bottom quartile. Girls and minority students appear to be adversely affected by the CAHSEE requirement

[13] 1 out of 48 statistical tests yields a $p<.05$, less than we would expect by chance.

much more than white and male students (controlling for 9th and 10th grade scores, free/reduced-price lunch status, and ELL status). This is a troubling finding and bears investigation.

VII. Explaining the Race and Gender Differences in CAHSEE Effects on Graduation

In this section we investigate several possible explanations for the differential race and gender effects of the CAHSEE policy. First, we investigate whether the pattern of racial effects results from differences in the schools attended by minority and white students. Second, we discuss the possibility that the differential effects result from racial or gender differences in tacit graduation requirements prior to the implementation of the CAHSEE, testing the argument that the CAHSEE may have had larger impacts on some groups because it effectively raised the bar for graduation more for some groups than for others. And third, we investigate whether the patterns of racial and gender effects result from a *stereotype threat* phenomenon (Steele & Aronson, 1995) that causes minority students and females to pass the CAHSEE exam at lower rates than white and male students of comparable demonstrated academic skill. We find the data refutes the first two of these hypotheses, but is highly consistent with the third.

Before discussing these different potential mechanisms, however, one additional piece of information is useful. Although large racial and gender differences in graduation are evident in Table 5, we find no evidence of racial or gender differences in graduation effects when we fit a version of Model (4) that includes interactions of the CAHSEE policy impact with race or gender (results not shown). That is, conditional on students' passing status on the CAHSEE in 10th grade, the effect of the CAHSEE graduation requirement does not differ by race or gender (or poverty or ELL status, for that matter). This finding implies that low-achieving girls and students of color graduate at lower rates because they pass the CAHSEE at lower rates than similarly low-achieving white and male students, not because failing the exam is more detrimental to these groups.

To test whether this conclusion is consistent with our data, we fit models predicting passing status on the CAHSEE in 10th grade, conditional on students' 8th, 9th, and 10th grade ELA and Math CST scores (see Table 6).[14] Table 6 indeed shows significant differences in CAHSEE passing status by race, gender, and ELL status among students in the bottom quartile of ELA skills, controlling for prior and current CST scores. Notably, passing rates for Black, Hispanic, and Asian students are 10-16 percentage points lower than those of observationally similar white students in the bottom quartile. For Asian students, the differential passing rates are driven by the ELA portion of the CAHSEE; for Black and Hispanic students, the differential passing rates result from differences on both the ELA and math CAHSEE tests. In addition, passing rates among girls are 5 percentage points lower than those of boys, a pattern driven by the math portion of the exam. Given that the passing rate among white, male, non-poor, non-ELL students in the bottom quartile is only 37%, these estimated differences in passing rates by race and gender are substantial. Moreover, the magnitude of these differences in passing rates is similar to the racial and gender differences in graduation effects (see Table 5). This similarity further supports the hypothesis that the racial and gender differences in CAHSEE effects are the result of differences in CAHSEE passing rates rather than differential effects of the policy.

Table 6 about here

We next consider the extent to which our results are consistent with each of the three potential explanations for the racial and gender differences in CAHSEE effects evident in Table 5: differential school quality; differential graduation requirements; and differential performance on the CAHSEE exam.

---

[14] We use only the 2004 and 2005 10th grade cohorts here in order to include three years of CST scores (the CST tests were not taken by the 2003 cohort in 8th grade). In addition, we include controls for math CST scores as well as for ELA CST scores, despite the fact that the math CST tests taken by a given student in 8th through 10th grade depend on which math course a student is enrolled in, making math CST scores not strictly comparable across students. We do so to control as well as possible for prior academic skills, and a noisy measure of math skill is likely better than no measure at all. In addition, we limit the sample to students in the bottom quartile of 9th grade ELA scores (rather than 10th grade ELA scores as we have done above) because we want to make sure we are conditioning on a sample defined prior to the administration of the CAHSEE test.

Differential School Quality. It is possible that minority students experience the most negative CAHSEE effects because these students are concentrated in schools that less successfully prepare low-achieving students to succeed on the CAHSEE (perhaps because these schools do not provide good remediation, CAHSEE preparation, or other supports). To investigate the differential school quality hypothesis, we conduct two sets of analyses. First, we fit a series of models like those in Table 5 that include a set of school-by-policy fixed effects.[15] These models effectively compare the effects of the CAHSEE on students of different racial/ethnic groups who attend the same schools. If minority students are concentrated in schools where the effect of the CAHSEE on graduation is particularly large and negative for all students in the school, then models that include school-by-policy fixed effects should substantially attenuate the negative coefficients on the race-by-policy interaction variables in Table 5. While this may appear a plausible explanation for CAHSEE effects on minority students, the differential school quality hypothesis cannot account for the gender differences in CAHSEE effects, because male and female students are generally equally distributed across schools.

Table 7 about here

When we add school-by-policy fixed effects to the model, the racial (and gender) differences in CAHSEE effects do not change at all (Table 7). This implies that the effect of the CAHSEE policy on graduation differs sharply even between white and minority students who attend the same school. Thus, the racial differences in CAHSEE effects cannot readily be attributed to differences in some aspect of school quality that reduces graduation rates in the presence of the policy. As a second test of the differential school quality hypothesis, however, we add school fixed effects to the models shown in Table 6 above (the models predicting CAHSEE passing status, conditional on

[15] That is, for each school *j*, we include two dummy variables in the model, one indicating that an observation comes from a student in 10th grade in 2003 (and who therefore is not subject to the CAHSEE requirement) in school *j*, and one indicating that an observation comes from a student in 10th grade in 2004 or 2005 (and who therefore is subject to the CAHSEE requirement) in school *j*. The inclusion of these dummy variables controls for any average change in outcomes between pre- and post-CAHSEE policy cohorts of students within the same school.

students' prior skills). Again we find no difference in the results; the racial (and gender) differences in passing rates are nearly identical to those shown in Table 6 when we compare students within the same schools (results not shown). This implies that the racial differences in passing rates cannot be attributed to differences in schools' ability to prepare low-achieving students to pass the CAHSEE in 10th grade.

Our results here do not imply there are no differences in any aspect of school quality among the schools attended by white and minority students (our analyses cannot test this possibility). School quality differences may play a role in producing the racial differences in average student skills (as measured by the CST tests) in 10th grade, but this is not a factor in our differential results, because our models show differential results by race, conditional on students' 10th grade (and prior) academic skills. Moreover, the results in Table 7 do not rule out the possibility that differential treatment of minority and white students within the same schools produces the pattern of differential effects by race and gender—if bottom quartile whites and males in a school received better instruction, remediation, support for passing the CAHSEE than similar minority and female students within the same school, this might produce differential effect patterns even among students in the same school. Nonetheless, none of our results yield any support for the hypothesis that between-school differences in educational environments causes the sizeable racial differences in CAHSEE policy effects.

<u>Implicit Differential Graduation Requirements.</u> Another possible explanation for the large racial and gender differences in CAHSEE effects on graduation rates is that, prior to the CAHSEE requirement, low-achieving minority students and girls were informally held to a lower standard for graduation than were similar white students and boys. For example, these students may have been counseled into lower-level classes that were easier to pass, and could therefore accumulate credits required for graduation but not acquire the same level of academic skills as their peers.

According to this argument, the CAHSEE requirement, by standardizing one aspect of graduation requirements, would have increased the bar for graduation for low-achieving minority and female students more than for similar white and male students. Several pieces of evidence argue against this hypothesis, however. First, our estimates condition on students' 9th and 10th grade academic skill, so we would need to believe there were different implicit pre-CAHSEE graduation requirements by race and gender for students with the same level of academic skill. Second, because we find no evidence that the differential effects are driven by between-school differences in effects, we would need to believe that there were racial and gender differences in graduation requirements for students within the same school, which is less likely than that there are differences in graduation requirements among schools. Third, if the effects were driven by differences in implicit graduation requirements, we would also expect these differences to show up as racial and gender differences in the effect of the policy on those who fail the test relative to those who pass the test, but we find no such racial or gender differences (as noted above). And finally, the differential requirement hypothesis cannot explain the sharp differences in passing rates on the CAHSEE. In sum, none of our evidence suggests that the observed pattern of effects results from differences in graduation requirements in effect prior to the CAHSEE policy.

Stereotype Threat. Table 6 shows that there are relatively large racial and gender differences in students' success at passing the CAHSEE in 10th grade, even conditional on a large vector of students prior scores on the CST tests (which are aligned to the same state standards as the CAHSEE). One possible explanation for such differences is the social psychological phenomenon known as *stereotype threat* (Steele & Aronson, 1995). The literature on stereotype threat argues that, in a situation where an individual's performance has the potential to confirm a negative stereotype about his or her group, individuals experience performance-related stress or anxiety, leading their performance to be biased in the direction of that stereotype (Steele, 1997; Steele & Aronson, 1995). In other words, students may underperform (relative to their true ability) if they feel under pressure to disprove a negative stereotype about their own group, such as “girls are bad at math” or “black students are not as smart as white students.” In Steele & Aronson’s (1995) seminal work, black students led to believe that an exam measured innate ability performed worse than those told they were simply taking a difficult test; white students, not facing a stereotype threat (because there is no negative societal stereotype regarding white students’ cognitive skill), performed the same under both conditions. Steele & Aronson lifted the threat simply by assuring students that the exam did not measure ability and therefore, implicitly, could not prove the negative stereotype. Although there is no consensus on the mechanism through which stereotype threat operates (Smith, 2004), recent work suggests that anxiety (Murphy, Steele, & Gross, 2007; Osborne, 2007) and/or stress (Blascovich et al., 2001) may play a key mediating role.

Research suggests that the effects of stereotype threat vary according to the salience of the at risk identity(Aronson et al., 1999; Gonzales, Blanton, & Williams, 2002; Keller, 2002; Shih, Pittinsky, & Ambady, 1999), what the student believes the exam measures (Brown & Josephs, 1999; Steele & Aronson, 1995), how difficult the exam is for them (Ben-Zeev, Fein, & Inzlicht, 2005; O'Brien & Crandall, 2003; Spencer, Steele, & Quinn, 1999) and how important the skill being measured is to the at-risk student’s self-perception (Aronson et al., 1999; Steele, 1997). The latter two factors are of particular relevance here.

The negative effects of stereotype threat tend to be strongest for those with the highest "domain identification”—that is, those to whom the exam is most important (Aronson et al., 1999; Keller, 2002; Steele, 1997). Steele in particular argues that domain identification is a measure of how important the skill being tested is to an at-risk student’s self-perception. Aronson et al. (1999) argue that domain identification may, more broadly, indicate that stereotype threat will be activated when it has repercussions for the at-risk student: “It therefore may be more correct to say that high motivation—a sense that something important is at stake—is the necessary factor in stereotype threat, not high identification per se" (Aronson et al., 1999, p. 43). Brown & Josephs (1999) demonstrated that the more a test outcome is associated with a clear statement about success or failure, the more it may impact performance. They demonstrated that women concerned with being identified as especially bad at math underperformed when told that a test identified only students who were particularly bad at math. In contrast, women told that the exam solely identified students who are especially strong in math did not underperform. This concern over the repercussions of an exam with high personal stakes and clearly demarcated success/failure results has direct relevance to the CAHSEE. We might expect negatively stereotyped students to underperform on the CAHSEE precisely because their own graduation status is at stake. In contrast, we would expect no (or less) underperformance on the CST tests, which do not carry direct consequences for students.

Although Steele and Aronson argue that stereotype threat is most likely to hurt the highest performing students (Steele, 1997; Steele & Aronson, 1995), others have argued that Steele's findings instead suggest that those "working at the frontier of their skills and knowledge" are at greatest risk for this phenomenon (Ben-Zeev et al., 2005, p. 175). In other words, those who are performing difficult tasks relative to their own skill, regardless of where they fall on the achievement spectrum, are susceptible to stereotype threat. Stereotype threat experiments that vary the difficulty of the task find that at-risk students perform better when performing easy tasks, but worse when performing difficult ones (Ben-Zeev et al., 2005; O'Brien & Crandall, 2003; Spencer et al., 1999; Walton & Spencer, in press). Given that the CAHSEE exam is relatively easy (68% of students pass both sections on the first attempt), we might expect stereotype threat to primarily affect the CAHSEE performance of low-achieving students, for whom the CAHSEE represents a significant challenge (81% of bottom quartile students fail the CAHSEE on their first attempt; see Table 4).

Although stereotype threat has been studied extensively in laboratory experiments, there is little evidence of the extent to which stereotype threat affects students in "real-world" situations. Most field work has not demonstrated stereotype threat directly, but instead focused on interventions to lift it for threatened students. In fact, a number of studies have found that techniques such as mentoring can significantly reduce the effects of stereotype threat (Good, Aronson, & Inzlicht, 2003; Walton & Spencer, in press). One experimental field study found that asking students to report their gender and ethnicity before the AP Calculus AB exam had a negative effect on women's scores, though the significance of this finding is subject to interpretation (Danaher & Crandall, 2008; Stricker & Ward, 2004).

Our data provide an opportunity to test whether stereotype threat affects students' performance on the CAHSEE. Prior research on stereotype threat suggests that it should be activated more by the CAHSEE, which carries high-stakes for the student, than the low-stakes CST tests, and that it should be activated more for low-achieving students than for high-achieving students. Moreover, the theory predicts a clear pattern of expected results: we would see lower than expected performance by black, Hispanic (and perhaps Asian)[16] students on the CAHSEE ELA test, and lower than expected performance by black, Hispanic, and female students on the CAHSEE math test. These patterns should be strongest for low-achieving students, for whom the content of the CAHSEE exam poses a substantial challenge.

This explanation fits the pattern of results described in Table 6 above. However, we provide a more careful test here by 1) using continuous CAHSEE scores rather than the binary pass/fail status used in Table 6; 2) comparing CAHSEE performance to contemporaneous CST performance; and 3) estimating the models for all four quartiles of students. We first estimate racial and gender

[16] The category "Asian American" is broad and diverse enough, especially in California, to belie the "model minority" stereotype. See Rumbaut & Portes (2001) for a thorough discussion of ethnicity and academic achievement. In particular, Asian students may feel at risk on the ELA exams, either due to recent arrival in the US and/or a concern over being perceived as not "fully American" (Cheryan & Monin, 2005).

differences in 10th grade ELA CST scores, conditional on 8th, 9th, and 10th grade ELA and math CST scores. We then estimate the corresponding differences in 10th grade ELA CAHSEE scores, using the same sample and model. We standardize both the CST and CAHSEE scores so that the differences are expressed in the same metric. If stereotype threat is activated for stereotyped groups by the high-stakes CAHSEE exam more than by the low-stakes CST exam, then the difference in race or gender gaps on the CAHSEE and the CST should be negative (implying that the stereotyped groups perform less well, on average, on the high-stakes CAHSEE test than we would predict on the basis of their prior on contemporaneous low-stakes CST test scores). We do the same for math scores as well. These difference-in-difference estimates indicate the extent of underperformance on the high-stakes CAHSEE test relative to the low-stakes CST test. One nice feature of these models is that the difference-in-difference estimates eliminate all potential bias due to measurement error in the prior CST scores.[17] Table 8 reports the results of these models for students in the bottom quartile of 9th grade ELA skill.

Table 8 about here

Table 8, like Table 6, shows results consistent with the stereotype threat hypothesis. In the ELA tests, racial differences are significantly larger on the 10th grade CAHSEE test than on the 10th grade CST (the difference-in-difference estimates range from -0.13 to -0.19 standard deviations). In the math tests, black-white and Hispanic-white differences are -0.19 and -0.14 standard deviations larger on the high stakes CAHSEE than on the low-stakes CST, but Asian-white differences are not significantly different on the two tests—in keeping with the lack of a negative stereotype about Asian students' math skills. In addition, the gender gap is significantly larger on the CAHSEE than

[17] For example, note that bottom-quartile black students perform .09 standard deviations lower on the 10th grade ELA CST test than observationally similar white students. However, given that there is measurement error in the 8th and 9th grade CST test scores used as control variables in these models, and that white students have higher average ELA skill than black students in the bottom quartile, we would expect that this estimate is biased downward from its true value. Because the same bias will affect the estimated black-white difference in CAHSEE scores (because the models are estimated on the same sample, using the same control variables), the difference-in-difference models will subtract off any such bias.

on the CST—again consistent with the stereotype threat prediction. There is no evidence of differential performance by free/reduced-price lunch eligibility, which is consistent with the fact that lunch eligibility status is not a readily visible characteristic, making it less susceptible to stereotype threat. Finally, there is a significantly larger gap between ELL and non-ELL students on the CAHSEE than the CST test. While this difference is larger on the ELA tests than the math tests—consistent with stereotype threat, which would predict larger effects on ELA performance for ELL students—it may also result from the fact that the CAHSEE includes a writing component and the CST does not, which would lead to lower performance by ELL students on the CAHSEE ELA test than their CST scores would predict.

Table 9 reports the difference-in-difference estimates for the ELA and math tests for each quartile of students. The pattern of results again largely corresponds to stereotype threat predictions—the difference-in-difference estimates are generally less negative (or even become positive) in the upper quartiles than the bottom quartile, though this is more true in the ELA results than in the math results.

Table 9 about here

The pattern of test score results closely mirrors those predicted by stereotype threat: minority students and girls do worse on the high-stakes CAHSEE exam in 10th grade than we would predict on the basis of their prior and current academic skill as measured by low-stakes CST tests in 8th, 9th, and 10th grade. As a final test of whether this might explain the differential passing rates on the 10th grade CAHSEE, we use the estimates from Table 9 to compute a "stereotype threat-adjusted" estimate of what each student's CAHSEE scores would be if there were no high-stakes stereotype threat effects. To do this, we use the difference-in-difference estimates in Table 9, multiply them by the vector of dummy variables indicating each student's demographic characteristics, and subtract the resulting number from the observed CAHSEE score. This process produces a "stereotype-threat-adjusted" CAHSEE score for each student by adding to each student's observed CAHSEE scores the amount by which our models suggest they were affected by stereotype threat. We then determine whether each student would have passed the CAHSEE in 10th grade had he or she received the stereotype-threat-adjusted math and ELA CAHSEE scores. This procedure does not eliminate all average racial and gender differences in CAHSEE scores, because there are racial and gender differences in 10th grade CST scores, conditional on prior scores (see Table 8); it only eliminates the portion of the differences in CAHSEE scores that we attribute to differences in scores between the high- and low-stakes tests. When we refit the models shown in Table 6 above (the models predicting passing rates) using this estimated passing status as the outcome, we no longer find any substantial differences in passing rates by race, gender, free/reduced-price lunch status, or ELL status (see Table 10). In other words, the racial and gender differences in CAHSEE passing rates can be fully explained by racial and gender differences in high- versus low-stakes test performance. The data are fully consistent with a stereotype threat account of differential performance of minorities and girls on high-stakes tests, leading to lower passing rates and, ultimately, to lower graduation rates for these students.

Table 10 about here

VIII. Discussion and Implications

Our analysis of student data from four large California school districts yields several key findings. First, the implementation of the exit exam requirement in California appears to have no positive effects (and may have negative effects) on student achievement. In addition the requirement has a large negative effect on graduation rates of students with low 10th grade academic skills. Our estimates suggest that 3.6-4.5 percent of all high school students fail to receive a diploma as a result of the exit exam requirement.[18] Given enrollments of roughly 500,000 10th grade students per cohort in California, this implies that between 18,000 and 22,500 California students per year fail to receive diplomas who would have received them had the exit exam not been required.[19]

Second, the effect of the CAHSEE requirement on graduation rates appears to be driven entirely by its effect on minority students. The great majority of the students failing to receive a diploma as a result of the exit exam are students of color. Graduation rates declined by 15 to 19 percentage points for low achieving black, Hispanic, and Asian students when the exit exam was implemented, and declined only one percentage point (not statistically significant) for similar white students. In addition, the negative effect of the exit exam on graduation rates was almost twice as large for girls as for boys. These differential effects are particularly troubling.

We are able rule out two explanations for these differential effects. First, we show that the differential results do not arise because the negative effects of the CAHSEE requirement are larger in schools attended by minority students. Second, we show that the differential effects of the policy are not due to a correction of lower implicit graduation requirements for girls and minority students. Instead, our results are consistent with a stereotype threat explanation. Minority students and girls are less likely than white students with similar patterns of scores on their (low-stakes) 8th-, 9th-, and 10th-grade CST scores to pass the high-stakes CAHSEE exam. And it is this stereotype-threat induced failure that appears to be responsible for the large racial and gender differences in the effects of CAHSEE requirement.

It is important to be clear that our conclusions regarding the role of stereotype threat in

---

[18] We compute this as follows. Using the lower-bound estimates from Table 3, and averaging over the four quartiles of students, we compute an overall effect on graduation rates of $0.112 \cdot 0.25 + 0.027 \cdot 0.25 + 0.005 \cdot 0.25 = 0.036$ (3.6 percentage points). Using the upper-bound estimates from Table 2A, we compute an overall effect on graduation rates of $0.147 \cdot 0.25 + 0.034 \cdot 0.25 = 0.045$ (4.5 percentage points).
[19] Note that this figure is very similar to that reported by state reports on the exit exam (Becker, Wise, & Watters, 2008).

producing the pattern of CAHSEE effects do not imply that the CAHSEE test is racially or gender biased. The CAHSEE exams were subjected to standard statistical tests to identify items that showed evidence of bias (i.e., differential item functioning tests were used to identify items that minority students or females, answered less often correctly than white students or males with the same level of skill as measured by all other items on the test), and such items were eliminated (Wise et al., 2000). Stereotype threat can occur even on a perfectly unbiased test (Walton & Spencer, in press).

Rather, women and minority students experience a confluence of factors that negatively affect their performance. First, they experience stress from two sources: fear of failing the test and concern about proving a negative stereotype. Second, these students are working on a task that taxes their ability and has high-stakes: low-performing students not under stereotype threat only pass the exam 37% of the time. While all low-performing students must grapple with concern over failing and the difficulty of the exam, only some students deal with the additional burden of stereotype threat.

Although the high-stakes nature of the test and the existence of negative societal stereotypes are not in the power of schools to change, researchers have found, in a few studies, evidence suggesting that the effects of stereotype threat may be able to be ameliorated or eliminated through classroom interventions to reduce the salience of race or gender in the testing situation and by teaching students that intelligence and academic skills are malleable characteristics rather than fixed traits (Cohen et al., 2006; Danaher & Crandall, 2008; Good et al., 2003; Stricker & Ward, 2004; Walton & Cohen, 2007; Walton & Spencer, in press). A forthcoming meta-analysis of classroom interventions by Walton and Spencer (in press), for example, finds that African American students receiving an intervention to eliminate stereotype threat outperformed white students of similar prior demonstrated ability by .17-.18 of a standard deviation, on average.

Certainly, however, more research is needed to test whether school and classroom interventions can effectively eliminate stereotype threat on high-stakes exams such as the CAHSEE.

Implications

These findings raise questions about the usefulness of exit exams as a method to improve student outcomes. If the goal of high school exit exams is to improve achievement of low-achieving students, there is no evidence the California high school exit exam has succeeded in this, at least among the first two cohorts subject to the exam. If however, the goal of the high school exit exam is to ensure that a high school diploma means something—that it signals to potential employers that graduates have a certain level of basic skills—then perhaps one might argue that reducing the number of students receiving diplomas is a necessary byproduct of such a policy. Put differently, if the goal of the policy is to distinguish between students with and without a basic level of skills, then necessarily some students should be denied a diploma. In order to accept this rationale, however, it is essential that the exit exam provide an unbiased signal, meaning the same thing regardless of race or gender. Given the clear evidence of differential performance on the CAHSEE by race and gender, even among students with the same prior and contemporaneous scores on low-stakes math and ELA tests, however, it is clear that the CAHSEE exam does not convey the same signal for minority and white students or for boys and girls. Our stereotype threat estimates imply that minority students require, on average, skills one sixth of a standard deviation higher than white students in order to pass the CAHSEE; girls require math skills one twelfth of a standard deviation higher in math than boys to pass the CAHSEE. If it takes a higher level of skill for minority and female students to pass a test than it does for white and male students, then the test cannot be said to convey a fair signal of students' level of basic skills. Therefore, if exit exam policies like California's are to be retained, it is imperative that they be accompanied by serious efforts to ameliorate their negative effects on minority students and girls. Such efforts might include interventions to reduce the activation of stereotype threat (see Cohen et al., 2006) or compensatory interventions designed to improve the performance of minority students and girls on the exams.

This study provides no evidence that the CAHSEE exam policy as currently implemented has any benefits for students. It does not serve students well, and appears to have sharply inequitable effects. Moreover, California, like the twenty-plus other states in the U.S. with exit exam policies, spends millions of dollars and a considerable amount of instructional time on exit exam test preparation, administration, and remediation. Our analysis suggests that, to date, this has been neither money nor time well spent.

Table 1: Selected Student Demographics, by Tenth Grade Cohorts

			Race/ Ethnicity				Student Designations		Gender
		Total	White	Asian	Black	Hispanic	ELL	Free/ Reduced	Female
		N	%	%	%	%	%	%	%
4 Districts	2003	20,985	22.2%	28.0%	13.5%	36.3%	19.9%	56.9%	51.4%
	2004	21,739	21.1%	28.0%	13.4%	37.5%	19.0%	60.5%	51.2%
	2005	21,502	20.6%	27.9%	13.9%	37.7%	18.2%	64.5%	51.5%

		Prior Achievement Measures							
			9th Grade English Language Arts CST Score				10th Grade English Language Arts CST Score		
		Mean	25th %ile	50th %ile	75th %ile	Mean	25th %ile	50th %ile	75th %ile
4 Districts	2003	329.3	290	329	366	326.6	288	325	364
	2004	341.1	304	342	380	332.0	293	329	371
	2005	339.2	300	337	378	332.5	289	332	372
State	2003	321.4	276	317	360	324.0	282	322	364
	2004	332.5	287	331	374	328.1	283	324	368
	2005	330.6	287	328	372	328.4	282	327	372

		Outcome Descriptives							
		Present in Spring of 11th Grade		Present in Spring of 12th Grade		11th Grade ELA CST Score		Graduation by Spring of 12th Grade	
		%	N	%	N	Mean	N	%	N
4 Districts	2003	88.8%	20985	78.2%	20985	328.9	17548	70.1%	14940
	2004	88.5%	21739	77.4%	21739	331.4	18190	68.0%	15233
	2005	88.1%	21502	77.0%	21502	332.0	18060	67.0%	15131

Note: Students are assigned to a cohort based on the calendar year in which they first took the CAHSEE. For 99.4 percent of the sample, this occurred in Spring of 10th grade. However, in limited cases, students took the CAHSEE in an earlier grade level. Usually this occurred when a student was retained in 9th grade for a second year. Students are designated as "ELL" if they were not a native English speaker, were not initially fluent in English, and had not been redesignated fluent in Spring of 10th grade. Students who are designated special education status are not included in these analyses. A student was categorized as a special education student if he or she had ever received special education services or had been designated special education status at any point after 8th grade. Percentages of students designated eligible for the free or reduced lunch program are indicated above, however these percentages only pertain to three of the four districts in which this information was provided. One of the four districts also did not provide graduation status outcomes for its students, and therefore graduation rates are reported only among students in the other three districts. Since students are defined by their 10th grade cohort, all students in the analytic sample have 10th grade ELA CST scores. However, not all members of a given cohort have 9th grade ELA CST scores. In cohort 2003, 87.1% of the students have 9th grade ELA CST scores, whereas 89.4% and 89.6% of students in cohorts 2004 and 2005, respectively, have 9th grade ELA CST scores.

Table 2A: Estimated Change in Persistence, Achievement, and Graduation between Cohort 2003 and 2004/05, by Quartile, Using Cohort-Specific 10th-Grade CST percentiles

	Estimated Difference 2003 to 2004/05		
	Model 3a	Model 3b	Model 3c
Quartile 1			
Presence, Spring 11	-0.010	-0.023 *	-0.023 *
	(0.012)	(0.009)	(0.009)
Presence, Spring 12	-0.031 *	-0.047 ***	-0.047 ***
	(0.012)	(0.010)	(0.010)
ELA CST 11 Score	-4.985 ***	-5.951 ***	-6.109 ***
	(1.024)	(0.951)	(0.988)
Graduation by Spring 12	-0.135 ***	-0.149 ***	-0.147 ***
	(0.018)	(0.016)	(0.016)
Quartile 2			
Presence, Spring 11	-0.011	-0.012	-0.012
	(0.008)	(0.007)	(0.007)
Presence, Spring 12	-0.021 *	-0.023 **	-0.023 **
	(0.009)	(0.008)	(0.008)
ELA CST 11 Score	-3.218 ***	-4.497 ***	-4.596 ***
	(0.937)	(1.087)	(1.090)
Graduation by Spring 12	-0.025 *	-0.035 ***	-0.034 ***
	(0.010)	(0.009)	(0.009)
Quartile 3			
Presence, Spring 11	-0.003	-0.003	-0.002
	(0.005)	(0.005)	(0.005)
Presence, Spring 12	-0.003	-0.005	-0.004
	(0.007)	(0.006)	(0.006)
ELA CST 11 Score	1.393	1.294	1.277
	(1.061)	(1.101)	(1.106)
Graduation by Spring 12	-0.003	-0.003	-0.002
	(0.007)	(0.007)	(0.007)
Quartile 4			
Presence, Spring 11	-0.007	-0.005	-0.005
	(0.004)	(0.003)	(0.003)
Presence, Spring 12	-0.013	-0.010	-0.010
	(0.008)	(0.006)	(0.006)
ELA CST 11 Score	10.369 ***	10.665 ***	10.653 ***
	(1.081)	(1.058)	(1.052)
Graduation by Spring 12	0.009	0.013	0.013
	(0.024)	(0.025)	(0.025)
Covariates Included		x	x
District Fixed Effects			x

Note: Model 3a estimates average difference in outcome between the 2003 10th grade cohort and the combined 2004 and 2005 cohorts for a student at the midpoint of the corresponding quartile of the 10th-grade ELA CST score distribution. Model 3b is identical, but includes a set of covariates, including race/ethnicity, gender, free/reduced-price lunch status, ELL status, and 9th-grade ELA CST scores. Morel 3c adds district fixed effects to model 3b. Standard errors are corrected for clustering of students within schools. * p<.05; ** p<.01; *** p<.001.

Table 2B: Estimated Change in Persistence, Achivement, and Graduation between Cohort 2003 and 2004/05, by Quartile, Using 2004-normed 10th-Grade CST percentiles

	Difference 2003 to 2004/05		
	Model 3a	Model 3b	Model 3c
Quartile 1			
Presence, Spring 11	-0.006	-0.026 **	-0.027 **
	(0.011)	(0.009)	(0.009)
Presence, Spring 12	-0.027 *	-0.054 ***	-0.054 ***
	(0.012)	(0.011)	(0.011)
ELA CST 11 Score	-4.789 ***	-8.636 ***	-8.834 ***
	(1.012)	(0.918)	(0.958)
Graduation by Spring 12	-0.129 ***	-0.164 ***	-0.162 ***
	(0.017)	(0.016)	(0.016)
Quartile 2			
Presence, Spring 11	-0.014 *	-0.017 **	-0.017 *
	(0.007)	(0.006)	(0.006)
Presence, Spring 12	-0.028 **	-0.034 ***	-0.033 ***
	(0.009)	(0.009)	(0.009)
ELA CST 11 Score	-7.188 ***	-10.507 ***	-10.617 ***
	(0.909)	(1.067)	(1.071)
Graduation by Spring 12	-0.040 ***	-0.056 ***	-0.056 ***
	(0.009)	(0.009)	(0.008)
Quartile 3			
Presence, Spring 11	-0.011 *	-0.009	-0.009
	(0.005)	(0.005)	(0.005)
Presence, Spring 12	-0.016 *	-0.017 *	-0.016 *
	(0.007)	(0.006)	(0.006)
ELA CST 11 Score	-4.557 ***	-6.994 ***	-7.021 ***
	(1.120)	(1.196)	(1.205)
Graduation by Spring 12	-0.018 *	-0.018 *	-0.017 *
	(0.008)	(0.007)	(0.007)
Quartile 4			
Presence, Spring 11	-0.010 *	-0.008	-0.008
	(0.005)	(0.004)	(0.004)
Presence, Spring 12	-0.016 *	-0.014 *	-0.014 *
	(0.008)	(0.006)	(0.006)
ELA CST 11 Score	1.910	1.589	1.591
	(1.055)	(1.126)	(1.136)
Graduation by Spring 12	0.007	0.011	0.011
	(0.029)	(0.031)	(0.031)
Covariates Included		x	x
District Fixed Effects			x

Note: Model 3a estimates average difference in outcome between the 2003 10th grade cohort and the combined 2004 and 2005 cohorts for a student at the midpoint of the corresponding quartile of the 10th-grade ELA CST score distribution. Model 3b is identical, but includes a set of covariates, including race/ethnicity, gender, free/reduced-price lunch status, ELL status, and 9th-grade ELA CST scores. Morel 3c adds district fixed effects to model 3b. Standard errors are corrected for clustering of students within schools. * p<.05; ** p<.01; *** p<.001.

Table 3: Difference-in-Difference Estimates of the Effects of CAHSEE Policy

	Quartile 1 (low)	Quartile 2	Quartile 3	Quartile 4 (high)
Presence Spring 11				
Cohort 2004 or 2005	-0.044 ***	-0.018 **	-0.005	-0.004
	(0.013)	(0.007)	(0.005)	(0.003)
Fail 10th Grade CAHSEE	-0.038 **	-0.027 ***	-0.011	0.041 *
	(0.012)	(0.007)	(0.010)	(0.016)
Fail x Cohort 2004/05	0.017	0.004	-0.011	-0.093 **
	(0.014)	(0.008)	(0.014)	(0.032)
Constant	0.897 ***	0.938 ***	0.959 ***	0.969 ***
	(0.021)	(0.011)	(0.008)	(0.007)
Adjusted R^2	0.044	0.021	0.014	0.013
N	8775	14110	15737	15259
Presence, Spring 12				
Cohort 2004 or 2005	-0.050 *	-0.024 *	-0.007	-0.013 *
	(0.022)	(0.012)	(0.007)	(0.006)
Fail 10th Grade CAHSEE	-0.067 ***	-0.043 **	-0.049 *	-0.076
	(0.019)	(0.013)	(0.020)	(0.061)
Fail x Cohort 2004/05	-0.012	-0.014	-0.019	-0.031
	(0.022)	(0.014)	(0.023)	(0.072)
Constant	0.723 ***	0.817 ***	0.846 ***	0.911 ***
	(0.032)	(0.026)	(0.037)	(0.038)
Adjusted R^2	0.061	0.035	0.031	0.039
N	8775	14110	15737	15259
ELA CST 11 Score				
Cohort 2004 or 2005	-5.585 *	-4.918 ***	1.189	10.681 ***
	(2.511)	(1.280)	(1.203)	(1.093)
Fail 10th Grade CAHSEE	-9.784 ***	-8.116 ***	-7.670 ***	-1.679
	(2.281)	(1.156)	(1.336)	(4.609)
Fail x Cohort 2004/05	-0.947	-0.654	-2.673	-11.232
	(2.530)	(1.412)	(2.085)	(6.646)
Constant	274.584 ***	301.595 ***	334.977 ***	388.373 ***
	(3.012)	(1.812)	(1.589)	(2.217)
Adjusted R^2	0.135	0.199	0.248	0.388
N	6881	12270	14430	14367
Graduation by Spring 12				
Cohort 2004 or 2005	-0.063 *	-0.024 *	-0.009	0.008
	(0.026)	(0.012)	(0.008)	(0.026)
Fail 10th Grade CAHSEE	-0.088 ***	-0.094 ***	-0.092 ***	-0.153 *
	(0.025)	(0.014)	(0.020)	(0.069)
Fail x Cohort 2004/05	-0.138 ***	-0.058 ***	-0.047 *	-0.026
	(0.032)	(0.015)	(0.022)	(0.079)
Constant	0.627 ***	0.726 ***	0.803 ***	0.886 ***
	(0.032)	(0.031)	(0.038)	(0.040)
Adjusted R^2	0.121	0.073	0.041	0.037
N	8775	14110	15737	15259

Note: Estimates are based on a model like model 3b in Table 2A, with the addition of the "Fail 10th-grade CAHSEE" variable and its interaction with the policy cohort indicator variable. Models estimates average difference in outcome for a typical student at the midpoint of the relevant quartile of the 10th-grade ELA CST score distribution. Quartiles are defined using the cohort-specific percentile measure, though results are substantively identical in models (not shown) using the 2004-cohort-based percentiles. Standard errors are corrected for clustering of students within schools. * p<.05; ** p<.01; *** p<.001.

Table 4: Estimated Upper- and Lower- Bounds of CAHSEE Effects on Graduation, by Quartile

	Quartile 1 (low)	Quartile 2	Quartile 3	Quartile 4 (high)
Change in Graduation Rate for those who Initially *Passed* CAHSEE	-0.063 *	-0.024 *	-0.009	0.008
Change in Graduation Rate for those who Initially *Failed* CAHSEE	-0.201 ***	-0.082 ***	-0.056 *	-0.018
Proportion of Students Who Initially Failed CAHSEE in Spring 10	0.809	0.468	0.104	0.011
Upper Bound Estimate of CAHSEE Effect on Graduation Rate [(row 1)*(1-row 3) + (row 2)*(row 3)]	-0.175 ***	-0.051 ***	-0.014	0.008
Lower Bound Estimate of CAHSEE Effect on Graduation Rate [(row 4) - (row 1)]	-0.112 ***	-0.027 ***	-0.005	0.000

Note: Estimated changes in graduation rates for passers and failers computed from estimates in Table 3. Proportion of students failing the exam in 10th grade refers to students failing at least one portion of the exam in 10th grade. Denominator is all students with valid CAHSEE scores in 2003-2005 cohorts in our sample. Upper and lower bound estimates refer to upper- and lower-values in absolute value. * p<.05; ** p<.01 *** p<.001.

Table 5: Effects of CAHSEE Policy on Graduation Rate, by Demographic Characteristics and Quartile of 10th-Grade ELA CST Score

	Quartile of 10th Grade ELA CST			
	Quartile 1 (low)	Quartile 2	Quartile 3	Quartile 4 (high)
By Race/ Ethnicity Group				
White	-0.011	-0.032	-0.013	-0.010
	(0.034)	(0.029)	(0.016)	(0.010)
Black	-0.191 ***	-0.050 *	-0.020	0.051
	(0.036)	(0.020)	(0.020)	(0.044)
Hispanic	-0.146 ***	-0.027	-0.015	0.026
	(0.021)	(0.014)	(0.012)	(0.050)
Asian	-0.173 ***	-0.038 **	0.025 *	0.025
	(0.029)	(0.012)	(0.011)	(0.037)
N	9553	14694	16159	15496
p(F-test)	0.001	0.825	0.031	0.428
By Gender				
Female	-0.193 ***	-0.055 ***	-0.004	0.004
	(0.024)	(0.013)	(0.010)	(0.025)
Male	-0.116 ***	-0.011	-0.001	0.024
	(0.014)	(0.014)	(0.011)	(0.028)
N	9553	14694	16159	15496
p(F-test)	0.000	0.023	0.846	0.074
By English Language Learner Status				
ELL Students	-0.165 ***	-0.042 *	0.006	-0.132
	(0.026)	(0.017)	(0.034)	(0.106)
Non-ELL Students	-0.130 ***	-0.031 **	-0.003	0.013
	(0.019)	(0.011)	(0.007)	(0.026)
N	9553	14694	16159	15496
p(F-test)	0.285	0.602	0.800	0.189
By Free/Reduced-Price Lunch Eligibility Status				
Eligible	-0.176 ***	-0.040 **	0.000	0.056
	(0.021)	(0.013)	(0.011)	(0.075)
Not Eligible	-0.142 **	-0.041	0.007	0.018
	(0.045)	(0.023)	(0.012)	(0.026)
N	6135	9642	10047	9222
p(F-test)	0.471	0.982	0.607	0.461

Note: Each panel represents estimates from a different model. All estimates from models that correspond to model 3c in Table 2A, with interactions between the specific demographic group variables and the policy-cohort indicator variable. The p-value at the bottom of each panel is the p-value from the F-test of the null hypothesis that the coefficiennts are the same for each of the specified demographic groups. Standard errors corrected for the clustering of students within schools are shown in parentheses under each coefficient. * p<.05; ** p<.01; *** p<.001.

Table 6: Estimated Associations Between Student Characteristics and CAHSEE Passing Rates, by CAHSEE Test, Students in Bottom Quartile of 9th Grade ELA Skill

	ELA	Math	Both
Black	-0.112 ***	-0.170 ***	-0.156 ***
	(0.029)	(0.030)	(0.025)
Hispanic	-0.083 **	-0.114 ***	-0.108 ***
	(0.027)	(0.027)	(0.023)
Asian	-0.107 ***	-0.029	-0.099 ***
	(0.030)	(0.030)	(0.025)
Female	-0.009	-0.092 ***	-0.049 ***
	(0.012)	(0.013)	(0.011)
FRPL	-0.034	0.005	-0.020
	(0.022)	(0.022)	(0.019)
ELL	-0.144 ***	-0.022	-0.082 ***
	(0.015)	(0.015)	(0.013)
Constant	0.484 ***	0.511 ***	0.369 ***
	(0.026)	(0.026)	(0.022)
Adj. R^2	0.236	0.293	0.227
N	4325	4325	4325

Note: Models are linear probability models (fit via OLS) of passing status on indicated test in Spring of 10th grade; coefficients indicate the average difference in passing rates between the indicated group and the reference category (white, male, non-poor, non-ELL students), controlling for students' ELA and Math CST scores in 8th, 9th, and 10th grade. Models include district fixed effects. Sample includes all students in 10th grade in 2004 and 2005 with non-missing CAHSEE and CST scores and with 9th grade ELA CST scores in bottom quartile of state score distribution. Results (coefficients and standard errors) are virtually identical if we fit a logit model instead of a linear probability model. * p<.05; ** p<.01; *** p<.001

Table 7: Effects of CAHSEE Policy on Graduation, by Demographic Characteristics and Quartile of 10th-Grade ELA CST SCORE, with School-by-Year Fixed Effects

	Quartile of 10th Grade ELA CST			
	Quartile 1 (low)	Quartile 2	Quartile 3	Quartile 4 (high)
By Race/ Ethnicity Group				
White	-0.018	-0.041	-0.002	0.023 **
	(0.037)	(0.026)	(0.014)	(0.008)
Black	-0.181 ***	-0.030	-0.025	0.043 *
	(0.026)	(0.021)	(0.017)	(0.018)
Hispanic	-0.142 ***	-0.031 *	-0.014	0.015
	(0.016)	(0.012)	(0.011)	(0.012)
Asian	-0.165 ***	-0.036 *	0.018	0.007
	(0.024)	(0.016)	(0.012)	(0.009)
N	9553	14694	16159	15496
p(F-test)	0.002	0.980	0.138	0.278
By Gender				
Female	-0.189 ***	-0.055 ***	-0.001	0.012
	(0.017)	(0.011)	(0.008)	(0.006)
Male	-0.111 ***	-0.010	-0.003	0.025 ***
	(0.014)	(0.012)	(0.009)	(0.007)
N	9553	14694	16159	15496
p(F-test)	0.000	0.005	0.915	0.149
By English Language Learner Status				
ELL Students	-0.163 ***	-0.048 **	-0.018	-0.077
	(0.016)	(0.016)	(0.029)	(0.072)
Non-ELL Student:	-0.123 ***	-0.028 **	-0.001	0.020 ***
	(0.016)	(0.010)	(0.006)	(0.005)
N	9553	14694	16159	15496
p(F-test)	0.056	0.290	0.560	0.182
By Free/Reduced Lunch Eligibility Status				
Eligible	-0.196 ***	-0.034	-0.001	0.026
	(0.026)	(0.018)	(0.015)	(0.015)
Not Eligible	-0.148 ***	-0.025	0.011	0.018
	(0.033)	(0.021)	(0.015)	(0.013)
N	6135	9642	10047	9222
p(F-test)	0.167	0.708	0.466	0.553

Note: Each panel represents estimates from a different model. All estimates from models that correspond to models in Table 5, with the addition of school-by-policy cohort fixed effects. The p-value at the bottom of each panel is the p-value from the F-test of the null hypothesis that the coefficiennts are the same for each of the specified demographic groups. Standard errors corrected for the clustering of students within schools are shown in parentheses under each coefficient. * p<.05; ** p<.01; *** p<.001.

Table 8: Estimated Between-Group Differences in Test Scores on High- and Low-Stakes 10th Grade Tests, Graduating Classes of 2006 and 2007, by Race, Gender, Free/Reduced-Price Lunch Eligibility, and ELL Status

	ELA Test Scores (10th Grade)			Math Test Scores (10th Grade)		
	Low-Stakes Test (CST)	High-Stakes Test (CAHSEE)	Difference-in-Difference	Low-Stakes Test (CST)	High-Stakes Test (CAHSEE)	Difference-in-Difference
Black-White Difference	-0.091 **	-0.221 ***	-0.131 ***	-0.145 ***	-0.335 ***	-0.191 ***
	(0.034)	(0.038)	(0.042)	(0.036)	(0.039)	(0.046)
Hispanic-White Difference	-0.043	-0.172 ***	-0.129 ***	-0.103 **	-0.239 ***	-0.136 ***
	(0.031)	(0.035)	(0.038)	(0.033)	(0.035)	(0.043)
Asian-White Difference	0.025	-0.166 ***	-0.191 ***	0.007	-0.075	-0.082
	(0.035)	(0.038)	(0.042)	(0.037)	(0.039)	(0.047)
Female-Male Difference	0.052 ***	-0.003	-0.034	0.018	-0.083 ***	-0.080 ***
	(0.015)	(0.015)	(0.018)	(0.016)	(0.016)	(0.020)
FRPL Eligible-non-Eligible Difference	0.007	-0.022	-0.029	0.049	0.019	-0.030
	(0.026)	(0.028)	(0.031)	(0.027)	(0.029)	(0.035)
ELL-non-ELL Difference	-0.071 ***	-0.221 ***	-0.150 ***	0.018	-0.031	-0.049 *
	(0.018)	(0.020)	(0.022)	(0.019)	(0.020)	(0.024)

Note: All models control for students 8th and 9th grade ELA and Math CST scores, student fixed effects, district-by-test fixed effects, and student covariates (race/ethnicity, gender, free/reduced-price lunch eligibility, and ELL status) interacted with test. Sample includes all students in 10th grade in 2004 or 2005 with non-missing test scores and who scored in the bottom quartile of the state test score distribution on their 9th grade ELA CST test (N=4,325). * p<.05; ** p<.01; *** p<.001.

Table 9: Estimated Difference in Between-Group Differences in Test Scores on High- and Low-Stakes 10th Grade Tests, Graduating Classes of 2006 and 2007, by Quartile of 9th Grade ELA CST Score

	ELA Test Score Difference-in-Difference				Math Test Score Difference-in-Difference			
	Quartile 1 (Low)	Quartile 2	Quartile 3	Quartile 4 (High)	Quartile 1 (Low)	Quartile 2	Quartile 3	Quartile 4 (High)
Black-White Difference	-0.131 ***	-0.076 **	-0.070 ***	-0.050 *	-0.191 ***	-0.120 ***	-0.103 ***	-0.197 ***
	(0.042)	(0.025)	(0.020)	(0.023)	(0.046)	(0.029)	(0.026)	(0.031)
Hispanic-White Difference	-0.129 ***	-0.032	-0.011	-0.028	-0.136 ***	-0.090 ***	-0.075 ***	-0.084 ***
	(0.038)	(0.022)	(0.016)	(0.017)	(0.043)	(0.025)	(0.021)	(0.023)
Asian-White Difference	-0.191 ***	-0.119 ***	-0.067 ***	-0.014	-0.082	-0.040	-0.053 *	-0.053 *
	(0.042)	(0.024)	(0.017)	(0.016)	(0.047)	(0.028)	(0.022)	(0.021)
Female-Male Difference	-0.034	-0.026 *	0.038 ***	0.044 ***	-0.080 ***	-0.088 ***	-0.093 ***	-0.051 **
	(0.018)	(0.012)	(0.011)	(0.012)	(0.020)	(0.014)	(0.014)	(0.016)
FRPL Eligible-non-Eligible Difference	-0.029	-0.008	0.026	0.036 *	-0.030	-0.007	0.008	-0.071 **
	(0.031)	(0.019)	(0.016)	(0.017)	(0.035)	(0.023)	(0.021)	(0.022)
ELL-non-ELL Difference	-0.150 ***	-0.069 ***	-0.029	-0.134	-0.049 *	-0.054 **	-0.149 ***	-0.245
	(0.022)	(0.016)	(0.034)	(0.101)	(0.024)	(0.019)	(0.043)	(0.134)
Sample Size	4325	7614	8987	9354	4325	7614	8987	9354

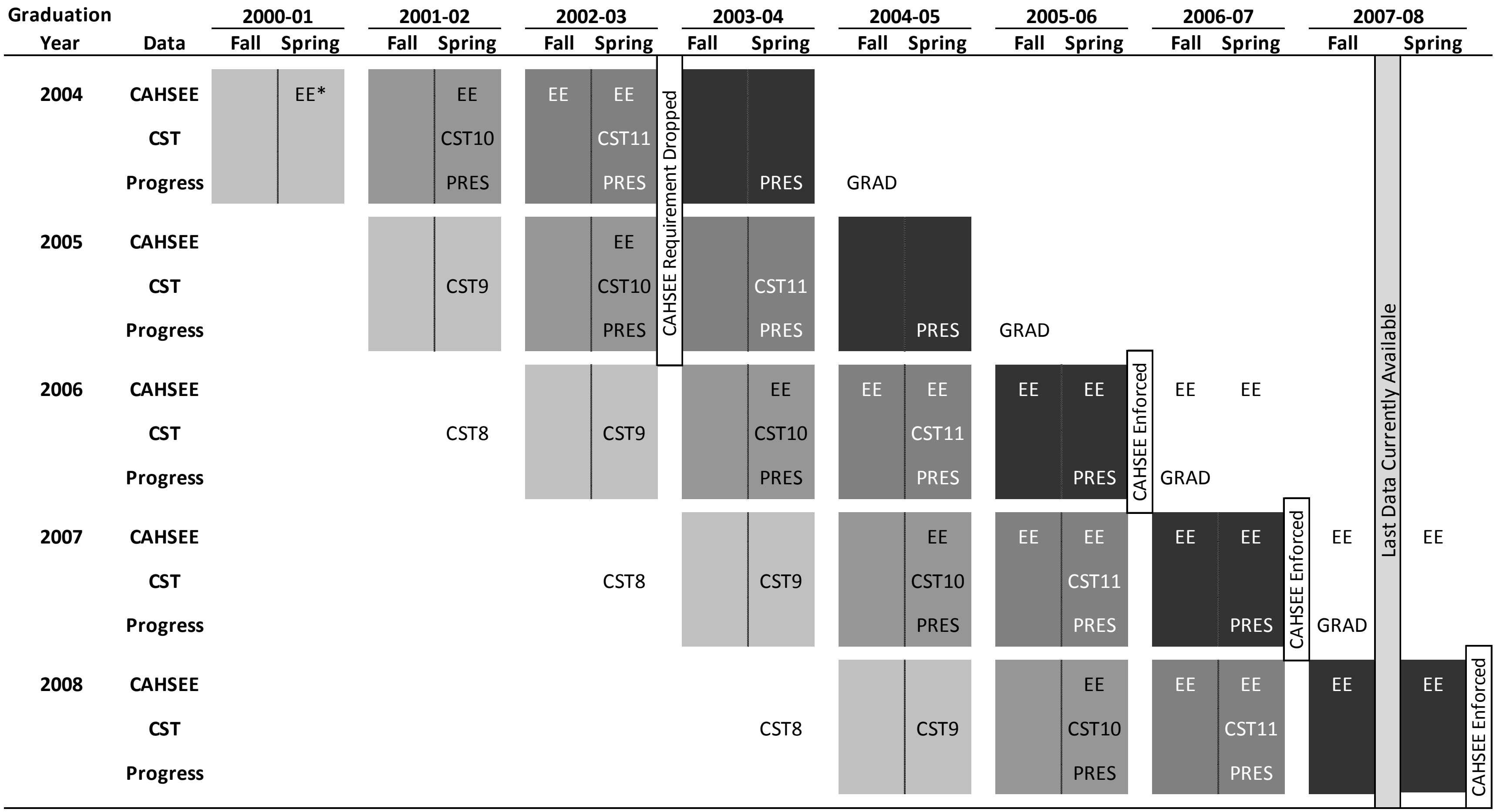
Note: All models control for students 8th and 9th grade ELA and Math CST scores, student fixed effects, district-by-test fixed effects, and student covariates (race/ethnicity, gender, free/reduced-price lunch eligibility, and ELL status) interacted with test. Sample includes all students in 10th grade in 2004 or 2005 with non-missing test scores and who scored in the relevant quartile of the state test score distribution on their 9th grade ELA CST test. * p<.05; ** p<.01; *** p<.001.

Table 10: Estimated Associations Between Student Characteristics and Stereotype-Threat-Adjusted CAHSEE Passing Rates, by CAHSEE Test, Students in Bottom Quartile of 9th Grade ELA Skill

	ELA	Math	Both
Black	-0.020	-0.056	-0.036
	(0.031)	(0.031)	(0.028)
Hispanic	0.002	-0.020	-0.007
	(0.029)	(0.029)	(0.025)
Asian	0.021	0.021	0.019
	(0.032)	(0.032)	(0.028)
Female	0.020	-0.047 ***	-0.013
	(0.013)	(0.013)	(0.012)
FRPL	-0.019	0.021	0.001
	(0.024)	(0.023)	(0.021)
ELL	-0.034 *	0.008	-0.027
	(0.016)	(0.016)	(0.014)
Constant	0.389 ***	0.446 ***	0.269 ***
	(0.028)	(0.028)	(0.025)
Adj. R^2	0.213	0.264	0.229
N	4325	4325	4325
p(F-test)	0.464	0.013	0.096

Note: Models are linear probability models (fit via OLS) of passing status on indicated test in Spring of 10th grade; coefficients indicate the average difference in stereotype-threat-adjusted passing rates between the indicated group and the reference category (white, male, non-poor, non-ELL students), controlling for students' ELA and Math CST scores in 8th, 9th, and 10th grade. Models include district fixed effects. Sample includes all students in 10th grade in 2004 and 2005 with non-missing CAHSEE and CST scores and with 9th grade ELA CST scores in bottom quartile of state score distribution. Results (coefficients and standard errors) are virtually identical if we fit a logit model instead of a linear probability model. F-test indicates test of null-hypothesis that all race groups have equal passing rates. * p<.05; ** p<.01; *** p<.001.

Figure1: Timeline for CAHSEE Administrations and Availability of CST, Persistence, and Graduation Data, High School Classes of 2004-2008



Legend

=Grade 9

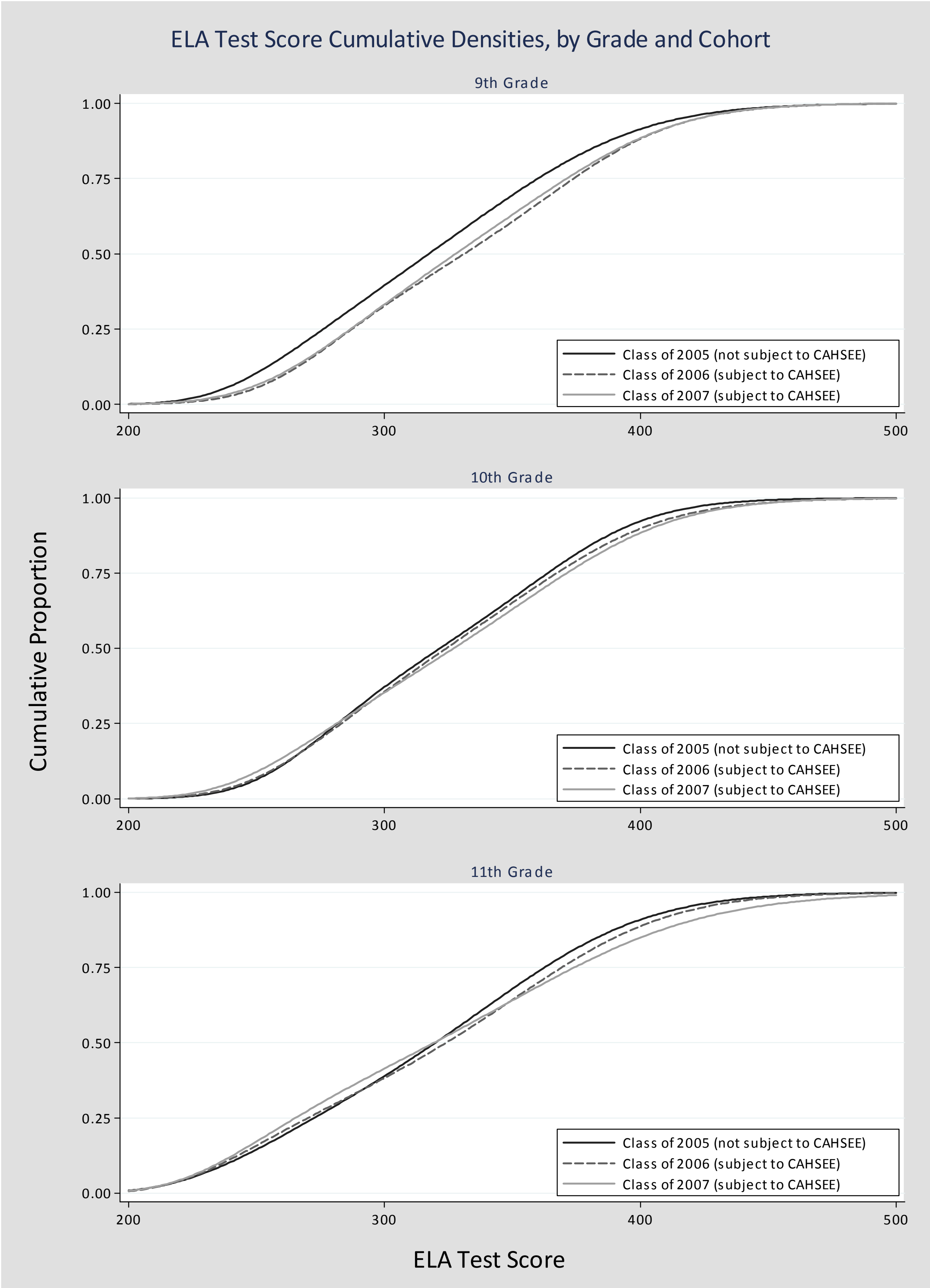
=Grade 10

=Grade 11

=Grade 12

Note: Figure illustrates the timeline for available information on each of the five graduating classes of 2004 through 2008. The shading indicates the grade level of student in a given class who has not been retained. "EE" indicates that students who had not yet passed the CAHSEE were administered an exit exam in the given semester. In any cell with "EE", students took the CAHSEE with the belief that they would be required to pass a CAHSEE exam in order to graduate. "CST#" indicates that students took the California Standards test in the #th grade year (CSTs are always administered in spring semesters). "PRES" indicates that we have an indicator variable for whether a student was present in a given semester. "GRAD" indicates the semester when on-time members of the graduating class should have first been coded as having graduated. *Asterisk indicates that, for this cohort, the CAHSEE was first administered to volunteer 9th graders.

Figure 2



ELA Test Score Cumulative Densities, by Grade and Cohort
9th Grade
10th Grade
11th Grade
Cumulative Proportion
1.00
0.75
0.50
0.25
0.00
200
300
400
500
Class of 2005 (not subject to CAHSEE)
Class of 2006 (subject to CAHSEE)
Class of 2007 (subject to CAHSEE)
ELA Test Score

Figure 3

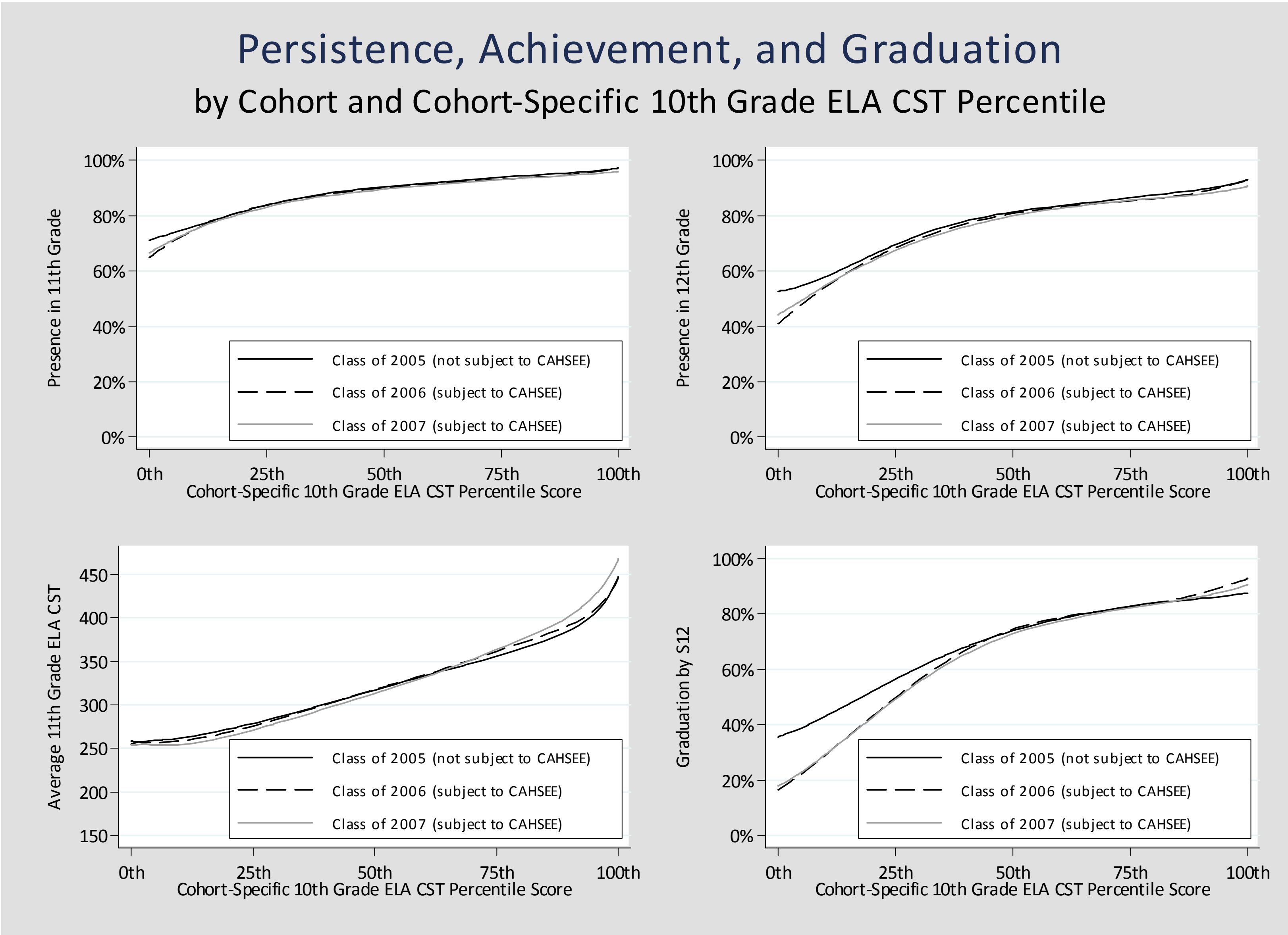
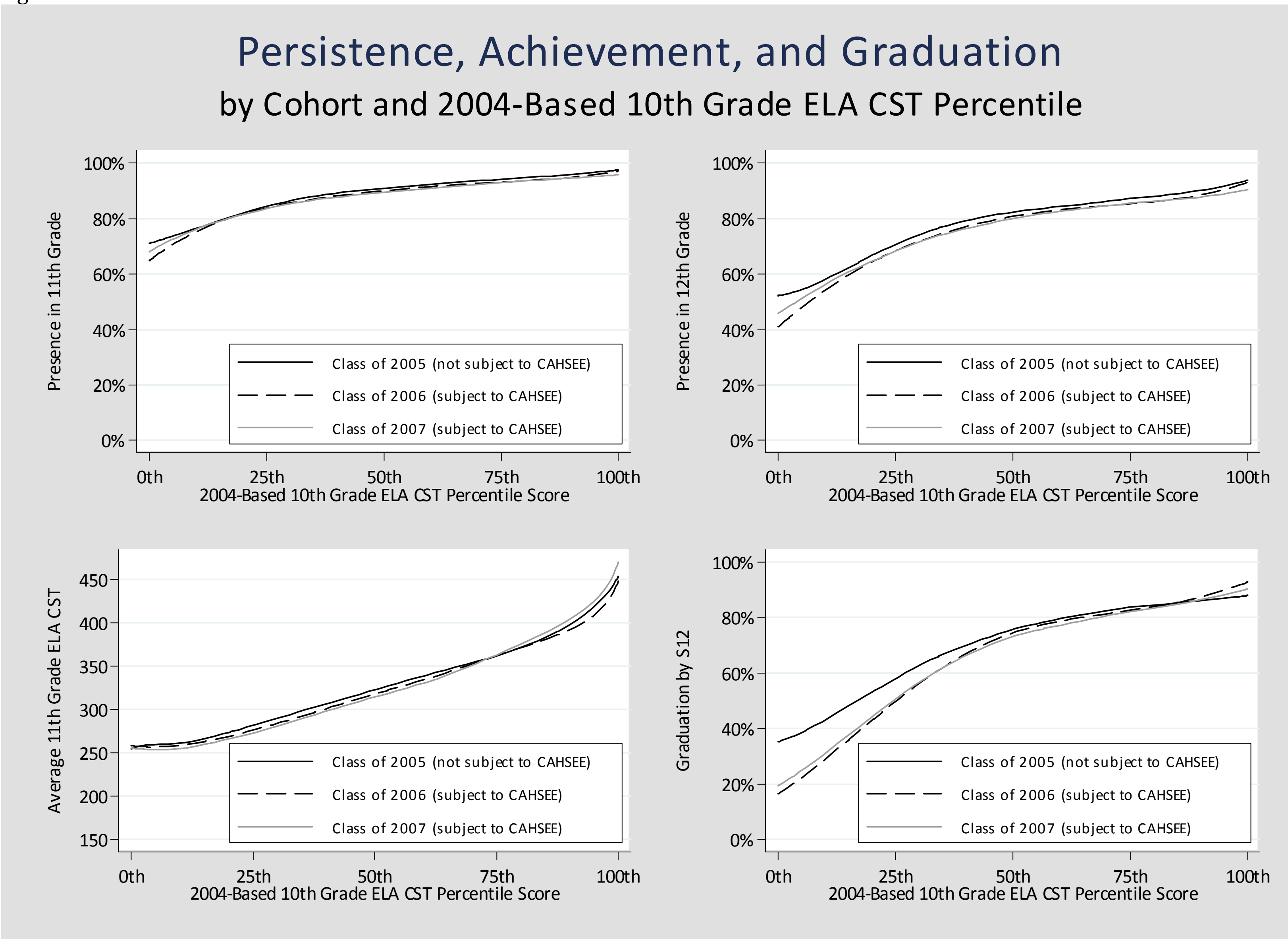

Persistence, Achievement, and Graduation
by Cohort and Cohort-Specific 10th Grade ELA CST Percentile
Presence in 11th Grade
Presence in 12th Grade
Average 11th Grade ELA CST
Graduation by S12
Cohort-Specific 10th Grade ELA CST Percentile Score
0th
25th
50th
75th
100th
0%
20%
40%
60%
80%
100%
150
200
250
300
350
400
450
Class of 2005 (not subject to CAHSEE)
Class of 2006 (subject to CAHSEE)
Class of 2007 (subject to CAHSEE)

Figure 4:



Persistence, Achievement, and Graduation
by Cohort and 2004-Based 10th Grade ELA CST Percentile
Presence in 11th Grade
Presence in 12th Grade
Average 11th Grade ELA CST
Graduation by S12
100%
80%
60%
40%
20%
0%
450
400
350
300
250
200
150
0th
25th
50th
75th
100th
2004-Based 10th Grade ELA CST Percentile Score
Class of 2005 (not subject to CAHSEE)
Class of 2006 (subject to CAHSEE)
Class of 2007 (subject to CAHSEE)

References

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35*, 29-46.

Becker, D. E., Wise, L. L., & Watters, C. (2008). *Independent Evaluation of the California High School Exit Exam (CAHSEE): 2008 Evaluation Report*. Sacramento, CA: Human Resources Research Organization.

Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology, 41*, 174-181.

Bishop, J. H., & Mane, F. (2001). The impacts of minimum competency exam graduation requirements on high school graduation, college attendance and early labor market success. *Labour Economics, 8*(2), 203-222.

Blascovich, J., Spencer, S. J., Quinn, D., & Steele, C. M. (2001). African Americans and high blood pressure: The role of stereotype threat. *Psychological Science, 12*(225-229).

Brown, R. P., & Josephs, R. A. (1999). A burden of proof: stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76*(2), 246-257.

Center on Education Policy. (2004). *State High School Exit Exams: States Try Harder But Gaps Persist*. Washington, DC: Center on Education Policy.

Center on Education Policy. (2005). *How have high school exit exams changed our schools? Some perspectives from Virginia and Maryland.* . Washington, DC: Center on Education Policy.

Cheryan, S., & Monin, B. (2005). "Where are you really from?": Asian Americans and Identity Denial. *Journal of Personality and Social Psychology, 89*, 717-730.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the Racial Achievement Gap: A Social-Psychological Intervention. *Science, 313*(5791), 1307-1310.

Danaher, K., & Crandall, C. S. (2008). Stereotype threat in applied settings re-examined. *Journal of Applied Social Psychology, 38*(6), 1639-1655.

Dee, T. S., & Jacob, B. A. (2006). *Do High School Exit Exams Influence Educational Attainment or Labor Market Performance?* (Working Paper).

Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin, 28*, 659-670.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Applied Developmental Psychology, 24*, 645–662.

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis, 23*(2), 99-121.

Keller, J. (2002). Blatant stereotype threat and women's math performance: Self-handicapping as a strategic means to cope with obtrusive negative performance expectations. *Sex Roles, 47*(3/4), 193-198.

Lavy, V., Paserman, M. D., & Schlosser, A. (2008). *Inside the Black of Box of Ability Peer Effects: Evidence from Variation in Low Achievers in the Classroom*: National Bureau of Economic Research.

Martorell, P. (2005). Do high school exit exams matter? Evaluating the effects of high school exit exam performance on student outcomes. University of California, Berkeley.

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science and engineering settings. *Psychological Science, 18*(10), 879-885.

O'Brien, L. T., & Crandall, C. S. (2003). Stereotype Threat and Arousal: Effects on Women's Math Performance. *Personality and Social Psychology Bulletin, 29*, 782-789.

Osborne, J. W. (2007). Linking stereotype threat and anxiety. *Educational Psychology, 27*(1), 135-154.

Ou, D. (2009). *To Leave or Not to Leave? A Regression Discontinuity Analysis of the Impact of Failing the High School Exit Exam*. London: Centre for Economic Performance, London School of Economics.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2008). *The consequences of High School Exit Examinations for Struggling Low-Income Urban Students: Evidence from Massachusetts*: National Bureau of Economic Research.

Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2008). *High Stakes, No Effects: Effects of Failing the California High School Exit Exam*. Stanford, CA: Working Paper Series, Institute for Research on Educational Policy and Practice, Stanford University.

Reardon, S. F., & Galindo, C. (2002, April). *Do High-Stakes Tests Affect Students' Decisions to Drop Out of School? Evidence from NELS.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Shih, M., Pittinsky, T., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10*(1), 80-83.

Smith, J. L. (2004). Understanding the process of stereotype threat: A review of mediational variables and new performance goal directions. *Educational Psychology Review, 16*(3), 177-206.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4-28.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613-629.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 67*(5), 797-811.

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34*(4), 665-693.

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit and achievement. *Journal of Personality and Social Psychology, 92*(1), 82-96.

Walton, G. M., & Spencer, S. J. (in press). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*.

Warren, J. R., & Edwards, M. R. (2005). High School Exit Examinations and High School Completion: Evidence From the Early 1990s. *Educational Evaluation and Policy Analysis, 27*(1), 53-74.

Warren, J. R., Jenkins, K. N., & Kulick, R. B. (2006). High School Exit Examinations and State-Level Completion and GED Rates, 1972-2002. *Educational Evaluation and Policy Analysis, 28*(2).

Wise, L. L., Becker, D. E., Harris, C. D., Sun, S., Wang, X., & Brown, D. G. (2004). *California High School Exit Examination (CAHSEE): Year 5 Evaluation Report*. Sacramento, CA: Human Resources Research Organization.

Wise, L. L., Harris, C. D., Brown, D. G., Becker, D. E., Sun, S., & Coumbe, K. L. (2003). *California High School Exit Examination (CAHSEE): Year 4 Evaluation Report*. Sacramento, CA: Human Resources Research Organization.

Wise, L. L., Harris, C. D., Sipes, D. E., Hoffman, G. R., & Ford, P. J. (2000). *High School Exit Examination (HSEE): Year 1 Evaluation Report.* Sacramento, CA: Human Resources Research Organization.