Disparate Impacts of Teacher Certification Exams

Christa Deneault*

Evan Riehl†

Jian Zou‡

November 24, 2025

Abstract

We use Texas administrative data to assess the long-standing claim that teacher certification exams have a disparate impact on prospective teachers from underrepresented minority (URM) groups. In a regression discontinuity design, we find that failing a certification exam delays entry into teaching, ultimately costing the average candidate who marginally fails $10,000 in forgone earnings. These costs fall disproportionately on URM teaching candidates because they are both more likely to fail and have more negative causal effects of failing. To assess the effectiveness of the exams as a screen for teaching quality, we exploit a reform that significantly increased the difficulty of Texas' elementary certification exam. We find that the distribution of teacher value-added among those who passed the easier pre-reform exam is similar to that for those who passed the harder post-reform exam, showing that the harder exam was not any better at screening out low-quality teachers. Taken together, our findings show that certification exams have a disparate impact on URMs in the sense that they impose disproportionate economic costs that are unjustifiable given limited differences in teaching quality.

*Deneault: Federal Reserve Bank of Dallas (email: christa.deneault@dal.frb.org)

†Riehl: Department of Economics and ILR School, Cornell University (email: eriehl@cornell.edu)

‡Zou: ILR School, Cornell University (email: jz2326@cornell.edu)

The views expressed in this article are those of the authors and do not necessarily reflect the opinions or official position of the Federal Reserve Bank of Dallas, Federal Reserve System, Texas Education Agency, the Texas Higher Education Coordinating Board, Texas Workforce Commission, or the state of Texas. This research was supported by a grant from the Spencer Foundation. We are grateful to Holly Kosiewicz, Zhixiang Lu, and members of the Texas Education Research Center Advisory Board for help in accessing and working with the data. All errors are our own.

1 Introduction

Economists have extensively examined labor market discrimination from the perspective of individual agents, such as firms engaging in taste-based or statistical discrimination.[1] By contrast, relatively little attention has been paid to how systemic or institutional features of labor markets may interact to generate unequal outcomes across groups of prospective workers (Bohren et al., 2025b; Small and Pager, 2020). A salient example is certification exams for public school teachers: the requirement that prospective teachers pass content and pedagogy exams creates racial disparities in access to the profession given differences in test-taking abilities between white and underrepresented minority (URM) candidates.[2]

Importantly, racial disparities in certification exam pass rates may be considered illegal discrimination under Title VII of the Civil Rights Act even if these exams are not motivated by discriminatory intent. The central legal question is whether such policies have *disparate impact*, meaning that they impose a disproportionate cost on members of a particular racial group that is not justified by differences in job performance between racial groups. Since the 1980s, numerous lawsuits have been filed alleging that certification exams have a disparate impact on Black and Hispanic prospective teachers. Yet identifying disparate impacts is empirically difficult for two main reasons. First, it is hard to quantify the costs of licensing exams on prospective teachers because few datasets contain careers outcomes for individuals who both passed and failed the exams. Second, teaching quality is not observed for individuals who do not pass the exams, making it hard to conclusively determine whether licensing exams are an effective screen for job performance.[3] As a result, many legal challenges have faltered due to lack of evidence, and most U.S. states continue to require certification exams despite large racial disparities in pass rates.

We address these challenges to provide new evidence that teacher certification exams have a disparate impact on Black and Hispanic prospective teachers. We use rich administrative data from Texas and multiple quasi-experimental designs to present causal evidence on the two components in the legal definition of disparate impact: racial disparities in the effects of certification exams on prospective teachers' careers, and limited effectiveness of these exams as a screen for teacher quality.

We begin by providing the first causal evidence on the economic costs of failing certifica-

[1] For recent examples see Bohren et al. (2025a), Kline et al. (2022), Benson et al. (2024), Lepage (2024), among others.

[2] Many papers document racial differences in test scores at all levels of education. For example, Fryer Jr and Levitt (2006) demonstrates that black-white test score gaps begin to diverge as soon as third grade.

[3] Existing studies typically correlate licensing exam performance with teacher value-added (e.g., Goldhaber, 2007; Clotfelter et al., 2010), but this work cannot overcome the censoring issue that arises because value-added is not observed for those who fail the exams.

tion exams for prospective teachers. Using a regression discontinuity (RD) design, we find that narrowly failing a license exam delays entry into the teaching profession for both whites and URMs, with larger effects for URMs. In the year following their first certification exam attempt, whites are 9 percentage points (pp) less likely to be working in a public school while URMs are 15pp less likely if they narrowly fail. Although these employment effects decline over time as candidates retake exams, URMs remain 3pp less likely to be employed in Texas public schools ten years later. Because of these delaying and discouraging effects, we find that narrowly failing an exam imposes a cost to long-term cumulative earnings: whites earn $7,800 less and URMs earn $12,200 less over a ten year period.[4] Importantly, URM candidates are also more likely to experience these earnings losses since they have much lower first-attempt pass rates than white candidates (57 percent vs. 83 percent).

Although failing is costly, especially for URMs, such costs may be justifiable if individuals who fail the exams would be significantly worse teachers than those who pass. An inherent issue in measuring disparate impacts is that outcomes such as productivity are censored, meaning we do not observe productivity for those who are kept out of the profession. Arnold et al. (2022) overcome this censoring issue in the context of bail judge release decisions by exploiting that fact that some judges are extremely lenient. This fact, combined with the quasi-random assignment of defendants to judges, allows the researchers to observe average uncensored outcomes and thus estimate disparate impacts of judge release decisions. This type of identification is not possible in our setting because there is no such lenient "decision maker," i.e., the exam pass rates are far from 100 percent, particular for URM candidates. This measure has potential applications in other settings that have policy variation but no lenient decision makers.

Instead, we utilize quasi-experimental variation from a certification exam regime change to create a new measure of disparate impact that we call *policy-relevant disparate impact*. In 2010, Texas increased the difficulty of its certification exams for elementary teachers by replacing its EC-4 certification exam (early childhood through grade 4) with an EC-6 exam (early childhood through grade 6). This reform made it significantly harder to become an EC-4 teacher because these prospective teachers also had to learn grade 5–6 content to pass the exams. We define policy-relevant disparate impact to be the pass rate difference between URM and white prospective teachers of equal levels of productivity for those who took the harder EC-6 certification exam *and who would have passed the easier EC-4 exam*. We consider this group "policy-relevant" because it represents the disparate impact of the

[4]These earnings are the present value of cumulative earnings over a 10 year period. This is estimated as the sum of discounted nominal earnings through for the 10 years following taking the exam using a 5% discount rate.

more difficult exam within the population of individuals who would have been eligible to teach if the easier exam had remained in place.

Using a potential outcomes framework, we define the assumptions that are necessary to estimate policy-relevant disparate impact and show they are plausible in our setting. Our analysis uses teacher test score value-added as a measure of teaching productivity (Chetty et al., 2014a).[5] We assume exam pass rates and the distribution of teacher value-added is unrelated to the timing of the certification exam reform. In other words, the underlying distribution of prospective teachers is independent of this policy shock to elementary certification exams.[6] We also rely on an assumption that anyone who passed a harder certification exam on their first attempt could have also passed an easier certification exam on their first try. We provide evidence supporting the plausibility of our assumptions and robustness to our primary specification, including a differences-in-differences-inspired estimator that nets out potential time trends in the composition of prospective teachers.

Our main estimate of policy-relevant disparate impact indicates that, among those who would have passed their certification exams during the easier exam regime, URM prospective teachers were 14.6pp (11.7pp) less likely to pass the harder exam than white candidates with the same math (ELA) value-added. Combining this result with our RD earnings estimates suggests that the move to a harder certification exam costs URM prospective candidates approximately $8.5 million in present value earnings in any given year.[7]

Notably, our estimate of policy-relevant disparate impact measure closely matches the raw pass rate differences between URMs and whites. We show that the experience-adjusted distributions of value-added for those who passed under the easier exam and harder exam are nearly identical for both URMs and whites, suggesting that the increased difficulty of the certification exam did little to screen out poor performing teachers. Further, correlations between value-added and certification exams are negligible, further indicating that certification exams have little predictive power of teacher effectiveness. While these methods point to the conclusion that value-added is uncorrelated with certification exams, none fully circumvent

[5] Value-added has been shown to be a well validated measure of teacher influence on student achievement, but only reflects one dimension of student outcomes. We also consider non-cognitive value-added and find our results to be similar.

[6] Among those who gain access to the teaching profession we still may not observe their value-added, which can only be computed for math and ELA subjects in certain grade levels and if a teacher remains in the profession. One additional assumption we make is that the likelihood of observing value-added conditional on having passed is similar in the pre-/post- regime periods. For example, teachers are not more or less likely to be teaching fourth grade math and ELA subjects in the harder exam regime relative to easier exam regime.

[7] There are approximately 4,700 URMs who attempt a certification exam yearly. Combined with our estimate of policy-relevant disparate impact this implies about 700 URMs affected by the hard exam relative to the easy exam regime. We use our RD estimates of a loss in PV discounted earnings over 10 years to be $12,200 for a total of approximately $8.5 million.

the issue of censoring nor do they account for supply-side adjustments when certification exams become more difficult. For example, our results suggest that a harder certification exam may discourage some prospective teachers from entering the teaching profession. How schools deal with this change in supply when demand for teachers remains fixed is an open question.

To ascertain reduced-form effects on net, we conclude by examining the impact of the certification exam regime change using an RD-DiD design. Our DiD additionally leverages the exam regime change by comparing elementary to middle school outcomes in a pre/post framework, while the RD captures shocks to schools' labor supply through teacher departures. Together these isolate how demand for new teachers under easier or harder certification exam regimes affects teacher composition and student achievement.

Consistent with our earlier analysis, our RD-DiD estimates show when schools experience a vacancy under stricter licensing regime, they employ 2pp fewer URM teachers, offsetting them with white teachers. This compositional shift also results in a redistribution of income: white teachers collectively earn, on average, $6,000 more under the stricter exam regime, while URM teachers earn collectively less. Yet these changes in teacher composition do not translate to higher test scores for students. We find no evidence of improvement in student standardized test scores when the teacher certification exams became more difficult, and in some specifications we find that students are made slightly worse off.

We make three contributions to the existing literature. First, we estimate both the costs to failed access to a licensed occupation and the heterogeneous costs across race and ethnicity. Some work with respect to teachers has explicitly studied how certification stringency heterogeneously affect subpopulations' supply but have found mixed outcomes. Law and Marks (2009) find an increase in Black participation following certification, Larsen et al. (2020) and Law et al. (2023) find no impact of changes to licensing stringency on diversity of the teacher workforce, and Angrist and Guryan (2008) and Chung and Zou (2025) find it leads to fewer diverse teachers. Other work has examined the wage gains for licensed workers by race or sex and finds that the earnings gap shrinks for licensed workers (e.g., Blair and Chung, 2025; Yan, 2024; Xia, 2021). However, these papers lack information about the costs of licensing barriers, and comparing wages of licensed to unlicensed workers (or even an increase post-license) has a selection issue: those who want to become licensed may be different than those who do not. Our RD design circumvents selection issues by examining teachers who narrowly fail the exam to show certification exams are a costly barrier in terms of employment and earnings, particularly for URMs. We also add to the literature on effects of certification stringency on supply by showing the composition of teachers changes in favor of whites when license exams become more difficult.

Second, we contribute to a burgeoning literature on quantifying disparate impacts by defining a policy-relevant disparate impacts measure. Recent work has measured disparate treatment in bail decisions, foster care, and policing, among others (e.g., Arnold et al., 2022; Baron et al., 2024; Goncalves and Mello, 2021). We contribute to this literature by modifying and extending the work from Arnold et al. (2022), which we believe can be used as a rubric for future work where population-level quasi-experimental variation is not available. Instead, our measure relies on changes in subpopulations affected by a policy change, a situation common in other settings.

Finally, we provide new and more convincing evidence that certification exams are unrelated to value added and that changes to certification exam stringency have little impact on student achievement on net (including supply side effects). Previous work has correlated value added with certification exams in an attempt to address whether it is predictive of teacher productivity and has found that it is weakly correlated (Goldhaber, 2007; Aaronson et al., 2007; Kane et al., 2008; Clotfelter et al., 2010; Goldhaber and Hansen, 2010; Shuls and Trivitt, 2015; Hendricks, 2016; Goldhaber et al., 2017; Cowan et al., 2020).[8] But this approach falls short of observing all teachers who initially fail and are discouraged from the profession, does not take into consideration potential supply-side impacts of changes to exam stringency, and may simply represent a lack of correlation between scores and value-added among those who passed. In a different approach using RD, Orellana and Winters (2023) find failing a certification pushes out those who would have been *better* teachers.[9] Their method requires strong assumptions about value-added for those who never enter the teaching profession, and they do not have information on race. We add to this literature by examining a certification exam reform, which allows us to rule out meaningful differences in the value added across teachers who pass differentially difficult exams. Our RD-DiD measure also incorporates potential supply-side adjustments and points to similar evidence: harder certification exams did not improve student achievement yet resulted in a less diverse workforce.

[8]Only a subset of these papers consider that certification exams may be differentially predictive across race/ethnicity. In all cases the finding is a weak correlation between value added and certification. But there is mixed evidence about the predictiveness across race/ethnicity. For example, Goldhaber and Hansen (2010) finds license exams are more predictive of student test achievement for white teachers than Black teachers, while Cowan et al. (2020) finds the opposite. We find that certification is weakly predictive for whites but not predictive for URMs.

[9]Other work has examined quasi-experimental changes to licensing on teacher quality but cannot connect teachers to students. See Angrist and Guryan (2008), Larsen et al. (2020), and Chung and Zou (2025).

2 Background and Data

2.1 Requirements for teacher certification in Texas. Certification is a requirement for teachers in public schools in Texas.[10] To become a certified teacher in Texas, a candidate must hold a bachelor's degree, complete teacher training, and pass two certification exams.[11] These two certification exams include a content exam, which is aligned with the candidate's intended teaching subject and grade level, and a Pedagogy and Professional Responsibilities (PPR) exam, a general test on classroom management. Since the 2002-03 academic year, Texas administers these two types of exams under the Texas Examinations of Educator Standards (TExES).[12] The TExES exams employ a scaled scoring system, with a passing threshold set at 240 (with a maximum score of 300). If candidates do not pass their first attempt at a license exam, they may retake. Exams are offered year round and cost a little over $100 to complete as of 2025.[13]

Table 1 presents averages for demographics, TExES exam performance, and earnings among first-time certification exam takers and comparison populations, including first-year teachers, bachelor's degree (BA) recipients, and K-12 students between 2002 and 2019. Women are over-represented in teaching relative to BA earners and K-12 students. In terms of racial and ethnic composition, there are no notable differences across exam takers, new teachers, and BA graduates; however, these groups are disproportionately more white relative to the general population. Income for first-year teachers is higher than the average for new college graduates, but within five years the average college graduate outpaces earnings for teachers.

2.2 Differences in test scores between URMs and whites. Historically, white students have outperformed Blacks and Hispanics on standardized tests, and these differences have been documented as early as at the start of formal schooling (Jencks and Phillips, 1998; Fryer Jr and Levitt, 2006). Among first-time certification test takers, we find URMs score an average of 0.4 standard deviations lower on their high school math scores than whites - see Table 1. This trend follows prospective teachers to their certification exams. On average, the pass rate gap between whites and URMs is 26pp for content exams and 14pp for PPR exams - see Table 1. This difference at the mean masks a general shift in the distribution of test scores. Figure A1 shows the proportion of each race/ethnicity in each

[10] Uncertified teachers are allowed to teach on a per permit basis in emergency situations - see TEC§21.055 (Templeton et al., 2022).

[11] See sections TEC§21.003 and TEC§21.044. Since 2008, Texas also requires fingerprinting and background checks for the full certification (Agency, 2022).

[12] Prior to 2003, Texas used the Examination for the Certification of Educators in Texas (ExCET).

[13] Fees per exam are provided here: https://www.tx.nesinc.com/PageView.aspx?f=GEN_Tests.html.

score among prospective teachers' first attempt. At lower test scores URMs are significantly over-represented relative to the proportion of test-takers whereas whites make up over 80 percent of the test takers in the highest scores (whites make up 59 percent of all test-takers). Thus, the distribution of test scores is heavily skewed to the left for URMs and to the right for whites.

While several reasons for these differences in performance on standardized tests have been posited, including racial segregation, school quality, and test-score scaling, there are few definitive answers as to why it exists and how to close it (Jencks and Phillips, 1998; Fryer Jr and Levitt, 2006; Card and Rothstein, 2007; Penney, 2017; Bond and Lang, 2018). In this paper, we take these test-taking differences at entry into the labor market as given. We remain agnostic as to *why* these gaps occur and acknowledge that they may represent previous discrimination or unequal impacts experienced prior to taking a teacher certification exam.

2.3 Texas administrative data. We use administrative data from four Texas agencies:

- **Texas Education Agency (TEA)**. TEA provides data on all public school teachers in Texas from 1996–2019. These data include each teacher's demographics, salary, years of teaching experience, school of employment, and full-time equivalent (FTE) years associated with their teaching grade(s) and subject(s). We also use TEA data on grade 3–8 state standardized tests from 1994–2019, which include student test scores and demographics.[14] We can connect student test scores to groups of teachers at the school/grade/subject level for all years from 1996–2019. From 2012–2019, we also observe classroom identifiers that allow us to connect students to individual teachers and compute teacher value-added estimates.

- **Texas State Board for Educator Certification (SBEC)**. SBEC provides data on all teacher certification exams in Texas between 2002 and 2019.[15] This data includes the exam date, test type, test description, test score, test result (pass/fail), certification route, the in-state educator preparation program that approved the educator to test, and the state in which the educator was prepared. For our analysis, we restrict the sample to content tests of the TExES, the primary certification exam type since 2002, and exclude PPR exams. We use the exam date to identify the first exam taken by each candidate, ensuring that we focus on initial exam performance. The test description

[14]Texas had three testing regimes during this time period: Texas Assessment of Academic Skills (TAAS, 1994–2002), Texas Assessment of Academic Skills (TAKS, 2003–2011), and State of Texas Assessments of Academic Readiness (STAAR, 2012–2019).

[15]This dataset also contains certification exams back to 1986.

enables us to distinguish between specific exams, allowing us to compare certification performance and examinees' outcomes within the same exams.

- **Texas Higher Education Coordinating Board (THECB)**. We use THECB data on all bachelor's degree graduates from public (1992–2019) and private colleges (2003–2019) in Texas. We observe each graduate's college, major, and graduation year.

- **Texas Workforce Commission (TWC).** TWC provides administrative earnings records for all individuals working at registered firms in Texas from 1992–2021. We use this data to measure individuals' annual and cumulative earnings in the years before and after they transition into teaching careers. All earnings are expressed in constant 2019 dollars, deflated using the U.S. Consumer Price Index from the Bureau of Labor Statistics. We calculate the present value (PV) of cumulative earnings as the sum of discounted nominal earnings over 1, 5, or 10 years after taking the exam, using a 5% discount rate. This measure better captures the economic cost of failing the exam by expressing the earnings loss in dollar terms at the time of the exam.

We accessed the data via the University of Texas at Dallas Education Research Center (ERC), which contains unique personal identifiers that allow us to connect all datasets at the individual level. With these links, we can follow teachers from their teacher certification exams into teaching jobs and observe their earnings, regardless of whether they become teachers. Appendix B.1 provides details on the variables that are derived from the administrative data.

3 The costs of failing a certification exam on prospective teachers' careers

3.1 Regression model. We begin our analysis by using a local linear regression discontinuity (RD) model to estimate the causal impacts of failing a teacher certification exam on prospective teachers' careers:

$$Y_{ie} = \beta D_{ie} + \alpha x_{ie} + \psi D_{ie} x_{ie} + \gamma_{t(i)e} + \varepsilon_{ie} \quad \text{if } |x_{ie}| \leq h^Y. \tag{1}$$

Y_{ie} is an outcome for individual i who took certification exam e. The variable of interest, D_{ie}, is an indicator for failing the certification exam on the first attempt. We include an interaction between D_{ie} and the running variable, x_{ie}, which is individual i's certification exam score on the first attempt normalized to equal zero for the lowest passing score (240).

To restrict identification to individuals taking similar exams, we include fixed effects, $\gamma_{t(i)e}$ for exams, e, interacted with the year in which individual i took the exam, $t(i)$. Our regression samples include only applicants whose admission scores are within h^Y standard deviations of the admission threshold. Our results use the Calonico et al. (2019) bandwidth computed separately for each outcome Y. We weight observations in equation (1) using a triangular kernel following the default options in the `rdrobust` package of Calonico et al. (2019). We cluster standard errors at the individual level, as roughly 0.8 percent of individuals in our sample appear in the data multiple times because they took more than one certification exam on the same day.

We estimate equation (1) for all certification exam takers and separately for two groups: white and URM candidates. The separate estimates for these groups show how failing a teacher certification exam might differently affect teacher candidates by race. We also jointly estimate the model with interaction terms and report p-values to indicate whether differences in effects across the two racial groups are statistically significant.

3.2 Identification assumptions and balance tests. The main RD identification assumption is that candidates' exam scores are effectively randomly assigned near the thresholds. This assumption is supported by balance tests that show that exam-taker characteristics do not change discontinuously at the passing threshold. Appendix Table A2 presents estimates from RD regressions that use demographic characteristics, K-12 enrolled schools and programs, high school achievement, and prior year earnings as dependent variables. Most RD estimates are statistically insignificant, and we cannot reject the hypothesis that these coefficients are jointly equal to zero for the full sample ($p = 0.242$), white ($p = 0.737$), and URM candidates ($p = 0.524$).

There is a discontinuous change in the density of certification exam scores at the passing threshold, but this likely reflects how TExES scores are scaled rather than score manipulation. According to a TExES technical manual, the formula that converts raw scores (proportion of correct answer) to scale scores differs depending on whether the raw score is above or below the passing standard.[16] This means that the relationship between underlying ability and scale scores is likely to differ above and below the threshold for any particular exam, and thus scale scores may be more "spaced out" on one side of the threshold. Indeed, we find that the relationship between exam takers' high school math scores and their certifi-

[16]TExES scores are scaled from 100–300 with 240 representing the minimum passing score. The scaling formula is as follows. "For raw scores greater than or equal to the minimum passing (cut) score obtained by a standard-setting study: Scaled score = 240 + [60*(raw score − raw cut score) / (max raw score − raw cut score)]. For raw scores less than the minimum passing score: Scaled score = 100 + 140*(raw score)/(raw cut score)." See "Texas Educator Certification Program Technical Manual, Valid through August 31, 2018," downloaded in October 2024 at: https://tea.texas.gov/texas-educators/preparation-and-continuing-education/program-provider-resources/texastechnicalmanual8.31.18.pdf

cation exam scores tends to be flatter below the passing threshold than above it, consistent with the lower density of scores below the threshold. Appendix Figure A2 shows that if we rescale failing scores to have the same linear relationship with high school math scores as passing scores in the vicinity of the threshold, we no longer find a discontinuous change in the density of scores using the test recommended by Cattaneo et al. (2020). Appendix Table A5 shows that our main RD results are essentially unchanged when we use these rescaled scores as our running variable (i.e., the rescaling affects the slope coefficient ψ but not the RD coefficient β). Further, we are not aware of any mechanism through which exam takers could manipulate their scores near the passing threshold.

3.3 The effects of failing a certification exam on short– and long–run teacher outcomes. Table 2 presents RD estimates of the impacts of failing a certification exam on candidates' teaching careers. In Panel A, candidates who marginally fail, on average, are 11.8pp (column C) less likely to be employed as a teacher one year later. URM candidates experience larger negative effects (15pp, a 28.6% decrease relative to the mean of the dependent variable) compared to white candidates (8.7pp, a 17.5% decrease), and the differential effects are statistically significant ($p = 0.000$). In addition, marginally-failed candidates have less years of teaching experience, where URM candidates experience larger effects (25.5% reduction relative to the dependent variable mean) than white candidates (16.6% reduction). The RD differences on teaching experience between URM and white candidates are statistically significant ($p = 0.000$).[17] Figure A3 graphically illustrates these RD results, displaying plots of teaching outcomes across certification exam scores separately for white (black) and URM (red) candidate groups. These plots visually show outcome disparities by race and ethnicity across the passing threshold.

For earning effects, on average, marginally failed candidates receive $2,831 less one year after the exam. The negative effects are more prominent for URM candidates ($3,942, a 9.5% reduction relative to the dependent variable mean) than white candidates ($1,919, a 5% reduction). The unequal effects between the URM and white candidates are statistically different from zero ($p = 0.000$). Similarly, marginally-failed candidates experience $4,062 less cumulative earnings (in present value) one year after the exam, and the effects are significantly larger ($p = 0.000$) for URM candidates ($5,672, a 10.4% reduction relative to the dependent variable mean) than white candidates ($2,532, a 5.2% reduction).[18] Figure 2

[17] As a supplement to results on teaching outcomes, Table A3 shows RD effects of failing a certification exam on exam outcomes, which includes indicators on whether pass the certification by now, whether pass any certification exam by now, the number of days to pass this exam, and the number of days to pass any exam, by white and URM groups. Results in Table A3 confirm that marginally-failed candidates are less likely to pass both current and any certification exam and take more days to pass both current and any exam. Again, URM candidates experience statistically larger impacts on exam outcomes than white candidates.

[18] Table A4 details the effects of failing a certification exam on earnings outcomes, including zero-earnings

graphs the associated RD plots of earning outcomes across exam scores separately for white (black) and URM (red) groups, illustrating outcome disparities by race beyond the passing threshold.

Given the ability for prospective teachers to retake, we examine the longer-run effects of failing their first attempt in Panels B and C. Five years after their first certification exam (Panel B), significant negative career impacts persist for marginally failing candidates overall and within URM and white subgroups. Over this extended period, RD effects diminish slightly for the likelihood of employment as a teacher, annual earnings, and cumulative earnings, while the effects on teaching experience grow larger. In addition, the differences between URM and white candidates in these RD effects remain statistically significant for teaching and earnings outcomes. Ten years post-exam (Panel C), most RD effects continue to be significant, but differences between URM and white candidates become statistically insignificant. Figure 3 provides a graphical illustration of the RD coefficients yearly for the first 10 years post-first attempt by white and URM groups, using a balanced panel sample.

These findings indicate that failing a certification exam substantially delays or deters entry into teaching, resulting in losses in both teaching experience and cumulative earnings. On average, marginally failed candidates lose around $9,700 in cumulative earnings over the ten years following the exam, with significantly larger negative impacts for URM candidates.

3.4 Heterogeneity. Table A6 examines heterogeneity within URM groups by further separating the URM group into Black and Hispanic subsamples. Results indicate that marginally-failed Black teacher candidates are less likely to become teachers, have less teaching experience, and have less annual and cumulative earnings, compared to their Hispanic counterparts. The RD effect differences on teaching outcomes are statistically significant in one and five years after the certification exam.

In addition to heterogeneity across race groups, Table A7 investigates RD effects of failing a certification exam by socioeconomic status (SES), where teacher candidates who participated in the free or reduced-price lunch when they were students are defined as low SES. Results show that, compared to high SES teacher candidates, low SES candidates who marginally failed the exam experience larger negative effects — are less likely to become a teacher, have less teaching experience, and have less annual and cumulative earnings. The unequal effects on teacher employment indicator, teaching experience, and cumulative earnings one year after the exam are statistically different from zero.

However, the difference in RD effects across whites and URMs is not explained solely

samples and teaching-specific earnings. The results show a consistent pattern as found in Table 2, reinforcing the robustness of RD effects on earning outcomes. Additionally, for earnings from teaching, statistically significant racial differences in the impact on teaching earnings are evident in the short term (one year after the exam) but become statistically insignificant over the longer run (five or ten years after the exam).

by SES. Table A8 further separates the group by race and SES, revealing racially heterogeneous effects within both the high- and low- SES groups. These findings confirm that the RD differential effects between white and URM groups are not driven by differences in socioeconomic status.

3.5 Extrapolated effects. While RD estimates indicate the impacts of failing a certification exam on outcomes near the passing score, the results are less informative on the impacts for candidates away from the cutoff. We thus conduct an extended exercise to examine the extrapolated effects. To do so, we use the observed relationship between the scores and outcomes using samples with an exam score range of 240 and 260, and employ a linear extrapolation to project the outcomes for candidates who are below the passing score. Figure A6 visually illustrates the scatter plots and the extrapolated linear lines for the baseline outcomes between white and URM candidates.

Figure 4 summarizes the RD effects and extrapolated effects of the three samples — exam takers with a score within ten points of the passing score, all exam takers with a score below the 240 cutoff, and all exam takers — for teacher employment and cumulative earnings. The results show that, while failing the exam is costly for candidates just at the margin (the RD effects), failing the exam is more costly for people far below the cutoff (the extrapolated effects), especially for the URM group. Moreover, the larger heterogeneous effects among all failing exam takers, compared to those who failed by 10 points, indicate that URM candidates are disproportionately represented farther from the passing cutoff.[19]

4 Disparate impacts of certification exams

This section asks whether certification exams have a *disparate impact* on prospective teachers from URM groups, which is the central legal question in the lawsuits that allege racial discrimination in certification exams. The legal concept of disparate impact does not depend on whether a policy has discriminatory *intent*; rather, the question is whether a policy has unjustifiable negative *impacts* on a protected group.[20] In our context, a disparate impact ex-

[19]Table A9 provides the extrapolated effect estimates of failing certification exam on the baseline outcomes in one, five, and ten years after the exam. Table A10 replicates the exercise but computes the extrapolated effect estimates using the sample that generates the lower bounds on the white-URM gap for the extrapolation, rather than the sample with a score ranging between 240 and 260. Our results still hold when using the lower bounds on the white-URM gap for extrapolation, as in Table A10.

[20]The U.S. Supreme Court Case *Griggs v. Duke Power Co.* (1971) established disparate impacts under Title VII of the civil rights act for employment law (Arnold et al., 2022). In particular, the opinion states "The [Civil Rights] Act proscribes not only overt discrimination but also practices that are fair in form, but discriminatory in operation. The touchstone is business necessity. If an employment practice which operates to exclude [minorities] cannot be shown to be related to job performance, the practice is prohibited." See https://tile.loc.gov/storage-services/service/ll/usrep/usrep401/usrep401424/usrep401424.pdf

ists if two conditions are met: 1) the exams impose a disproportionate cost on URM teaching candidates; and 2) exam performance is not demonstrably related to teaching performance. Having demonstrated the first condition through our RD results in Section 3, we now turn to an analysis of the relationship between certification exam performance and teaching quality.

The main challenge in testing for disparate impact is that, by design, teaching performance is not observed for individuals who do not pass the certification exams. Arnold et al. (2022) show how researchers can overcome a similar data censoring issue to estimate disparate impact in the context of judge bail release decisions. Their analysis exploits the fact that some judges are extremely lenient in their release decisions, which, combined with the random assignment of defendants to judges, allows them to estimate average uncensored outcomes. While their empirical evidence is compelling, their method cannot be directly replicated in settings such as ours that do not have highly lenient decision makers. For example, more than 40 percent of URM exam takers fail their content exam on their first attempt (Table 1), so it is difficult to estimate overall average teaching quality in the entire pool of URM exam takers. We therefore modify the framework in Arnold et al. (2022) to focus on what we call *policy-relevant disparate impact*, which we can estimate by exploiting a reform of the TExES certification exam for elementary teachers. We begin by describing the TExES reform and its impacts on exam passing rates, and then we use variation from the reform to compute policy-relevant disparate impact.

4.1 TExES elementary exam reform. Our disparate impact analysis takes advantage of a reform that increased the difficulty of the TExES certification exams for elementary school teachers. Panel A of Table 3 shows the TExES exams that led to elementary teaching certificates in core subjects between 2002 and 2015. From 2002–2010, Texas offered a Generalist EC-4 exam for certification in early childhood (EC) education through fourth grade. In 2010, the state replaced this exam with the Generalist EC-6 exam, which expanded the certification grade range up to sixth grade. The EC-6 exam was harder than the EC-4 exam because the EC-6 exam covered both more material and more advanced content. Column (E) of Table 3 shows that 82.7 percent of Generalist EC-4 exam takers passed on their first try, while only 65.5 percent of Generalist EC-6 exam takers did so. At the same time, Texas also replaced the Bilingual Generalist EC-4 exam and the English as a Second Language (ESL)/Generalist EC-4 exam with EC-6 versions of these exams, which also increased their difficulty.[21] Contrastingly, the exams for middle school grade levels did not change from

[21]Column (F) of Table 3 shows that the first-attempt pass rates on the Generalist EC-6 exams remain 18–27pp lower in regressions that control for exam takers' high school math and ELA scores, showing that the lower pass rates are not driven by a time trend in the average ability of exam takers. In 2015, the Generalist EC-6 exams were replaced with the Core Subjects EC-6 exam. While the Generalist EC-4 and EC-6 exams produced a single score that determined the pass/fail decision, individuals who took the Core Subjects EC-6

2002–2015. Panel B shows that individuals could be certified to teach grades 4–8 by passing either a Generalist 4-8 exam or by passing subject-specific exams (e.g., Mathematics 4-8).

Panel A of Table 4 shows that the TExES reform caused passing rates on the elementary certification exams to decline substantially, particularly for URM candidates. Prior to the reform (2004–2009), 93 percent of white teaching candidates who took a EC-4 exam passed on their first attempt (column A) as compared with 68 percent of URM teaching candidates (column B). Columns (C)–(E) shows DiD coefficients from regressions that compare Generalist EC-4/EC-6 exam takers (treated group) to individuals who took all other TExES teacher certification exams that were offered in 2004–2015 (control group).[22] The adoption of the EC-6 exams reduced first-attempt passing rates by 14.3pp for white candidates (column D) and by 20.4pp for URM candidates (column E). Although some individuals retook the exams, the reform reduced the proportion of white and URM exam takers who ever passed the Generalist exams by 4.6pp and 12.0pp, respectively. As a result, the URM share of exam takers who passed on the first try fell by 4.3pp, and the URM share of candidates who ever passed fell by 2.6pp (Panel B of Table 4). Panels A–D of Figure 5 display event study coefficients that correspond to the results in Table 4, which show that passing rates declined immediately with reform in 2010 and remained lower in subsequent years. The TExES reform also increased the average number of exam attempts and the average length of time that candidates took to pass the exam (Panel A of Table 4).[23]

The more difficult exam also increased the average ability of individuals who passed on their first attempt, although the average ability of individuals who ever passed increased

were required to pass each subject separately (Math, ELA, Natural Science, and Social Science), and exam takers could retake individual subjects if they failed. We focus our analysis on the switch from the Generalist EC-4 to EC-6 exam because the changes in exam structure are more substantial for the Core Subjects exam.

[22]Specifically, our DiD specification for Table 4 is:

$$Y_i = \gamma_{e(i)} + \gamma_{t(i)} + \beta \text{Generalist}_{e(i)} \text{Post}_{t(i)} + \varepsilon_i, \tag{2}$$

where Y_i is an outcome for exam taker i. The sample includes individuals who took one of the elementary certification exams in Panel A of Table 3 (Generalist EC-4/EC-6, Bilingual Generalist EC-4/EC-6, or ESL/-Generalist EC-4/EC-6) or any of the 62 other TExES teacher certification exams that were offered in every year from 2004–2015. We include fixed effects for exam years, $\gamma_{t(i)}$, and TExES exams, $\gamma_{e(i)}$. We treat each EC-4/EC-6 pair as a single exam for the $\gamma_{e(i)}$ fixed effects. $\text{Generalist}_{e(i)}$ is an indicator for taking one of the Generalist exams. $\text{Post}_{t(i)}$ is an indicator for exam years $t(i) \geq 2010$, when the EC-6 exams were offered. Columns (C)–(E) of Table 4 show the DiD coefficients β estimated separately for all, white, and URM exam takers. Figure 5 shows event study estimates from a version of equation (2) that computes separate $\beta_{t(i)}$ coefficients for each exam year $t(i)$ (omitting 2009).

[23]In Panel C of Table 4, we collapse our data to the exam × year level, and our outcome variables are the log number of exam takers and exam passers. The switch from the EC-4 to EC-6 exam had no significant impact on the total number of first-time exam takers (a 0.035 log point increase), which is notable given that, all else equal, the EC-6 certificate should have been more popular since it allowed individuals to teach a wider range of grades. Yet the number of individuals who passed on the first attempt declined by 18 percent (−0.18 log points), with a particularly large reduction for URM candidates (−0.44 log points).

only modestly. In Panel D of Table 4, we do not find evidence that TExES reform affected the average high school math score of individuals who attempted the Generalist exams, but the switch from the EC-4 to EC-6 exam led to a 0.08 SD increase in the math ability of individuals who passed the exam on their first try, with larger effects for URM candidates (see also Panels E–F of Figure 5). Although this is consistent with the motivation for more difficult exams, the average high school math school of individuals who ever passed increased by only 0.015 SDs, which is driven by a 0.048 SD increase among URM exam takers.[24]

4.2 Policy-relevant disparate impact. This subsection defines our measure of disparate impact and shows how we estimate it (see Appendix C.1 for details). We begin by considering a population of prospective elementary school teachers characterized by their race, potential exam performance, and potential teaching quality. Let $R_i \in \{U, W\}$ denote the race of individual i, where U is URM and W is white. Let $T_i \in \{E, H\}$ denote whether individual i took an easy or hard test as their first certification exam. In our context, $T_i = E$ represents the easier EC-4 TExES exam and $T_i = H$ represents the harder EC-6 exam. Let D_i^E and D_i^H be potential outcomes that indicate for whether individual i would pass the easy and hard exams on their first try. We let $D_i = \mathbb{1}\{T_i = E\}D_i^E + \mathbb{1}\{T_i = H\}D_i^H$ be the observed indicator of whether individual i passed their exam. Lastly, we let Y_{it}^* be individual i's potential teaching quality in year t if they were to become a teacher. Potential teaching quality Y_{it}^* exists for everyone, but we only observe it for individuals who pass their exam and enter the teaching profession.[25] We use teacher value-added as a measure of teaching quality and compare teachers at the same level of potential experience $e(t)$, defined as the number of years since individuals took their first certification exam. We denote observed value-added by Y_{it}.

Given this notation, we define the *policy-relevant disparate impact* (PRDI) at a particular value of teaching quality $Y_{it}^* = y$ and level of potential experience $e(t) = e$ as:

$$\begin{aligned} \text{PRDI}(y, e) \equiv \quad & \Pr(D_i^H = 1 | R_i = U, T_i = H, D_i^E = 1, Y_{it}^* = y, e(t) = e) \\ & - \Pr(D_i^H = 1 | R_i = W, T_i = H, D_i^E = 1, Y_{it}^* = y, e(t) = e). \end{aligned} \tag{3}$$

Equation (3) says there is disparate impact if the pass rate on the hard exam, $\Pr(D_i^H = 1)$, differs between URM (U) and white (W) exam takers who have the same level of value-added, $Y_{it}^* = y$, measured $e(t) = e$ years after the exam. Our measure of disparate impact is policy-relevant in the sense that it corresponds to a particular population that is affected

[24] The results in Panel D of Table 4 are similar if we use exam takers' high school ELA scores.

[25] Although some individuals who initially fail retake the exams and go on to teach, we exclude these individuals from our analysis because of potential selection bias.

by the switch from an easy to a hard exam: individuals who took the hard exam ($T_i = H$) and who *would have passed* the easy exam if they had taken it instead ($D_i^E = 1$). This differs from the population average definition of disparate impact in Arnold et al. (2022), which is equivalent to dropping the $T_i = H$ and $D_i^E = 1$ conditions in equation (3). We focus on policy-relevant disparate impact because it allows for more credible identification assumptions in our context, and because it answers the specific policy question of whether the EC-6 exam (H) had a disparate impact relative to the EC-4 exam (E) for the population of exam takers affected by this reform.[26]

The key object to be estimated in equation (3) is $\text{PassRate}(r, y, e) \equiv \Pr(D_i^H = 1 | R_i = r, T_i = H, D_i^E = 1, Y_{it}^* = y, e(t) = e)$, i.e., the race-specific pass rate on the hard exam conditional on a particular level of value-added $Y_{it}^* = y$ and potential experience $e(t) = e$. By Bayes' rule, this term is equal to:[27]

$$\text{PassRate}(r, y, e) = \Pr(D_i^H = 1 | R_i = r, T_i = H, D_i^E = 1) \times \frac{\Pr(Y_{it}^* = y | R_i = r, T_i = H, D_i^E = 1, D_i^H = 1, e(t) = e)}{\Pr(Y_{it}^* = y | R_i = r, T_i = H, D_i^E = 1, e(t) = e)}. \tag{4}$$

The first term on the right-hand side of equation (4) is the hard exam pass rate in our policy-relevant population ($T_i = H$ and $D_i^E = 1$). To transform this into the pass rate at a particular value of teaching quality $Y_{it}^* = y$ and $e(t) = e$, this term is multiplied by the ratio of two distributions of teacher value-added, Y_{it}^*. The numerator is the distribution of value-added for the subset of our policy-relevant population who *actually passed* the hard exam ($D_i^H = 1$), while the denominator is the value-added distribution for the full policy-relevant population. None of the terms in equation (4) is observed in the data, for three reasons. First, we do not observe which hard exam test takers ($T_i = H$) would have passed the easy exam if they had taken it ($D_i^E = 1$). Second, value-added (Y_{it}^*) is censored for some individuals who would have passed the easy exam but instead took and failed the hard exam. Third, even among exam passers, we do not observe value-added for many individuals, either because they choose not to become teachers, because leave the profession after a period of time, or because they teach grades or subjects without standardized exams.

To identify these unobservable quantities, we make the following three assumptions.

[26] Our analysis will not directly speak to any potential disparate impact of certification exams on individuals who failed the easy exam.

[27] Equation (4) assumes $\Pr(D_i^H = 1 | T_i = H, D_i^E = 1, e(t) = e) = \Pr(D_i^H = 1 | T_i = H, D_i^E = 1)$ for all values of potential experience e. Heterogeneity in hard exam pass rates by potential experience is driven only by the timing of our data, which determines the years in which we observe teaching outcomes for both EC-4 and EC-6 exam takers. There is minimal heterogeneity on this dimension in our data, so we ignore it to simplify our analysis.

Assumption 1. *Any prospective fourth grade teacher who passed the hard exam would also have passed the easy exam:* $D_i^H = 1 \implies D_i^E = 1$.

Appendix Figure A7 provides support for Assumption 1. This figure plots scores for individuals who took *both* the EC-4 and EC-6 exams in 2010 — the transition year in which both exams were offered. Among individuals who took the EC-6 exam before the EC-4 exam, 94 percent earned a higher score on the EC-4 exam. This improvement in scores might partially reflect learning or greater familiarity with the exam on the second attempt, but among individuals who took the EC-4 exam before the EC-6 exam, we also find that 73 percent earned higher scores on the EC-4 exam. In this latter group, those whose did better on the EC-6 exam were individuals whose scores were well below the passing threshold on both exams.[28] More broadly, the EC-6 exam contained the same content as the EC-4 exam except that it also added material pertinent to grades 5–6, so we believe it is plausible that the exam was uniformly more difficult for prospective fourth grade teachers.

Assumption 2A. *Individuals' potential pass rates and value-added are independent of whether they took the easy exam or hard exam:* $T_i \perp\!\!\!\perp D_i^E, D_i^H, Y_{it}^*$.

Assumption 2A states that individuals are as good as randomly assigned to the EC-4 or EC-6 exams. We make this as our benchmark assumption because it simplifies the computation of disparate impact, and we believe it is plausible since the decision to take the EC-4 or EC-6 exam depended primarily on the year in which the individual wanted to begin teaching. But this assumption may be violated if there is a time trend in the potential pass rates or teaching quality of prospective fourth grade teachers. In our robustness analysis below, we relax Assumption 2A by using middle school teachers to control for potential time trends in pass rates or teaching quality (see Assumption 2B in Section 4.4).

Assumption 3. *Among exam passers, the ratio of potential value-added distributions on the hard/easy exams is equal to that for observed value-added distributions (for all* r*,* y*, and* e*):*

$$\frac{Pr(Y_{it}^* = y | R_i = r, T_i = H, D_i = 1, e(t) = e)}{Pr(Y_{it}^* = y | R_i = r, T_i = E, D_i = 1, e(t) = e)} = \frac{Pr(Y_{it} = y | R_i = r, T_i = H, D_i = 1, e(t) = e)}{Pr(Y_{it} = y | R_i = r, T_i = E, D_i = 1, e(t) = e)}$$

Assumption 3 states that, conditional on passing the exam, selection into having value-added is the same among individuals who took the easy and hard exams. While we cannot test this assumption directly, Appendix Table A12 shows that the characteristics of elementary teachers for whom we observe value-added are consistent with the effects of the TExES

[28] Indeed, among individuals who took the EC-4 exam before the EC-6 exam, 90 percent of those who passed the EC-6 exam also passed the EC-4 exam.

reform on the characteristics of exam passers (Table 4). Among teachers with value-added estimates, the URM share is 15pp lower for hard exam passers than for easy exam passers, and hard exam passers have high school math and ELA scores that are roughly 0.1 SDs higher on average. This suggests that there is not significant selection into having value-added conditional on passing the exam. Further, it shows that our disparate impact analysis compares value-added between two teaching populations that differ significantly in racial composition and average ability, as intended.[29]

Under these assumptions, the pass rate for $R_i = r$ at $Y_{it}^* = y$ and $e(t) = e$ is equal to:

$$\text{PassRate}(r, y, e) = \frac{\Pr(D_i = 1 | R_i = r, T_i = H)}{\Pr(D_i = 1 | R_i = r, T_i = E)} \times \frac{\Pr(Y_{it} = y | R_i = r, T_i = H, D_i = 1, e(t) = e)}{\Pr(Y_{it} = y | R_i = r, T_i = E, D_i = 1, e(t) = e)}. \tag{5}$$

We can estimate $\text{PassRate}(r, y, e)$ with two ratios of observable quantities: 1) the ratio of race-specific pass rates, $\Pr(D_i = 1)$, for the hard and easy exams; and 2) the ratio of race-specific distribution of value-added, Y_{it}, for exam takers who passed each exam.

To compute policy-relevant disparate impact, we estimate $\text{PassRate}(r, y, e)$ using equation (5), plug these values into equation (3) to compute $\text{PRDI}(y, e)$, and then average across values of value-added y and levels of potential experience e. As a benchmark, we compute an average measure of $\text{PRDI}(y, e)$ using the observed distribution of teacher value-added for URM teachers who passed the easy (EC-4) exam, i.e., $\Pr(Y_{it} = y | R_i = U, T_i = E, D_i = 1, e(t) = e)$ for each level of potential experience $e(t) = e$. Under our maintained assumptions, this choice means that our measure reflects disparate impact for the average URM candidate who took the hard exam and who would have passed the easy exam. In Appendix C.1, we show that this averaging choice means that our disparate impact measure is:

$$\text{PRDI} = \frac{\Pr(D_i = 1 | R_i = U, T_i = H)}{\Pr(D_i = 1 | R_i = U, T_i = E)} - \frac{\Pr(D_i = 1 | R_i = W, T_i = H)}{\Pr(D_i = 1 | R_i = W, T_i = E)} \times \text{AdjustmentFactor}, \tag{6}$$

where AdjustmentFactor is a function of the value-added distributions for both white and

[29] Even if there is selection into having value-added, this selection is policy-relevant in the sense that the choice of easier or harder certification exams can affect who passes the exam but not who becomes a teacher conditional on passing. A related consideration is that the sample of individuals who took the EC-6 (hard) exam includes some individuals who wished to become grade 5–6, while these individuals do not appear in the EC-4 (easy) exam sample. In the data we find that conditional on passing the EC-6 exam, individuals who go on to teach grades 5–6 have higher average high school test scores. Thus if anything, the fact that the EC-6 sample includes some prospective grade 5–6 would lead us to *underestimate* the magnitude of disparate impact for individuals who wished to teach grades EC–4, since the URM/white gap in passing rates would likely be even larger in this sample.

URM teachers.[30] In other words, our measure of policy-relevant disparate impact is equal to the *raw* ratio of pass rates on the hard and easy exams for URM candidates minus an *adjusted* ratio of hard/easy exam pass rates for white candidates, where the adjustment reflects differences in teaching quality in the two groups. We compute standard errors using an individual-level bootstrap that accounts for variation in both teacher value-added estimates and the sample used to compute disparate impact.

4.3 Estimating disparate impact. We focus our analysis of disparate impacts on fourth grade math and ELA teachers, for two reasons. First, the switch from the Generalist EC-4 and EC-6 exam had the most direct effects on grades K–4 because the certification exams for these grades because unambiguously more difficult.[31] Although fourth grade teachers could also have gained certification by passing the Generalist 4-8 exam, in practice the large majority of new fourth grade teachers took the Generalist EC-4 or EC-6 exams (see Appendix Table A11). Second, our main measure of teaching performance is teacher value-added on math and ELA standardized exams, and we cannot measure value-added in grades K–3 since Texas standardized testing begins in third grade. We think that test score based value-added is an ideal measure of teaching quality for our disparate impact analysis, both because it has been shown to be highly informative of long-run student outcomes (Chetty et al., 2014b) and because the main goal of content certification exams is to ensure that individuals know the core material so they can effectively teach it. We follow the methodology in Chetty et al. (2014a) to compute teacher value-added for fourth grade math and ELA teachers for the years in which our data includes classroom identifiers (2012–2019). Appendix B.2 provides details on our methods for calculating value-added.

As an initial descriptive fact, Figure 6 shows that the distributions of value-added are similar for fourth grade teachers who passed the EC-4 and EC-6 exams. Panels A and B plot kernel densities of math value-added for white and URM teachers, and Panels C and D show distributions ELA value-added for white and URM teachers. The solid lines depict individuals who took the easier EC-4 exam in 2003–2009 as their first certification exam, who passed on their first attempt, and who subsequently went on to teach fourth grade.

[30] Specifically, AdjustmentFactor is equal to:

$$\sum_e \sum_y \frac{\Pr(Y_{it} = y | R_i = U, T_i = E, D_i = 1, e(t) = e)}{\Pr(Y_{it} = y | R_i = W, T_i = E, D_i = 1, e(t) = e)} \times \Pr(Y_{it} = y | R_i = W, T_i = H, D_i = 1, e(t) = e) \times f(e),$$

where $f(e)$ is the probability mass function of potential experience e. If, for example, white teachers were more prevalent at higher levels of value-added than URM teachers and the proportion of higher value-added teachers was higher on the hard exam, then AdjustmentFactor would be less than one. This would adjust the white pass rate ratio downward, reducing the disparity in pass rates between URMs and whites.

[31] The effect of the TExES reform on grade 5–6 teachers is less clear because the EC-6 exam was a pathway to teaching these grades but the EC-4 exam was not.

Similarly, the dashed lines depict individuals whose first certification exam was the harder EC-6 exam in 2011–2015, who passed on their first attempt, and who then taught fourth grade.[32] The distributions of value-added overlap significantly between the two groups of exam passers. In fact, the means of the EC-4 distributions are 0.01–0.03 SDs *higher* than that for the EC-6 exam, as indicated by the vertical lines. This suggests that the harder EC-6 exam was not any more effective at screening out low-quality teachers.

Figure 7 shows that certification exam performance is only weakly related to value-added, particularly nearly the passing thresholds. Using the same samples as in Figure 6, we plot fourth grade teachers' math (Panel A) and ELA (Panel B) value-added against certification exam scores separately by race and whether individuals took the pre- or post-reform elementary certification exams. While there is a positive relationship between certification performance at value-added at scores above 270, scores in the 240–270 range are essentially uncorrelated with value-added on both subjects. URM teachers have systematically higher value-added than white teachers conditional on certification exam score. If the harder post-reform exam was effective at screening out lower-quality teachers, one would expect to see particularly low levels of value-added for teachers who marginally passed the easier pre-reform exam. Yet teachers who passed the pre-reform exam with scores close to the passing threshold (240) have, if anything, higher levels of value-added on average than individuals who passed the post-reform exam, particularly among URM teachers. This casts further doubt on the effectiveness of the harder EC-6 exam as a screen for teaching quality.

Figure 8 presents our benchmark estimates of policy-relevant disparate impact using math (Panel A) and ELA (Panel B) value-added as measures of teaching quality. Each panel shows the key terms in equation (6). The first two bars show the ratio of pass rates on the hard (EC-6) exam and the easy (EC-4) exam computed separately for URM and white exam takers. The third bar shows the white pass rate ratio multiplied by AdjustmentFactor, which accounts for differences in teaching quality between white and URM teachers. For our benchmarks estimates, we compute AdjustmentFactor by grouping teacher value-added into deciles, and we perform these calculations separately for each year of potential experience to ensure that EC-4 and EC-6 teachers are measured at the same number of years since passing the exams. The "Disparate Impact" bar in Figure 8 shows the different between the raw pass rate ratio for URM teachers and the adjusted pass rate ratio for white teachers, which is our estimate of policy-relevant disparate impact.

We find that the harder EC-6 exam had a large disparate impact on URM exam takers

[32]We observe EC-4 exam passers at higher levels of potential experience (i.e., years since the exam) than EC-6 exam passers, so we weight observations in this figure so that the average level of potential experience is the same in the two samples. We omit individuals who took their first certification exam in 2010 because both the EC-4 and EC-6 exams were offered in this year.

relative to the easier EC-4 exam. Panel A of Figure 8 shows that the EC-6 pass rate among individuals who would have passed the EC-4 exam was 66.7 percent for URM test takers and 82.2 percent for white test takers. Adjusting the white pass rate ratio to match the distribution of math value-added among URM teachers changes this pass rate only slightly to 81.3 percent. Thus benchmark estimate of disparate impact is a 14.6pp URM/white gap in EC-6 exam pass rates among individuals who would have passed the EC-4 exam. Panel B shows that adjusted pass rate ratio for White teachers is somewhat lower when we use ELA value-added as a measure of teaching quality (78.4 percent), but this still leaves a large and statistically significant disparate impact of 11.7pp. In both cases, adjusting for differences in teaching quality has only a small effect on our estimates of disparate impact, consistent with the descriptive results in Figures 6 and 7 showing that certification exam performance is only weakly related to value-added. Thus our computations show that the raw disparity in exam pass rates between white and URM prospective teachers is similar to estimates of disparate impact that control for teaching quality.

4.4 Robustness. This section shows that our benchmark disparate impact estimates are robust to alternative identification assumptions and to a variety of different specifications. Appendix Tables A13 and A14 show details on our disparate impact estimates.

We first consider the sensitivity of our results to Assumption 2A, which states that prospective teachers are as good as randomly assigned to the EC-4 or EC-6 exam. We probe this assumption in two ways. First, we estimate placebo disparate impact estimates using grade 7–8 teachers. For this we estimate policy-relevant disparate impact in the same way as we did for fourth grade teachers, but our sample includes individuals who took the middle school certification exams (Panel B of Table 3) and grade 7–8 teachers. Since the middle school certification exams did not change from 2002–2015, these placebo estimates allow us to test for time trends in exam pass rates and teacher value-added that might bias our estimates of disparate impact for fourth grade teachers.

Second, we use grade 7–8 teachers as a control group to explicitly relax Assumption 2A to one that is analogous to a parallel trends assumption in difference-in-differences analyses. Our disparate impact estimands are based on the *ratios* of pass rates and value-added between post-reform (hard) and pre-reform (easy) exam takers (equation 5), so our identification assumptions take the form of "equal ratios" rather than "parallel trends." Specifically, for this specification we replace Assumption 2A with the following assumption:

Assumption 2B. *For both URM and white exam takers, the following three conditions hold:*

(i) The ratio of pass rates for prospective grade 4 teachers on the post- and pre-reform exams would have been the same as that for prospective grade 7–8 teachers in the

absence of the TExES reform.

(ii) The average (across y and e) ratio of the value-added distributions for prospective grade 4 teachers who passed the post- and pre-reform exams would have been the same as that for prospective grade 7–8 teachers in the absence of the TExES reform.

(iii) Among prospective grade 4 teachers, the effect of the increase in exam difficulty from the TExES reform on the distribution of teacher value-added is independent (across values of y and e) of trends in teacher value-added in the absence of the reform.

Assumptions 2B(i) and 2B(ii) are the "equal ratios" assumptions, which state that, in the absence of the reform, pass rates and teacher value-added would have been similar for prospective grade 4 and grade 7–8 teachers for the pre-reform (2003–2009) and post-reform (2011–2015) periods. Since Assumption 2B(ii) is based on an *average* ratio different levels of value-added y and potential experience e, we also make Assumption 2B(iii), which states that the time trend in value-added is unrelated to the effect of the reform on the distribution of value-added.[33] This assumption is plausible given the descriptive evidence in Figures 6 and 7 that value-added is unrelated to exam performance for marginal passers. Further, Appendix Figure A8 shows that the distributions of teacher value-added are similar for grade 7–8 teachers who took the pre- and post-reform exams, suggesting minimal trends in teaching quality over this time period. Given Assumption 2B, we can identify policy-relevant disparate impact using an estimator that we call *ratio-of-ratios*, in which each of the three terms on the righthand side of equation (6) for prospective grade 4 teachers is divided by an analogous term computed using prospective grade 7–8 teachers. Appendix C.2 provides a formal description of Assumption 2B and our ratio-of-ratios estimator.

Figure 8 shows the results from our placebo and ratio-of-ratios estimates, which lends further support for our benchmark disparate impact estimates. We find small and statistically insignificant placebo disparate impact estimates using grade 7–8 teachers, suggesting that our results for fourth grade teachers are not driven by trends in pass rates or teacher value-added. These null effects on middle school teachers mean that our ratio-of-ratios estimates for grade 4 teachers are similar to our benchmark estimates. Specifically, our ratio-of-ratios estimates show a URM/white disparity in pass rates of 14.4pp controlling for math value-added and 9.1pp controlling for ELA value-added.

[33] An alternative to Assumptions 2B(ii) and 2B(iii) is to assume "equal ratios" in teacher value-added for *each* level of value-added y and potential experience e (rather than equals ratios on average). Appendix C.2 shows that this assumption also identifies our estimand of interest. Panel E of Appendix Tables shows that our main results are similar using this y- and e-specific equal ratios assumption, but the confidence intervals are large because when we use granular values of y, the estimator involves taking ratios of values computed with small sample sizes.

Figure 9 examines the robustness of our results to our methodological decisions for computing disparate impact. This figure shows estimates of disparate impact from the following specifications:

A. Recall that in our benchmark specification, values of teacher quality y are defined by deciles of the value-added distribution. Since value-added is a continuous variable, we explore the sensitivity of our results to different levels of grouping value-added in specifications B–E.

B. We group value-added into 20 quantiles rather than into deciles.[34]

C. We estimate kernel densities of value-added and then group the estimated density values into 100 percentiles.

D. We define two value-added groups: above and below the median.

E. We define two value-added groups: the bottom 10% and the top 90%.

F. Since value-added is a *relative* measure of teacher quality, a potential concern is that our value-added estimates may be biased toward zero. This bias could arise if, for example, teachers who passed the EC-6 exam were higher quality than teachers who passed the EC-4 exam and if EC-6 passers also worked in years when many other EC-6 passers we also employed. To address this potential concern, we estimate a leave-out measure of value-added where the scale for test scores is defined only by experienced teachers who were certified before *either* the EC-4 or EC-6 exams were in place. In this specification, the value-added estimates for our sample of EC-4 and EC-6 passers reflect a teacher's performance relative to this set of experienced teachers, which eliminates the source of potential bias toward zero.

G. Since teacher quality varies in other dimensions besides test score value-added, we follow Jackson (2018) in defining a measure of each teacher's non-cognitive using an index of grade retention, attendance, and disciplinary incidents.

H. Our benchmark disparate impact estimates average across levels of teacher value-added y and potential experience e using the distribution of these values for URM teachers who passed the EC-4 exam. In specification H, we instead average across the distribution of these values for *all* teachers who passed the EC-4 exam.

[34] In specification B, in some cases there are no teachers with a given race, exam cohort, and potential experience level who appear in a particular quantile. When this happens, we combine this quantile with the adjacent quantile for all race and exam cohort cells at that experience level to ensure that all values y have a non-zero mass of teachers.

I. In our ratio-of-ratios estimator, a potential concern is that the introduction of the EC-6 exam may have affected pass rates on the grade 4–8 middle school certification exams. Specifically, some individuals who would have taken the grade 4–8 exams to teach grades 5–6 may have instead taken the EC-6 exam when it was offered. As an alternative, we use grade 9 teachers as a control group in our ratio-of-ratios estimator using high school math/ELA certification exams and teacher value-added for end-of-course exams in Algebra I and English I.

Figure 9 shows that our disparate impact estimates are highly robust across all of these specifications. Across specifications, our disparate impact estimates based on Assumption 2A range from 12.3pp to 15.4pp using math value-added as a measure of teacher quality and from 11.1pp to 17.3pp using ELA value-added (Panel A). Similarly, our ratio-of-ratios estimates based on Assumption 2B range from 12.2pp to 16.4pp controlling for math value-added and 9.1pp to 16.0pp controlling for ELA value-added (Panel B). The robustness of our findings is due to the fact that teacher value-added is essentially unrelated to certification exam performance (Figures 6 and 7). The weak relationship between teaching quality and exam performance means that our method of controlling for value-added make little substantive difference; our measure of disparate impact is essentially equal to the raw observed difference in pass rates on the EC-4 and EC-6 exams. Taken together with our findings from Section 3 on the career consequences of failing a certification exam, our results show that the harder EC-6 exam imposed a disproportionate costs on URM exam takers without corresponding benefits for teaching quality.

5 Overall effects of certification exams on teachers and students

Although our results above document both high costs and limited benefits of certification exams, reforms that significantly change the difficulty of certification exams may have additional impacts on teachers and students through supply side channels. Since K–12 education is mandatory and essentially free, the demand for public school teachers is roughly fixed regardless of the supply and quality of potential teachers. In other words, schools will still have to find individuals to fill open teaching vacancies even if harder certification exams reduce the size of the pool of potential teachers. Thus it is unclear how certification exam reforms will impact both the composition of the teaching workforce and student achievement, even if certification exams are not related to value-added.

This section develops a new identification strategy that allows us to examine the overall

impacts of certification exam difficulty on teachers and students. We identify a set of open teaching vacancies by using the departure of experienced teachers as a shock to schools' labor supply (Section 5.1). Our main specification is an RD-DiD regression model that combines variation from teacher departures with variation from the TExES exam reform (Section 5.2). Together these two sources of variation identify the effects on teacher composition and student achievement when a school is faced with a vacancy to fill under a harder or easier certification exam regime. Importantly, our estimates circumvent an issue, censoring value-added, faced in the previous literature by instead examining student achievement on net.

5.1 Teacher departures. To examine the supply-side impacts of the TExES reform, we bring in a second source of variation: the departure of experienced teachers. Intuitively, departures are a shock to schools' labor supply that create open teaching vacancies, which allows us to ask how demand for new teachers affects teacher composition under easier or harder certification exams. We define a *teacher departure* as an instance in which a math or ELA teacher with five or more years of experience leaves a given school. To isolate large changes in teacher composition, we restrict our analysis to cases in which the departing teacher taught one-third or more of the students in a given school, grade, and subject in the year before their departure.[35] This strategy is analogous to research that uses worker deaths as a shock to firm labor supply to ask how firms find substitutes for their workers (e.g., Jäger and Heining, 2022). Our strategy also follows Chetty et al. (2014a) in using teacher departures to identify impacts of teacher quality.

To use all possible variation from teacher departures, we create a "stacked" dataset for each teacher departure from a given school, grade, and subject. We first collapse our individual-level data to the school/grade/subject/year level. We let s denote school/grade/-subject triplets and use t to denote years, so the variables in our collapsed dataset are mean teacher characteristics at the st level. A given school/grade/subject may experience multiple teacher departures during the period of our data. Thus we let y denote the year of a teacher departure, and we "stack" our collapsed dataset so that st observations occur multiple times for each departure. (We drop st observations for which there is no teacher departure.) Lastly, we let $\tau_{ty} = t - y$ denote years *relative* to the teacher departure, where $\tau_{ty} = 0$ is the first year in which the teacher is no longer in the school.

Since teacher departures may be caused in part by school-specific trends, we use an RD

[35]For grades 7–8, we focus on two subjects: math or ELA (English + reading). For grades 3–4, our main specification uses a single core subject (the combination of math, ELA, science, social studies, and generic) since most elementary teachers teach all core subjects. We focus on departures of teachers with 5+ years of experience so that the departing teacher likely took a different certification exam than the exam that was currently offered to new teachers. See Appendix B.3.1 for details on our definition of teacher departures.

model to isolate sharp changes in teacher composition due to departures.[36] Specifically, we use our collapsed and stacked dataset to estimate a local linear RD regression:

$$Y_{st} = \beta \mathbf{1}\{\tau_{ty} \geq 0\} + \alpha \tau_{ty} + \psi \mathbf{1}\{\tau_{ty} \geq 0\}\tau_{ty} + \gamma_{sy} + \varepsilon_{sty} \quad \text{if } |\tau_{ty}| \leq h^Y. \tag{7}$$

The dependent variable, Y_{st} is an average teacher characteristic or student outcome at the school/grade/subject (s) and year (t) level. Our variable of interest is an indicator for years after the teacher departure, $\mathbf{1}\{\tau_{ty} \geq 0\}$. The running variable is years relative to the teacher departure, τ_{ty}, and we include an interaction between τ_{ty} and $\mathbf{1}\{\tau_{ty} \geq 0\}$. We include fixed effects for school/grade/subject/departure-year quadruplets, γ_{sy}, so that identification comes only from before and after variation within the same departure event. The regression includes years relative to departure, τ_{ty}, that are within the Calonico et al. (2019) RD bandwidth, h^Y, computed separately for each outcome Y.[37] Standard errors are clustered at the school level to allow for correlation in outcomes within the same school (e.g., if schools ask teachers to switch grades following a departure). The coefficient of interest, β, estimates the projected change in average outcomes in the first year after the teacher departure ($\tau_{ty} = 0$). For example, if departing teachers are less likely to be URMs than the teachers that replace them, we would find $\beta > 0$ for the outcome of the URM share of teachers.

5.2 RD-DiD specification. Our full specification combines our teacher departure RD specification with DiD variation in exposure to the TExES reform. Specifically, we estimate our RD regression (7) separately for grades that were more and less affected by the exam reform (treated/control groups) and for teacher departures that occurred before and after the reform (pre/post variation). Our treated group includes grades 3–4 because the vast majority of new teachers in these grades took the Generalist EC-4 or EC-6 exams during this period (see Appendix Table A11). We use grades 7–8 as our control group because neither the EC-4 exam nor the EC-6 exam qualifies one to teach these grades. Our pre-period includes teacher departures that occurred in $y \in$ 2005–2010, when the Generalist EC-4 exam was in place. Our post-period includes departures in $y \in$ 2011–2016, when the EC-6 exam was in place.[38] We estimate equation (7) separately for each pairwise combination of grade group $g \in \{$3–4, 7–8$\}$ and departure period $p \in \{$2005–2010, 2011–2016$\}$, which gives four RD coefficients

[36] Jäger and Heining (2022) use a DiD specification because worker deaths are plausibly exogenous. As we show below, a DiD model does not work as well in our case because teachers may choose to leave a school if its performance is on a negative trend. The RD specification helps to address this issue by separating sharp changes in teacher composition from longer-term trends that may cause teacher turnover.

[37] We weight observations in equation (7) by the product of a triangular kernel (based on the RD bandwidth) and the number of individual teachers or students that were used to compute the outcome variable, Y_{st}.

[38] The EC-6 exam was introduced in 2010, but we consider teacher departures in $y = 2011$ to be the first post-reform period since most individuals do not get a teaching job right away after passing the exam.

β_{gp}.

Lastly, we use these RD coefficients as the dependent variable in a simple DiD regression:

$$\beta_{gp} = \phi \text{Treated}_g + \delta \text{Post}_p + \theta \text{Treated}_g \text{Post}_p + \varepsilon_{gp}, \tag{8}$$

where Treated_g is a dummy for grades 3–4 and Post_p is a dummy for teacher departures in $y \in$ 2011–2016. Equation (8) gives our RD-DiD specification. The coefficient of interest, θ, shows how the effects of teacher departures (as estimated by the RD coefficients) changed with the TExES reform in grades 3–4 relative to grades 7–8. For example, if the TExES reform made it more difficult for elementary schools to hire URM teachers relative to middle schools, then we would find $\theta < 0$ for the outcome of the URM share of teachers.

Although equations (7)–(8) present our RD-DiD specification in two steps to build intuition, in practice we estimate a single-step specification by plugging equation (8) into equation (7).[39] Our single-step RD-DiD specification is equivalent to estimating the RD equation (7) separately for each gp pair and then constructing the DiD estimator: $\theta = \Big(\beta_{\text{Treated,Post}} - \beta_{\text{Treated,Pre}}\Big) - \Big(\beta_{\text{Control,Post}} - \beta_{\text{Control,Pre}}\Big)$.

The key identification assumption in our RD-DiD strategy is parallel trends *in the RD coefficients*. Specifically, we assume that the effects of teacher departures on teacher composition (and student outcomes in Section 5.4) as estimated by the RD coefficients would have trended in the same manner for grades 3–4 and grades 7–8 in the absence of the TExES reform. One might be concerned that the standard RD assumption of "no threshold manipulation" may be violated in our case if, for example, teacher departures cause students to change schools. But our RD-DiD specification allows for this type of behavioral response provided that it does not change differentially in response to the TExES reform. This is analogous to the "difference-in-IV" estimator in Alsan et al. (2025), in which violations of the standard IV assumptions are permissible provided that they are the same in the pre- and post-periods. Below we test the validity of our parallel trends assumption by using a variety of student characteristics as outcome variables, Y_{st}.

Our main output in Tables 5 and 6 show our RD-DiD estimates for the effects of the TExES reform on teacher composition and student achievement, respectively. Column (A)–

[39] In other words, our single-step RD-DiD specification is:

$$\begin{aligned} Y_{st} = \big(\phi \text{Treated}_g + \delta \text{Post}_p + \theta \text{Treated}_g \text{Post}_p\big)\mathbf{1}\{\tau_{ty} \geq 0\} + \\ \alpha_{gp}\tau_{ty} + \psi_{gp}\mathbf{1}\{\tau_{ty} \geq 0\}\tau_{ty} + \gamma_{sy} + \varepsilon_{sty} \quad \text{if } |\tau_{ty}| \leq h^Y_{gp}. \end{aligned} \tag{9}$$

Note that we estimate separate running variable coefficients α_{gp} and ψ_{gp} for each grade group g and departure period p pair. We also compute Calonico et al. (2019) RD bandwidths h^Y_{gp} separately for each gp pair. Grade groups g are a function of the school/grade/subject triplet s, i.e., $g = g(s)$. Departure periods p are a function of teacher departure years y, i.e., $p = p(y)$. Thus the fixed effects γ_{sy} subsume dummies for gp pairs.

(C) present results for grade 3–4 teacher departures. Column (A) shows the mean of each outcome in the year prior to the departure ($\tau_{ty} = -1$) in school/grades that experienced a departure in 2011–2016. Columns (B)–(C) show RD coefficients β from equation (7) estimated separately for departures in 2011–2016 (post-reform) and 2005–2010 (pre-reform). Similarly, columns (D)–(E) show coefficients β for grades 7–8 teacher departures in the post- and pre-reform periods, respectively. Column (F) shows our main object of interest, the RD-DiD coefficient θ from equation (8), which is equal to column (B) − column (C) − (column (D) − column (E)). Appendix Figures A10 – A12 show RD graphs for selected outcomes.

5.3 Effects of exam difficulty on teacher composition. We first show that our RD-DiD approach identifies open teaching vacancies with differential exposure to the Generalist EC-4 and EC-6 exams, as intended. Panel A of Table 5 shows that grade 3–4 teachers who departed in 2005–2010 were often replaced with teachers who took the EC-4 exam, whereas elementary teachers who departed in 2011–2016 were more likely to be replaced with EC-6 exam takers. Our RD-DiD estimate shows that the TExES reform reduced the proportion of newly-hired elementary teachers who took the EC-4 exam by 14.4pp and increased the proportion of newly-hired elementary teachers who took the EC-6 exam by 7.6pp. The reform also increased the difficulty of newly-hired elementary teachers' exams as measured by the average pass rate and the teachers' own exam scores (Appendix Table A16).

Our main finding on teacher composition is that the TExES reform reduced the URM share of newly-hired elementary teachers. The RD-DiD estimate for the URM share of teachers decreased by −1.9pp, which roughly matches our estimate for the URM share of exam passers from Table 4 (−1.5pp). This suggests that the more difficult EC-6 exam reduced the share of URMs in the pool of certified teachers. The TExES reform reduced the share of elementary teaching vacancies that were filled by new teachers (within five years of certification) by 3pp, driven primarily by a 2.5pp reduction in the share of new URM teachers. By contrast, the share of elementary teaching vacancies that were filled by white teachers who earned a certification more than five years ago increased by 1.7pp.

This change in teacher composition meant that total earnings declined for URM teachers in elementary schools and while total earnings for white teachers increased. Our outcome variables in Panel C of Table 5 are the sum of earnings to teachers in a given school/grade/-subject/year in cell.[40] Overall, we find a $6,300 increase in payment to white teachers, whereas the total payment to URM teachers fell by $2,300 with the reform.[41]

[40]Teacher departures typically lead to a decline in total teacher pay since the newly-hired teachers tend to make lower salaries than the departing teachers. But relative to middle schools, the TExES reform led to a $6,800 increase the total payment to teachers in elementary schools with departing teachers, consistent with the higher experience of the newly-hired teachers.

[41]We also consider teachers' year-to-year earnings gains as our outcome variable and find similar results

5.4 Effects of exam difficulty on student achievement. We find that our RD-DiD estimates are balanced along student composition. We find no systematic evidence that the TExES reform affected the number of students in school/grades with departing teachers or their demographic characteristics. Our estimates on demographic composition, including race, ethnicity and gifted status are insignificant, with one exception: the proportion of students at risk of dropping out ($\theta = -2.0pp$).

Turning to our main results on student achievement, we find that the harder elementary certification exams neither increased nor decreased student test scores. Panel B of Table 6 shows RD-DiD estimates on student math and ELA scores. Our outcome variables in these panels also include test score residuals, which are the residuals from a value-added-like regression that includes individual, school/grade mean, and school mean student characteristics and lagged test scores. The RD coefficients in columns (B)–(E) show that teacher departures have mostly small and insignificant effects on student achievement on average in either time period and for either grade level. Our RD-DiD estimates in column (F) also show that these RD coefficients did not change differentially between grades 3–4 and grades 7–8 with the TExES reform. We find positive estimates for ELA scores and ELA score residuals (0.007 SDs and 0.009 SDs) and negative estimates for math scores and math score residuals (-0.014 SDs and -0.000 SDs), but all estimates are statistically insignificant.

Table 7 shows test score effects using an alternative specification that exploits variation across math and ELA teacher departures within grades 3–4. Although the majority of grade 3–4 teachers teach all core subjects, some schools in our data have teachers that specialize in either math or ELA. In Table 7, we restrict our sample to school/grades with teachers who specialize in either subject and estimate our RD-DiD specification using the departure of one of these specialized teachers. Panel A asks how student math score changes relative to ELA test scores when the math teacher departs. Panel C asks how math scores change in school/grades with a departing math teacher relative to school/grades with a departing ELA teacher. Analogously, Panels B and D show ELA scores relative to math.[42]

These alternative specifications show evidence that the TExES reform *reduced* student achievement in math. In Table 7, the RD-DiD coefficients for demographics suggest our specifications that restrict identification to only grades 3–4 are balanced. In Panels A and C, we find negative and statistically significant RD-DiD coefficients for both math scores and math score residuals, with estimates ranging from -0.029 SDs to -0.055 SDs. In

(Appendix Table A16). Thus harder certification exams not only have disproportionate costs on prospective teachers from URM groups, but they also reduce the share of teaching salaries that go to URMs as a whole.

[42]Panels A–B exploit "within school" variation by comparing changes in test scores across the two subjects, while Panels C–D exploit "across school" variation by comparing changes in scores in a given subject across departing teachers who teach different subjects.

other words, the effect of math teacher departures on student math scores declined with the reform relative to ELA scores for the same students (Panel A) and relative to math scores at other schools with departing ELA teachers (Panel C). These results could be driven by the reduction in the supply for elementary teachers documented in Table 4, which may have made it harder for schools to find qualified math teachers. We do not find statistically significant effects on ELA test scores following the departure of ELA teachers in either Panel B or Panel D.

Our finding that the TExES reform did not improve student achievement is robust to other specifications. A caveat with our RD-DiD approach is that it only estimates effects in the near the cutoff, which is the first year of a teacher departure ($\tau_{ty} = 0$). In Appendix Table A17, we modify the RD regression (7) so that our RD-DiD specification estimates changes in outcomes from three years before to three years after the teacher departure ($\tau_{ty} = -3$ to $+3$). We continue to find small and statistically insignificant effects on student test scores in this specification, although there is imbalance on student demographics in some specifications (which is why we prefer the RD model that estimates effects at $\tau_{ty} = 0$). Our point estimates for math and ELA scores are relatively stable and statistically insignificant across a wide range of RD bandwidths, although we find demographic imbalance using the ELA score index with large bandwidth (Appendix Figure A14).

We do not find systematic patterns of heterogeneity in the RD-DiD estimates for student test scores by race, although we find a statistically significant 0.028 SD decline in math scores for white students (Appendix Table A18).

In sum, we find that the switch from the easier EC-4 exam to the harder EC-6 exam reduced the share of teachers from URM groups and without any measurable improvements in student achievement, and some evidence of negative effects on student math performance. These results corroborate our findings that certification exams impose disproportionate costs on URM prospective teachers (Section 3) and are not effective at screening out teachers with lower value-added (Section 4).

6 Conclusion

Dating back to the 1980s, numerous lawsuits have been filed alleging that teacher certification exams have a disparate impact on prospective teachers from underrepresented minority (URM) groups, which violates Title VII of the U.S. Civil Rights Act. The legal argument for disparate impact relies on two claims: 1) the exams impose disproportionate costs on URM teaching candidates because they have lower pass rates; and 2) the exams are not sufficiently related to teaching performance to justify the pass rate disparities.

Both of these claims are difficult to prove, which explains the mixed success of these lawsuits. While New York City recently agreed to pay $1.8 billion to Black and Hispanic teachers to settle a more than 25-year lawsuit over its certification tests,[43] similar lawsuits were withdrawn in Texas and Arkansas after plaintiffs failed to show that the exams imposed a significant cost on individuals who failed.[44] In a lawsuit between the Association of Mexican-American Educators and the State of California, both District and Circuit Court judges ruled in favor of the State, arguing that it had successfully shown that the exam was a valid measure of job performance.[45]

The academic literature too provides little conclusive evidence on the costs and benefits of certification exams. No study to date has estimated the economic costs of failing certification exams for prospective teachers. While there is a large literature that correlates certification exam performance with teacher value-added (e.g., Goldhaber, 2007; Clotfelter et al., 2010), these analyses suffer from the fundamental problem that teaching quality is not observed for candidates who do not manage to pass the exams. The vast majority of U.S. states continue to require that individuals pass certification exams to receive a teaching license.

This paper presented new evidence on both the economic costs and predictive validity of certification exams. Using administrative data from Texas on certification exams, K–12 teachers, and students, we found that individuals who marginally fail certification exams are delayed from starting their teaching careers, ultimately costing them roughly $10,000 in forgone earnings. Not only are Black and Hispanic teaching candidates more likely to bear these costs given their lower pass rates, but they also have more negative causal effects of failing. Ultimately, we found that increases in the difficulty of certification exams reduce the share of teachers who are Black and Hispanic, thus contributing to their under-representation in the profession relative to the racial and ethnic composition of students.

Our paper also showed that more difficult certification exams do not increase the quality of the teaching workforce. For this we exploited a reform that significantly increased the difficulty of Texas' certification exam for elementary teachers. This reform allowed us to observe the teaching quality of individuals who passed the easier pre-reform exam but would not have passed the harder post-reform exam, thus overcoming the overcoming the typical challenge inherent in measuring certification exam validity under plausible assumptions. We found that the distribution of teacher value-added was similar for teachers who went through the easier and harder exams, showing that the harder exam was not a more effective screen

[43] See Green (2023) and https://clearinghouse.net/case/11541/.

[44] The Texas and Arkansas lawsuits were withdrawn because most candidates who initially failed ended up passing after subsequent retaking the exam, and thus it was argued that the exams did not create a significant harm. See: https://www.edweek.org/education/arkansas-union-drops-suit-over-teacher-tests/1987/11.

[45] See: https://clearinghouse.net/case/10826/.

for teaching quality. A complementarity RD-DiD analysis showed that the switch to the harder elementary certification exam did not improve student achievement on standardized tests, and we found evidence of reductions in math scores in some specifications.

Taken together, our findings support the claim that certification exams have a disparate impact on URM teaching candidates. Within the population of prospective teachers who would have passed the easier pre-reform exam, our preferred estimates suggest that URM teachers had 10–15pp lower pass rates on the harder exams than white teachers with comparable teaching quality. Combined with our evidence on the career consequences of failing exams, this suggests that certification exams impose disproportionate economic costs on URM teaching candidates.

An important caveat is that our findings pertain to the elementary teacher certification exams in Texas. It is possible that certification exams are more informative for teaching quality at higher grade levels, but we cannot credibly test for such effects given the nature of the Texas exam reform that underlies our analysis. More broadly, examinations are widely used to determine certification and licensing in other professions, and they are also a requirement for high school graduation in some states. These exams typically exhibit large disparities in pass rates between white and URM exam takers, but it is similarly challenging to measure their validity and economic consequences. We hope future research will shed light on both the costs and potential benefits of certification exams in other settings.

References

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the chicago public high schools. *Journal of labor Economics*, 25(1):95–135.

Agency, T. E. (2022). Requirements for certified educators and non-certified employees. https://tea.texas.gov/texas-educators/investigations/fingerprinting/requirements-for-certified-educators-and-non-certified-employees. Accessed 2022.

Alsan, M., Barnett, A., Hull, P., and Yang, C. S. (2025). "Something works" in US jails: Misconduct and recidivism effects of the IGNITE program. *The Quarterly Journal of Economics*, 140(2):1367–1415.

Angrist, J. D. and Guryan, J. (2008). Does teacher testing raise teacher quality? evidence from state certification requirements. *Economics of Education Review*, 27(5):483–503.

Arnold, D., Dobbie, W., and Hull, P. (2022). Measuring racial discrimination in bail decisions. *American Economic Review*, 112(9):2992–3038.

Baron, E. J., Doyle Jr, J. J., Emanuel, N., Hull, P., and Ryan, J. (2024). Discrimination in multiphase systems: Evidence from child protection. *The Quarterly Journal of Economics*, 139(3):1611–1664.

Benson, A., Board, S., and Meyer-ter Vehn, M. (2024). Discrimination in hiring: Evidence from retail sales. *Review of Economic Studies*, 91(4):1956–1987.

Blair, P. Q. and Chung, B. W. (2025). Job market signaling through occupational licensing. *Review of Economics and Statistics*, 107(2):338–354.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2025a). Inaccurate statistical discrimination: An identification problem. *Review of Economics and Statistics*, 107(3):605–620.

Bohren, J. A., Hull, P., and Imas, A. (2025b). Systemic discrimination: Theory and measurement. *The Quarterly Journal of Economics*, page qjaf022.

Bond, T. N. and Lang, K. (2018). The black–white education scaled test-score gap in grades k-7. *Journal of Human Resources*, 53(4):891–917.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics*, 101(3):442–451.

Card, D. and Rothstein, J. (2007). Racial segregation and the black–white test score gap. *Journal of Public Economics*, 91(11-12):2158–2184.

Cattaneo, M. D., Jansson, M., and Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–79.

Chung, B. W. and Zou, J. (2025). Occupational licensing in us public schools: Nationwide implementation of teacher performance assessment. *Journal of Public Economics*, 244:105328.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2010). Teacher credentials and student achievement in high school a cross-subject analysis with student fixed effects. *Journal of Human Resources*, 45(3):655–681.

Cowan, J., Goldhaber, D., Jin, Z., and Theobald, R. (2020). Teacher licensure tests: Barrier or predictive tool? working paper no. 245-1020. *National Center for Analysis of Longitudinal Data in Education Research (CALDER)*.

Fryer Jr, R. G. and Levitt, S. D. (2006). The black-white test score gap through third grade. *American law and economics review*, 8(2):249–281.

Goldhaber, D. (2007). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of human Resources*, 42(4):765–794.

Goldhaber, D., Gratz, T., and Theobald, R. (2017). What's in a teacher test? assessing the relationship between teacher licensure test scores and student stem achievement and course-taking. *Economics of Education Review*, 61:112–129.

Goldhaber, D. and Hansen, M. (2010). Race, gender, and teacher testing: How informative a tool is teacher licensure testing? *American Educational Research Journal*, 47(1):218–251.

Goncalves, F. and Mello, S. (2021). A few bad apples? racial bias in policing. *American Economic Review*, 111(5):1406–41.

Green, E. (2023). The $1.8-billion lawsuit over a teacher test. *The New Yorker*.

Hendricks, M. (2016). Teacher characteristics and productivity: Quasi-experimental evidence from teacher mobility. *Available at SSRN 2822041*.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non-test score outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Jäger, S. and Heining, J. (2022). How substitutable are workers? evidence from worker deaths. NBER Working Paper No. 30629.

Jencks, C. and Phillips, M. (1998). *The Black-White test score gap*. Brookings Institution Press.

Kane, T. J., Rockoff, J. E., and Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education review*, 27(6):615–631.

Kline, P., Rose, E. K., and Walters, C. R. (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics*, 137(4):1963–2036.

Larsen, B., Ju, Z., Kapor, A., and Yu, C. (2020). The effect of occupational licensing stringency on the teacher quality distribution. Technical report, National Bureau of Economic Research.

Law, M. T., Marks, M., and Stern, T. (2023). Teacher testing standards and the new teacher pipeline. *Available at SSRN 4644356*.

Law, M. T. and Marks, M. S. (2009). Effects of occupational licensing laws on minorities: Evidence from the progressive era. *The Journal of Law and Economics*, 52(2):351–366.

Lepage, L.-P. (2024). Experience-based discrimination. *American Economic Journal: Applied Economics*, 16(4):288–321.

Orellana, A. and Winters, M. A. (2023). Licensure tests and teacher supply. Technical report, Working Paper. Retrieved from https://alexisorellana. github.io/assets/pdf ....

Penney, J. (2017). Test score measurement and the black-white test score gap. *Review of Economics and Statistics*, 99(4):652–656.

Shuls, J. V. and Trivitt, J. R. (2015). Teacher effectiveness: An analysis of licensure screens. *Educational Policy*, 29(4):645–675.

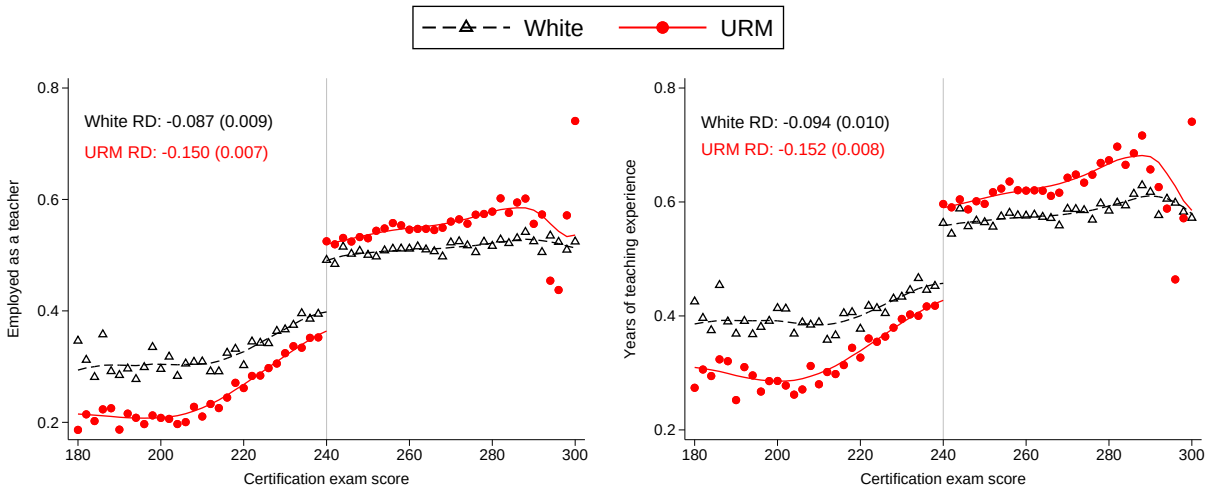
Small, M. L. and Pager, D. (2020). Sociological perspectives on racial discrimination. *Journal of Economic Perspectives*, 34(2):49–67.

Templeton, T., Sands, S. R., Lowrey, S., Horn, C. L., and Narvaez, J. (2022). Texas classroom teacher certification. Technical report, University of Houston Education Resource Center.

Xia, X. (2021). Barrier to entry or signal of quality? the effects of occupational licensing on minority dental assistants. *Labour Economics*, 71.

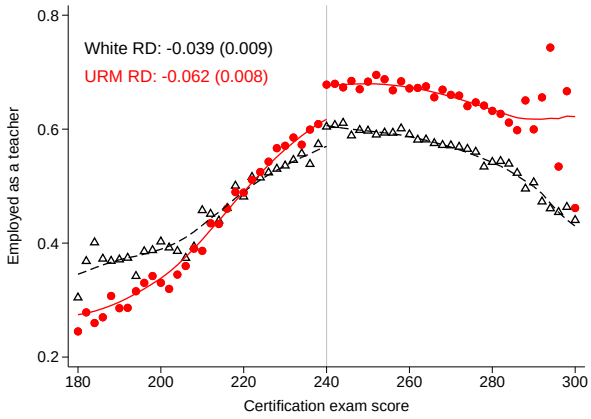
Yan, Z. (2024). The effect of occupational licensing on the gender wage gap. *Available at SSRN 5090345*.

Figures

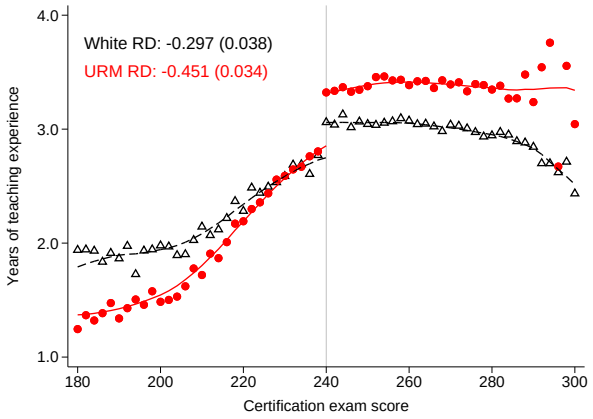


Panel A. Employed as a teacher 1 year later

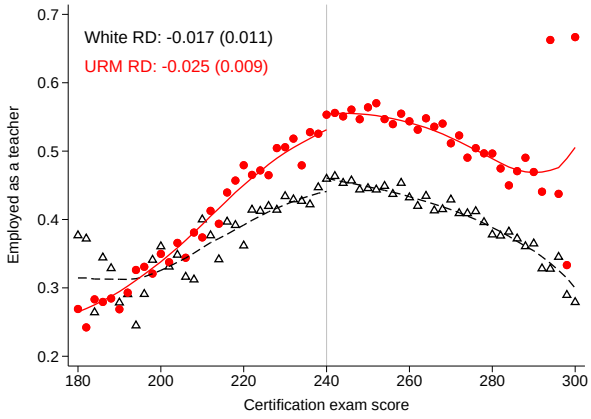
Panel B. Years of teaching experience 1 year later



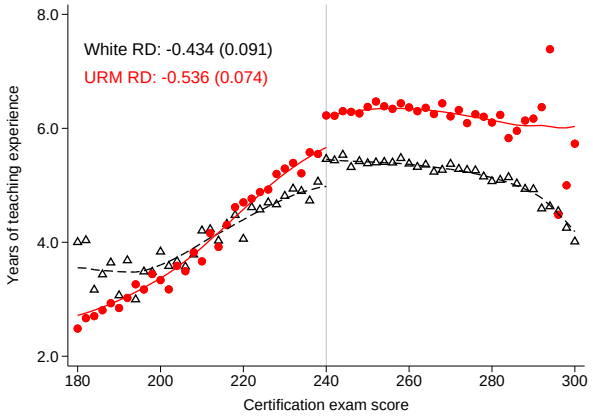
Panel C. Employed as a teacher 5 years later



Panel D. Years of teaching experience 5 years later



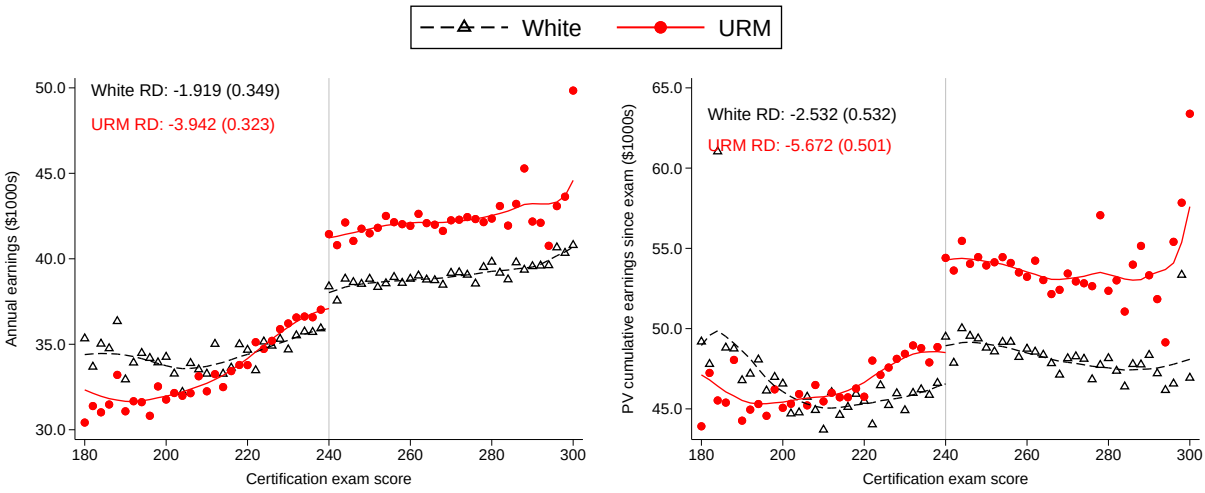
Panel E. Employed as a teacher 10 years later



Panel F. Years of teaching experience 10 years later

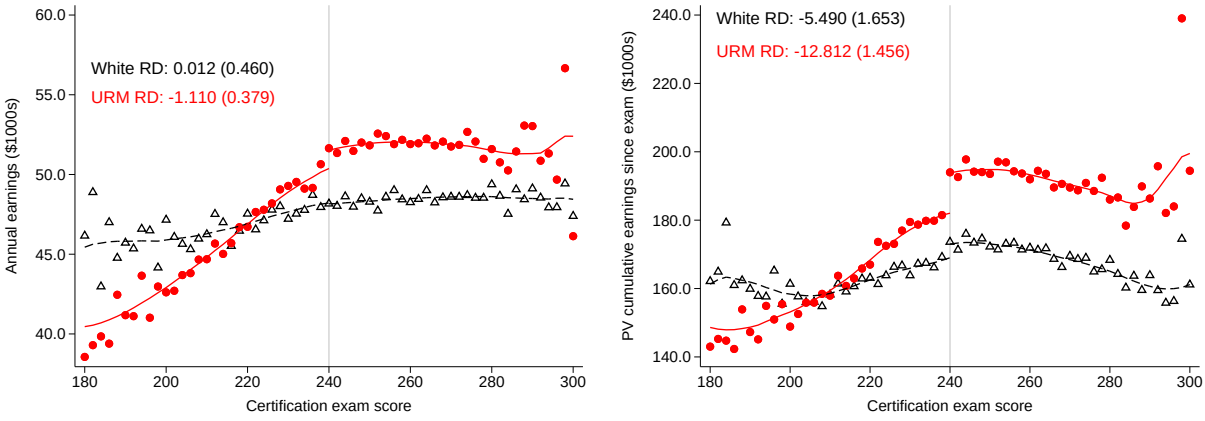
Figure 1: RD plots: Teaching outcomes by race

Notes: The figure presents the RD effects of failing a certification exam on teaching outcomes by white and URM groups. Panels A-F display the RD effects on teacher employment and years of teaching experience one, five, and ten years after the exam, respectively. The RD estimates are shown at the upper-left corner, with standard errors included in the parentheses.



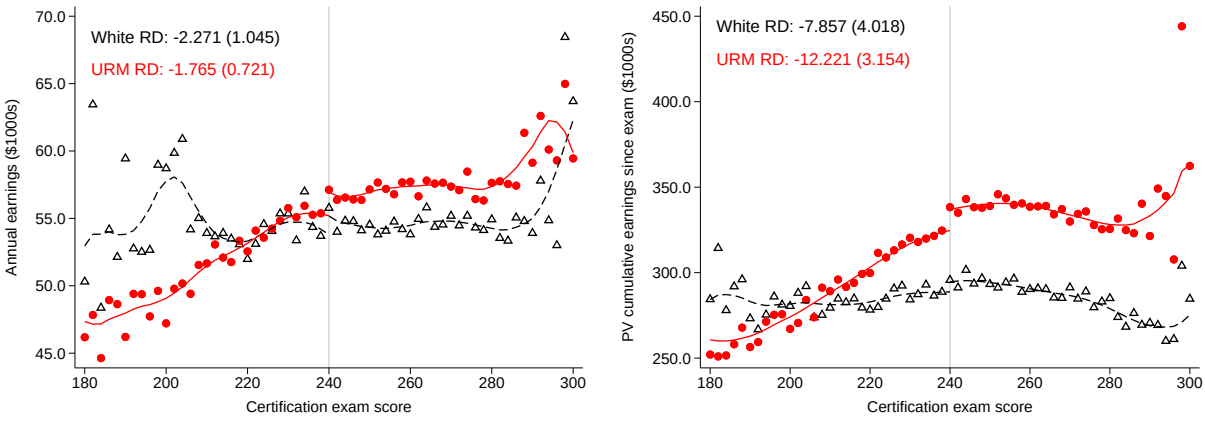
Panel A. Annual earnings 1 year later ($1000s)

Panel B. PV cum. earnings after 1 year ($1000s)



Panel C. Annual earnings 5 years later ($1000s)

Panel D. PV cum. earnings after 5 years ($1000s)

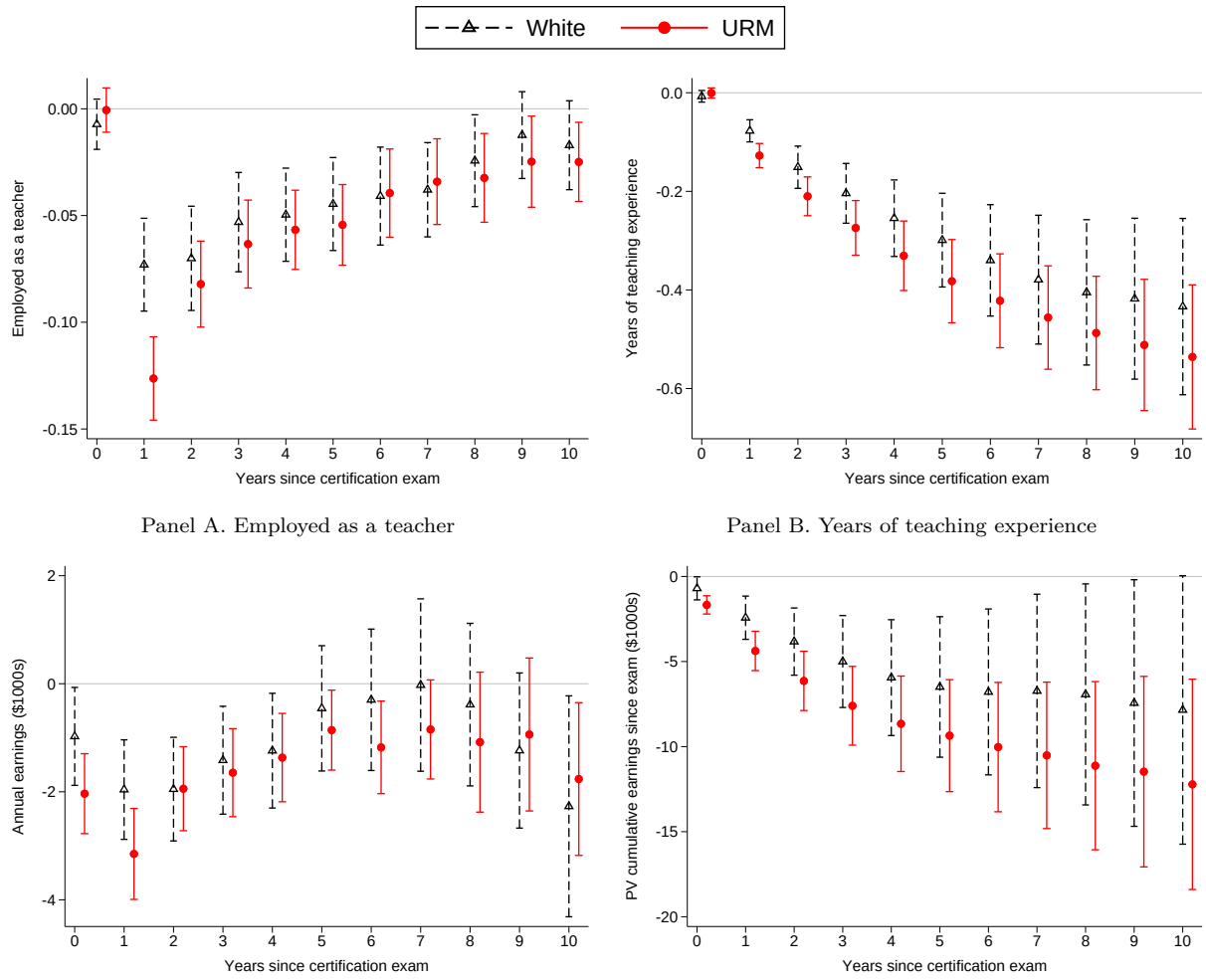


Panel E. Annual earnings 10 year later ($1000s)

Panel F. PV cum. earnings after 10 years ($1000s)

Figure 2: RD plots: Earnings outcomes by race

Notes: The figure presents the RD effects of failing a certification exam on earnings outcomes by white and URM groups. Panels A-F display the RD effects on annual earnings and the present value of cumulative earnings one, five, and ten years after the exam, respectively. The RD estimates are shown at the upper-left corner, with errors deviations included in the parentheses.



Panel A. Employed as a teacher

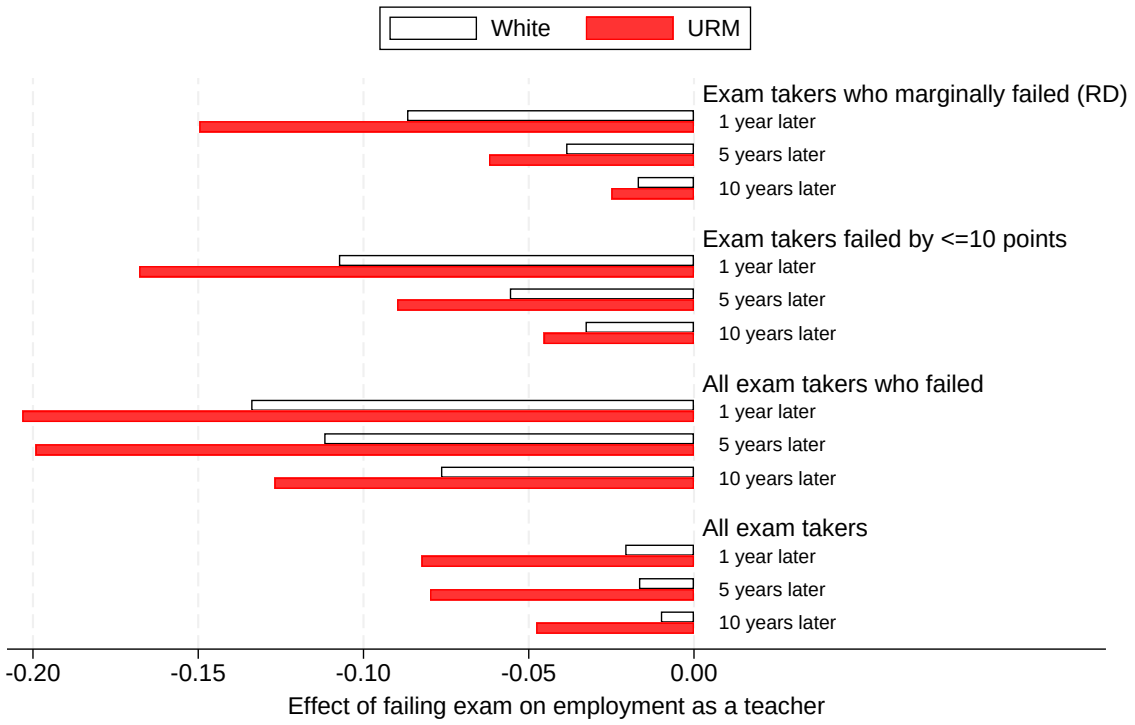
Panel B. Years of teaching experience

Panel C. Annual earnings ($1000s)

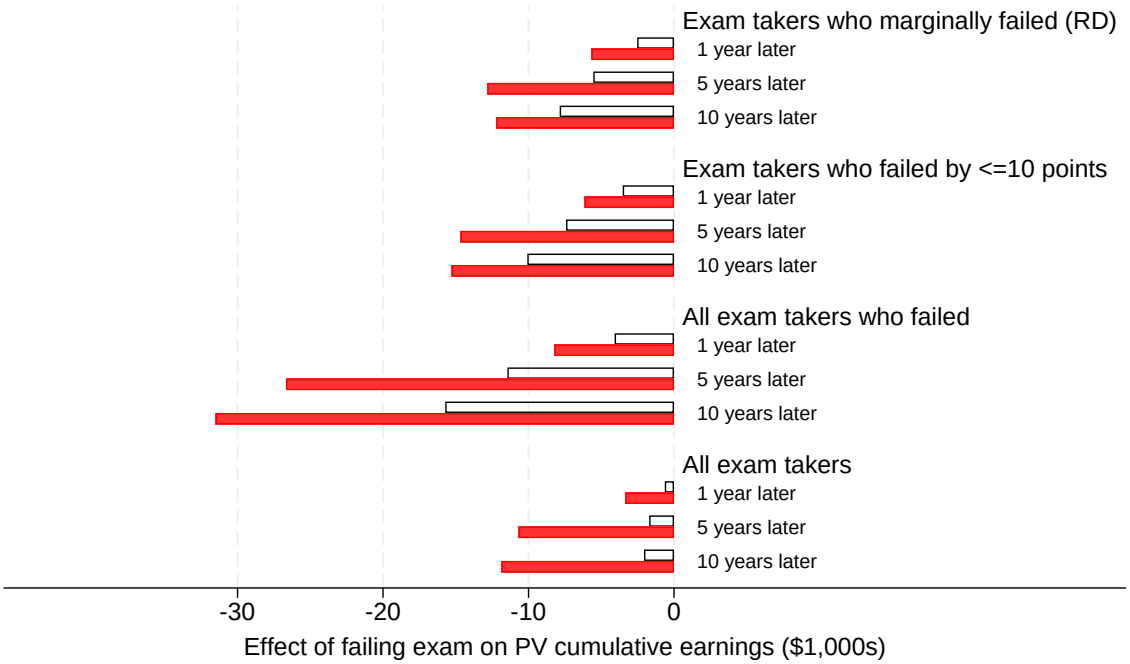
Panel D. PV cumulative earnings since exam ($1000s)

Figure 3: RD effects of failing a certification exam over time — Balanced panel

Notes: The figure presents the RD effects of failing a certification exam on teaching and earnings outcomes by white and URM groups. Panels A-F display the RD effects on teaching employment, years of teaching experience, annual earnings, and the present value of cumulative earnings, respectively. The figure shows lasting RD effects over each of the post-period up to ten years after failing the exam, using a balanced panel sample.



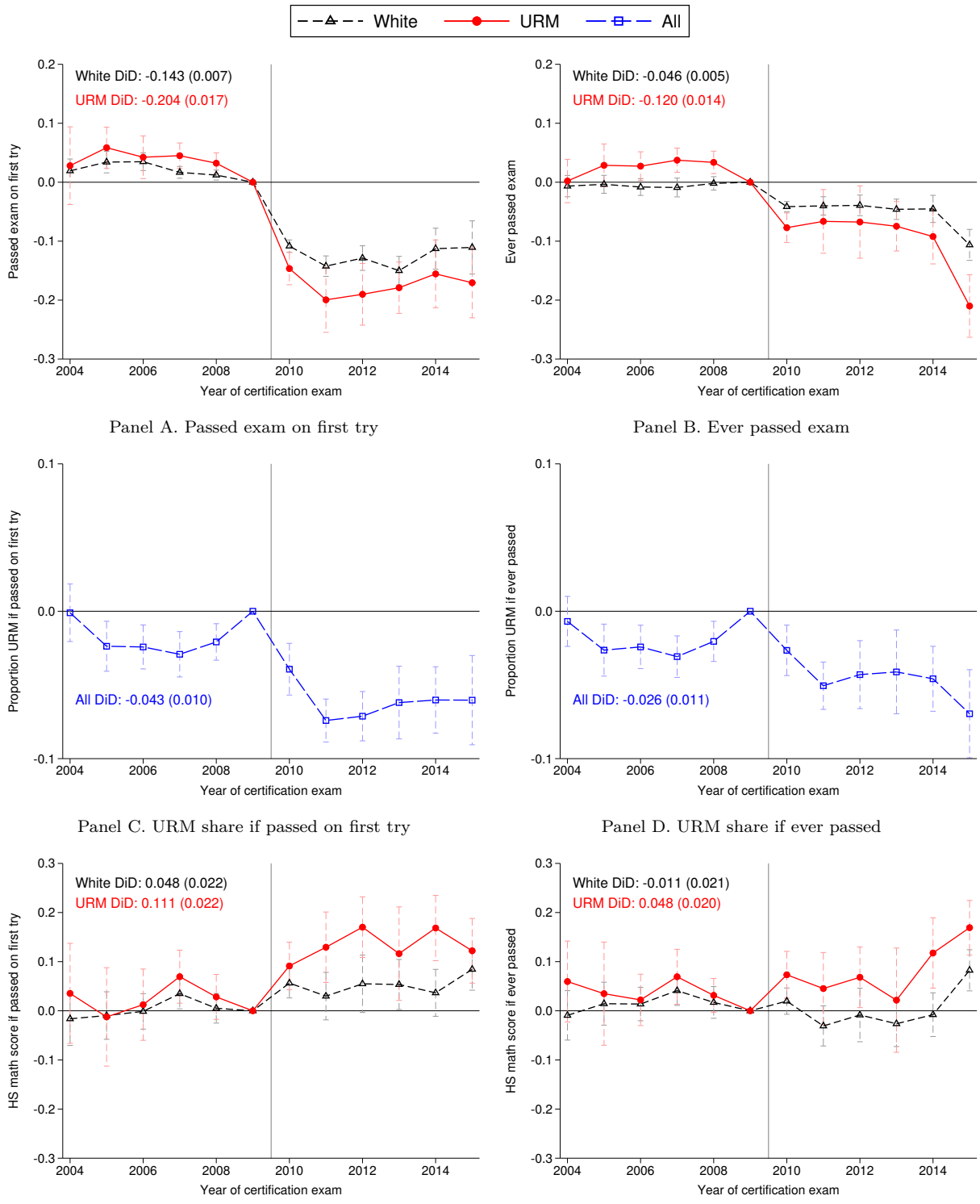
Panel A. Employed as a teacher



Panel B. PV cumulative earnings since exam ($1000s)

Figure 4: Extrapolated effects of failing certification exam on outcomes

Notes: The figure summarizes the baseline RD effects and extrapolated effects of the three samples — exam takers with a score within ten points of the passing score, all exam takers with a score below the 240 cutoff, and all exam takers — by white and URM groups in one, five, and ten years after the exam. Panels A and B display the effects on teaching employment and the present value of cumulative earnings, respectively.



Panel A. Passed exam on first try

Panel B. Ever passed exam

Panel C. URM share if passed on first try

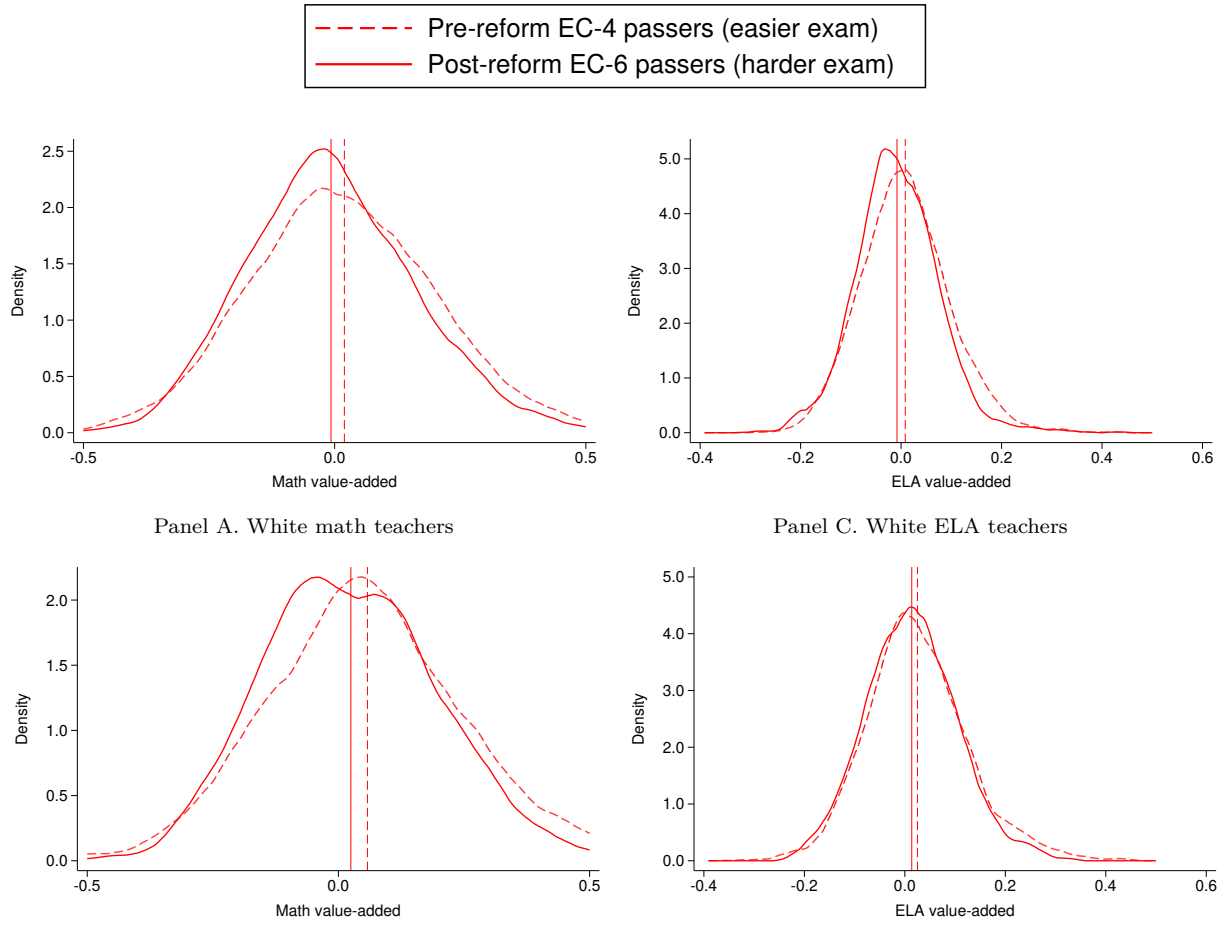
Panel D. URM share if ever passed

Panel E. HS math score if passed on first try

Panel F. HS math score if ever passed

Figure 5: Event studies: Effects of harder certification exams on exam outcomes

Notes: This figure shows the effects of the TExES reform on exam outcomes (Panels A–B) and the composition of individuals who took/passed the exams (Panels C–F). The graphs display estimates from an event-study version of equation (2) that computes separate $\beta_{t(i)}$ coefficients for each exam year $t(i)$ (omitting 2009). Each figure displays the β coefficient from equation (2) and with its standard error in parentheses. We estimate these regressions separately for white, URM, or all exam takers, as indicated by the legend. Dashed lines contain 95 percent confidence intervals using standard errors clustered at the exam level.



Panel A. White math teachers

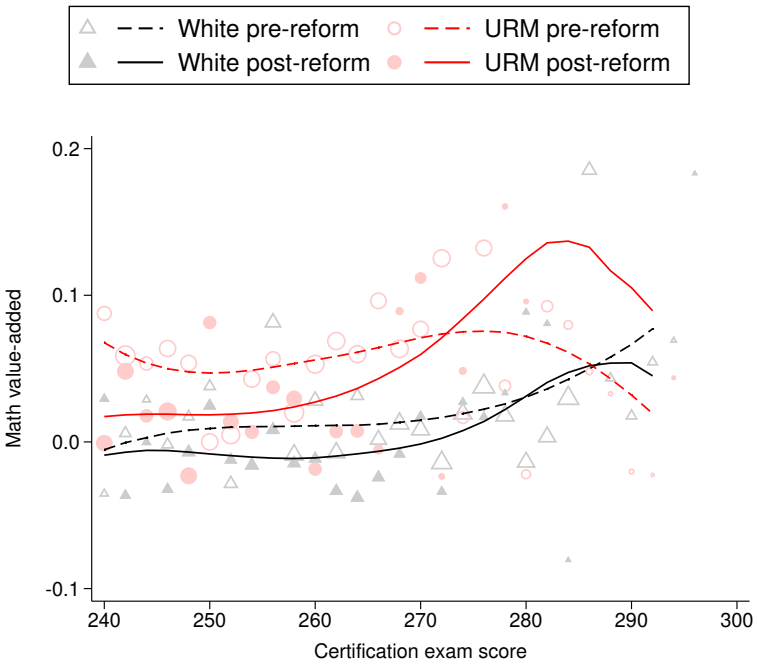
Panel C. White ELA teachers

Panel B. URM math teachers

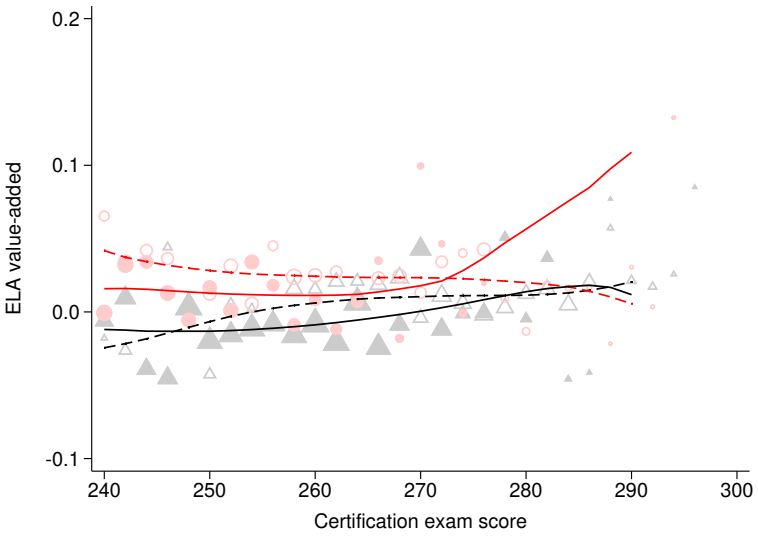
Panel D. URM ELA teachers

Figure 6: Value-added distributions for fourth grade teachers measured 3–8 years after the certification exam

Notes: This figure shows kernel densities of math (Panels A–B) and ELA (Panels C–D) value-added for white and URM individuals who passed different TExES certification exams. The sample includes individuals who took the EC-4 or EC-6 exams (Panel A of Table 3) as their first certification exam, who passed the exam, and who subsequently became fourth grade teachers. The graphs plot distributions of grade 4 value-added for individuals who took the EC-4 exam in 2003–2009 (dashed lines) or the EC-6 exam in 2011–2015 (solid lines). Panels A and C show estimates for white teachers, and Panels B and D show estimates for URM teachers. We measure value-added 3–8 years after taking the certification exam ("potential experience"), and weight observations so that the distribution of potential experience is the same for the two exams in each graph (using the exam with the minimum number of observations for each level of potential experience). Vertical lines denote the mean of each distribution.



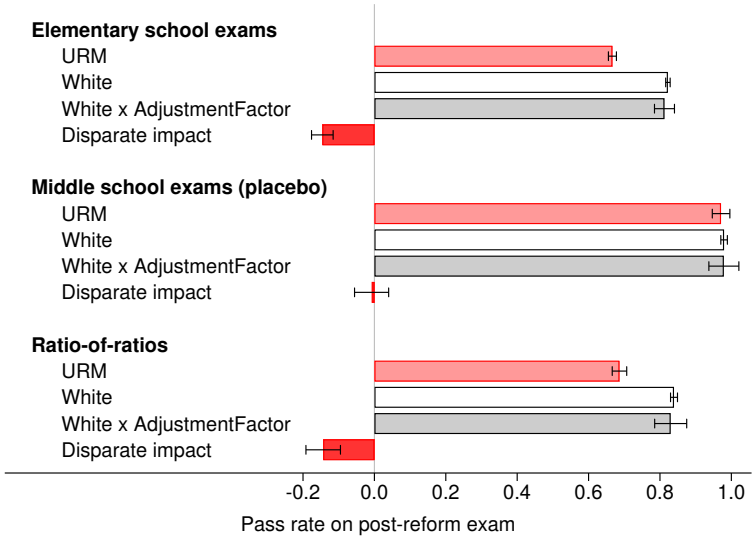
Panel A. Math value-added



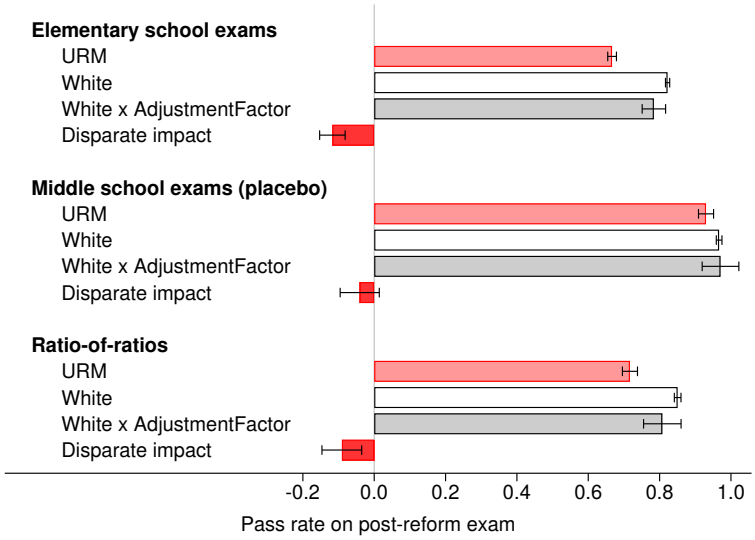
Panel B. ELA value-added

Figure 7: Value-added of fourth grade teachers by certification exam score

Notes: This figure shows the relationship between teacher value-added (y-axis) and certification exam scores (x-axis) for fourth grade math (Panel A) and ELA (Panel B) teachers. The sample includes individuals who took the EC-4 or EC-6 exams (Panel A of Table 3) as their first certification exam, who passed the exam, and who subsequently became fourth grade teachers. Markers depict average value-added within bins of certification exam scores, computed separately by race (white and URM) and for individuals who took the EC-4 exam in 2003–2009 or the EC-6 exam in 2011–2015, as indicated by the legend. We measure value-added 3–8 years after taking the certification exam ("potential experience"), and weight observations so that the distribution of potential experience is the same for the two exams in each graph (using the exam with the minimum number of observations for each level of potential experience). Lines depict predicted values from local linear regressions.



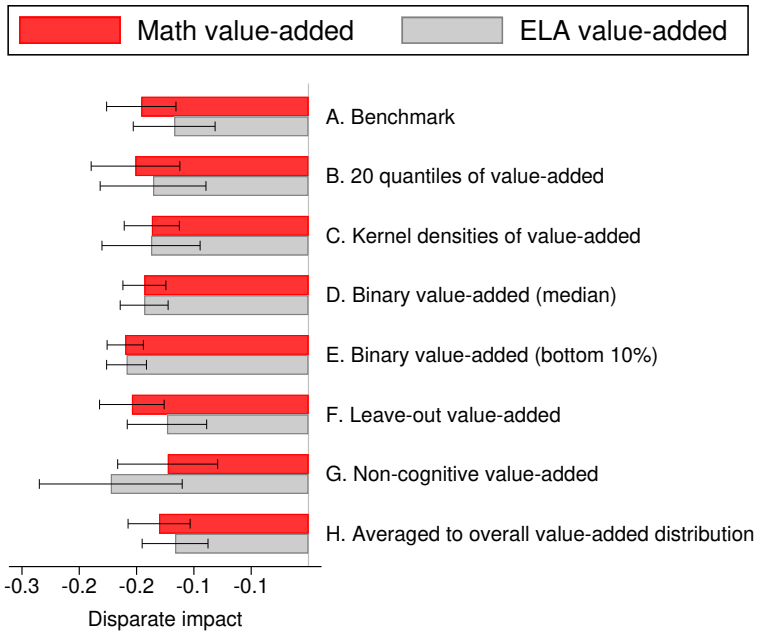
Panel A. Math value-added



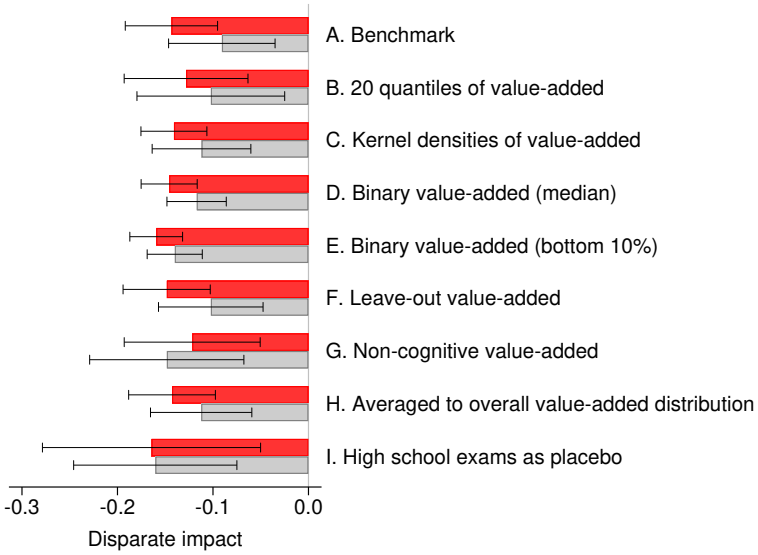
Panel B. ELA value-added

Figure 8: Policy-relevant disparate impact of EC-6 exam relative to EC-4 exam

Notes: This figure shows estimates of the policy-relevant disparate impact of the EC-6 exam relative to EC-4 exam using math value-added (Panel A) and ELA value-added (Panel B) as measures of teaching quality. For the elementary school certification exams, the first two bars show the ratio of pass rates on the EC-6 and EC-4 exam computed separately for URM and white exam takers. The third bar shows the white ratio of pass rates after it has been adjusted to match the distribution of value-added for grade 4 URM teachers, as defined by equation (6). The fourth bar shows our estimate of policy-relevant disparate impact, which is equal to the raw URM ratio of pass rates minus the adjusted white ratio of pass rates. We do the same calculation for middle school certification exams using pre-reform (2003–2009) and post-reform (2011–2015) exams to compute a placebo measure of disparate impact for grade 7–8 teachers. The last block in each panel reports our ratio-of-ratios estimator defined in Appendix C.2. For this we take ratio of URM, white, or adjusted white pass rates for grade 4 teachers and divide it by the corresponding ratio of pass rates for grade 7–8 teachers. Our ratio-of-ratios estimate of policy-relevant disparate impact is equal to the URM ratio-of-ratios pass rate minus the adjusted white ratio-of-ratios pass rate. Horizontal brackets denote 95 percent confidence intervals computed using bootstrap sampling at the individual level. See Appendix Tables A13 and A14 for details on these computations.



Panel A. Elementary school exams only



Panel B. Ratio-of-ratios

Figure 9: Disparate impact robustness checks

Notes: This figure shows estimates of the policy-relevant disparate impact of the EC-6 exam relative to EC-4 exam using a variety of specifications. Red bars present estimates using math value-added as a measure of teaching quality, and gray bars present estimates using ELA value-added as a measure of teaching quality. Panel A shows estimates from equation (6), which uses only grade 4 teachers. Panel B shows estimates from our ratio-of-ratios estimator defined in Appendix C.2, which uses teachers in higher grade levels as a control group. Specification A shows our benchmark results. Specifications B–E vary the level at which we group values of teacher value-added y. Specification F uses a leave-out measure of value-added where the scale for test scores is defined only by experienced teachers who were certified before *either* the EC-4 or EC-6 exams were in place. Specification G defines teacher value-added using a non-cognitive index based on grade retention, attendance, and disciplinary incidents. Specification H compute average disparate impact using the distributions of teacher value-added y and potential experience e for *all* teachers who passed the EC-4 exam (rather than for URM teachers). Specification I uses grade 9 teachers as a control group in our ratio-of-ratios estimator (rather than grade 7–8 teachers). Horizontal brackets denote 95 percent confidence intervals computed using bootstrap sampling at the individual level. See Section 4.4 for details on these specifications. See Appendix Tables A13 and A14 for details on the computations.

Tables

Table 1: Summary statistics for TExES exam takers and comparison populations in 2002–2019

	(A)	(B)	(C)	(D)	(E)	(F)
	First-time certification exam takers			First-year teachers	New BA graduates	K–12 students
	White	URM	All	All	All	All
Panel A. Demographics						
Female	0.777	0.742	0.762	0.747	0.581	0.487
White	1.000	0.000	0.588	0.590	0.524	0.326
Black	0.000	0.307	0.119	0.113	0.094	0.134
Hispanic	0.000	0.693	0.269	0.261	0.265	0.488
HS math score	0.626	0.250	0.483	0.552	0.676	0.043
HS ELA score	0.700	0.307	0.546	0.602	0.632	0.041
Panel B. TExES content exam performance						
Content exam score on first attempt	257.8	241.5	252.1	254.0		
Passed content exam on first attempt	0.833	0.573	0.739	0.787		
Ever passed first content exam	0.925	0.769	0.863	0.927		
Ever passed any content exam	0.959	0.851	0.913	0.979		
Panel C. TExES PPR exam performance						
PPR exam score on first attempt	266.6	254.6	262.5	263.3		
Passed PPR exam on first attempt	0.945	0.807	0.894	0.907		
Ever passed first PPR exam	0.992	0.955	0.976	0.987		
Ever passed any PPR exam	0.994	0.966	0.982	0.992		
Panel D. Teaching in public schools						
Taught in year of exam/first teaching/graduation	0.165	0.129	0.192	0.975	0.001	
Taught 1 year later	0.429	0.374	0.445	0.863	0.084	
Taught 5 years later	0.596	0.607	0.568	0.644	0.130	
Taught 10 years later	0.449	0.518	0.438	0.491	0.125	
Panel E. Annual earnings (2019 USD)						
Earnings in year of exam/first teaching/graduation	26,492	26,909	29,158	47,716	24,529	
Earnings 1 year later	37,195	37,481	39,286	48,491	38,482	
Earnings 5 years later	49,422	50,668	50,528	51,806	57,513	
Earnings 10 years later	54,631	55,832	55,016	56,504	75,787	
N (# unique individuals)	292,785	193,094	662,332	348,800	1,472,457	13,313,603

Notes: The table presents the summary statistics (mean) for TExES exam takers in the first three columns and comparison populations in the last three columns (first-year teachers, bachelor degree recipients, and K-12 students, respectively) between 2002 and 2019. The sample first-time certification exam takers is further separated into Whie and URM groups. Panel A displays demographic variables, including gender, race, and high school achievement. Panel B presents TExES content exam performance, including exam scores and passing rates. Panel C reports the TExES PPR exam performance, including the exam scores and passing rate. Panel D reports the fraction of individuals teaching in Texas public schools in the year of the exam and one, five, and ten years later. Panel E shows earnings (in 2019 US dollars), including earnings in the year of exam/first teaching/graduation and earnings one, five, and ten years later. "N (# unique individuals)" indicates the unique number of observation for each of the populations.

Table 2: RD effects of failing a certification exam on outcomes

	(A)	(B)	(C)	(D)	(E)	(F)
	Mean above threshold		RD coefficients			*P*-value: W=URM
	White	URM	All	White	URM	
Panel A. 1 year after certification exam						
Employed as a teacher	0.497	0.525	-0.118***	-0.087***	-0.150***	0.000
			(0.005)	(0.009)	(0.007)	
Years of teaching experience	0.565	0.597	-0.122***	-0.094***	-0.152***	0.000
			(0.007)	(0.010)	(0.008)	
Annual earnings ($1000s)	38.239	41.432	-2.831***	-1.919***	-3.942***	0.000
			(0.213)	(0.349)	(0.323)	
PV cumulative earnings since exam ($1000s)	49.084	54.467	-4.062***	-2.532***	-5.672***	0.000
			(0.321)	(0.532)	(0.501)	
N (Employed as a teacher)	16,683	16,896	506,345	227,740	156,499	384,239
Panel B. 5 years after certification exam						
Employed as a teacher	0.608	0.677	-0.048***	-0.039***	-0.062***	0.046
			(0.006)	(0.009)	(0.008)	
Years of teaching experience	3.076	3.342	-0.364***	-0.297***	-0.451***	0.003
			(0.026)	(0.038)	(0.034)	
Annual earnings ($1000s)	48.275	51.687	-0.552*	0.012	-1.110***	0.058
			(0.290)	(0.460)	(0.379)	
PV cumulative earnings since exam ($1000s)	173.604	194.717	-9.062***	-5.490***	-12.812***	0.001
			(1.039)	(1.653)	(1.456)	
N (Employed as a teacher)	14,528	14,425	442,510	201,864	133,272	335,136
Panel C. 10 years after certification exam						
Employed as a teacher	0.459	0.553	-0.020***	-0.017	-0.025***	0.587
			(0.007)	(0.011)	(0.009)	
Years of teaching experience	5.477	6.248	-0.464***	-0.434***	-0.536***	0.390
			(0.059)	(0.091)	(0.074)	
Annual earnings ($1000s)	54.772	56.658	-1.595***	-2.271**	-1.765**	0.697
			(0.559)	(1.045)	(0.721)	
PV cumulative earnings since exam ($1000s)	296.041	338.626	-9.668***	-7.857*	-12.221***	0.400
			(2.390)	(4.018)	(3.154)	
N (Employed as a teacher)	9,848	9,734	310,404	145,790	88,718	234,508

Notes: The table presents the RD estimates of the impacts of failing a certification exam on teaching and earning outcomes. Panels A-C indicate the RD effects in one, five, and ten years after the exam, respectively. The dependent variables include four outcomes: an indicator of whether being employed as a teacher, years of teaching experience, annual earnings, and the present value of cumulative earnings. Columns (A) and (B) display the dependent variable means of candidates with exam scores ten points above the threshold. Columns (C)-(E) show the RD coefficients using equation (1). Column (F) indicates the p-value testing the statistical differences of RD effects between white and URM groups in Columns (D) and (E), respectively. "N (Employed as a teacher)" indicates the sample size for the outcome "Employed as a teacher", serving as a proxy sample size for other outcomes. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: TExES certification exams for math and ELA teachers, 2002–2015

(A) TExES Exam	(B) First offered	(C) Last offered	(D) # exam takers	(E) First try pass rate	(F) Pass rate w/ HS score controls
Panel A. Elementary school exams					
Generalist EC-4	2002	2010	137,375	0.827	
Generalist EC-6	2010	2015	140,609	0.655	-0.180
Bilingual Generalist EC-4	2002	2010	29,397	0.642	
Bilingual Generalist EC-6	2010	2015	17,290	0.442	-0.242
English as a Second Language (ESL)/Generalist EC-4	2003	2010	8,577	0.892	
English as a Second Language (ESL)/Generalist EC-6	2010	2015	10,937	0.655	-0.265
Panel B. Middle school exams					
Generalist 4-8	2002	2015	83,059	0.895	
Bilingual Generalist 4-8	2002	2015	4,038	0.383	-0.462
English as a Second Language (ESL)/Generalist 4-8	2003	2015	3,785	0.640	-0.245
Mathematics 4-8	2002	2022	49,905	0.646	-0.275
English Language Arts and Reading 4-8	2002	2022	37,785	0.852	0.004
Mathematics/Science 4-8	2002	2022	8,635	0.696	-0.238
English Language Arts and Reading/Social Studies 4-8	2002	2022	9,994	0.842	-0.044

Notes: This table shows the TExES exams for elementary (Panel A) and middle (Panel B) school math and ELA teachers that were offered between 2002 and 2015. Column (A) shows the name of each exam. Columns (B) and (C) show the first and the last year in which exam was offered (up through the end of our data in 2022). Column (D) shows the total number of exam takers in 2002–2015. Column (E) shows the proportion of exam takers who passed on their first attempt at the exam. For column (F), we regress an indicator for passing the exam on the first attempt on a cubic in teachers' high school math and ELA scores plus dummies for each exam. We estimate these regressions separately for each pair of EC-4/EC-6 exams in Panel A and for all exams in Panel B, omitting the dummy for the first exam in each group. Column (F) shows the coefficients on the exam dummies in these regressions.

Table 4: Effects of harder certification exams on exam outcomes

	(A)	(B)	(C)	(D)	(E)	(F)
	Pre-reform means		DiD coefficients			*P*-value: White=URM
	White	URM	All	White	URM	
Panel A. Exam outcomes						
Passed exam on first try	0.930	0.677	-0.154***	-0.143***	-0.204***	0.000
			(0.011)	(0.007)	(0.017)	
Ever passed exam	0.979	0.849	-0.075***	-0.046***	-0.120***	0.000
			(0.009)	(0.005)	(0.014)	
Number of times taking this exam	1.102	1.521	0.375***	0.312***	0.594***	0.000
			(0.028)	(0.018)	(0.054)	
Days between first try and passing	9.311	48.476	31.650***	26.237***	54.117***	0.000
			(3.339)	(2.496)	(6.460)	
Panel B. Racial distribution of exam takers						
URM share of first-time exam takers		0.420	-0.008			
			(0.012)			
URM share of exam takers who passed on first try		0.346	-0.043***			
			(0.010)			
URM share of exam takers who ever passed		0.386	-0.026**			
			(0.011)			
Panel C. Number of exam takers						
Log number of first-time exam takers	8.552	7.671	0.035	0.057	-0.058	0.494
			(0.109)	(0.074)	(0.173)	
Log total number of exams taken	8.634	8.027	0.274**	0.299***	0.247	0.784
			(0.126)	(0.074)	(0.196)	
Log number passed on first try	8.482	7.300	-0.180	-0.103	-0.441**	0.050
			(0.111)	(0.070)	(0.169)	
Log number ever passed	8.533	7.513	-0.063	0.009	-0.240	0.169
			(0.118)	(0.069)	(0.181)	
Panel D. HS test scores of exam takers						
HS math score of first-time exam takers	0.516	0.208	-0.024	-0.035*	0.002	0.000
			(0.019)	(0.019)	(0.013)	
HS math score if passed on first try	0.570	0.372	0.077***	0.048**	0.111***	0.000
			(0.024)	(0.022)	(0.022)	
HS math score if ever passed	0.530	0.274	0.015	-0.011	0.048**	0.000
			(0.023)	(0.021)	(0.020)	
N (# exam takers)	38,185	28,674	393,452	180,292	118,633	

Notes: This table shows the effects of the TExES reform on exam outcomes (Panel A) and the composition of individuals who took/passed the exams (Panels B–D). Columns (A)–(B) show means of each outcome in the pre-reform years (2004–2009) for white and URM exam takers. Columns (C)–(E) show β coefficients from the DiD regression (2) estimated separately for all, white, and URM exam takers. Column (F) shows *P*-values from a test of equality of the white and URM coefficients in columns (D)–(E). Standard errors in parentheses are clustered at the exam level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: RD-DiD effects of harder certification exams on teacher composition

	(A)	(B)	(C)	(D)	(E)	(F)
		Grade 3–4 departures		Grade 7–8 departures		
	Post-reform mean at τ_{ty} = -1	Post-reform RD	Pre-reform RD	Post-reform RD	Pre-reform RD	RD-DiD
Panel A. Teachers' cert exams						
Took the Generalist EC-4 exam	0.415	-0.014***	0.146***	0.002	0.018***	-0.144***
		(0.005)	(0.005)	(0.003)	(0.002)	(0.008)
Took the Generalist EC-6 exam	0.134	0.107***	0.002***	0.029***	-0.000	0.076***
		(0.004)	(0.000)	(0.003)	(0.000)	(0.005)
Panel B. Teacher demographics						
Number of teachers	3.960	0.078***	-0.028	0.042***	0.012	0.077**
		(0.017)	(0.024)	(0.010)	(0.012)	(0.034)
URM	0.397	0.002	0.015***	0.021***	0.014***	-0.019**
		(0.003)	(0.003)	(0.004)	(0.004)	(0.008)
Recently-certified teacher (within 5 years)	0.223	0.190***	0.206***	0.250***	0.236***	-0.030***
		(0.005)	(0.006)	(0.007)	(0.006)	(0.012)
Recently-certified URM teacher	0.097	0.073***	0.069***	0.087***	0.058***	-0.025***
		(0.004)	(0.004)	(0.004)	(0.004)	(0.007)
Recently-certified White teacher	0.120	0.112***	0.126***	0.151***	0.156***	-0.008
		(0.004)	(0.005)	(0.005)	(0.006)	(0.010)
URM teacher certified 6+ years ago	0.300	-0.070***	-0.054***	-0.064***	-0.044***	0.004
		(0.004)	(0.004)	(0.005)	(0.004)	(0.008)
White teacher certified 6+ years ago	0.464	-0.113***	-0.142***	-0.169***	-0.180***	0.017*
		(0.004)	(0.005)	(0.006)	(0.006)	(0.010)
First-year teacher	0.047	0.085***	0.093***	0.097***	0.129***	0.024**
		(0.004)	(0.005)	(0.004)	(0.006)	(0.010)
First-year White teacher	0.026	0.053***	0.059***	0.060***	0.092***	0.026***
		(0.003)	(0.004)	(0.003)	(0.005)	(0.008)
First-year URM teacher	0.019	0.028***	0.031***	0.034***	0.032***	-0.004
		(0.002)	(0.003)	(0.003)	(0.003)	(0.005)
Panel C. Teacher pay						
Total salary for all teachers	197,199	-815	-7,997***	-2,246***	-2,611***	6,818***
		(811)	(1,094)	(523)	(547)	(1,584)
Total salary for URM teachers	79,801	30	603	2,002***	226	-2,349*
		(751)	(703)	(479)	(420)	(1,234)
Total salary for White teachers	113,641	-1,374*	-9,094***	-3,072***	-4,466***	6,327***
		(790)	(986)	(510)	(591)	(1,531)
N (# school/grade/departure/years)	4,887	105,829	86,102	123,706	107,812	423,449

Notes: This table displays RD and RD-DiD estimates of the effects of the TExES reform on teacher composition as defined by their certification exams (Panel A), demographics (Panel B), and pay (Panel C). Column (A) shows the mean of each outcome in the year prior to the teacher departure ($\tau_{ty} = -1$) in school/grades that experienced a departure in the post-reform years (2011–2016). Columns (B)–(E) show RD coefficients β from equation (7) estimated separately for grades 3–4 and 7–8, and for departures in 2011–2016 (post-reform) and 2005–2010 (pre-reform). Column (F) shows the RD-DiD coefficient θ from equation (8). Standard errors in parentheses are clustered at the school level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: RD-DiD effects of harder certification exams on student achievement

	(A)	(B)	(C)	(D)	(E)	(F)
	Grade 3–4 departures			Grade 7–8 departures		
	Post-reform mean at τ_{ty} = -1	Post-reform RD	Pre-reform RD	Post-reform RD	Pre-reform RD	RD-DiD
Panel A. Student demographics (balance tests)						
Number of students with test scores	74.268	-0.107	-0.819**	-0.642	-2.799***	-1.445
		(0.253)	(0.347)	(0.675)	(0.668)	(1.063)
Male	0.498	-0.004***	0.001	-0.000	0.001	-0.004
		(0.001)	(0.002)	(0.001)	(0.001)	(0.002)
White	0.262	0.003***	0.002	0.004***	0.002*	-0.000
		(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Hispanic	0.537	-0.002**	0.001	-0.004***	-0.001	-0.000
		(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Black	0.148	-0.000	-0.001	0.000	-0.002**	-0.001
		(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Economically disadvantaged	0.667	-0.002	0.004**	-0.010***	-0.000	0.004
		(0.002)	(0.002)	(0.001)	(0.001)	(0.003)
In gifted education	0.096	-0.004***	-0.000	-0.000	0.002*	-0.001
		(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
At risk of dropping out	0.485	-0.004	-0.000	-0.000	-0.017***	-0.020***
		(0.003)	(0.003)	(0.003)	(0.002)	(0.006)
Demographic index (Math score)	-0.000	0.002	-0.006***	0.010***	0.001	-0.001
		(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
Demographic index (ELA score)	-0.005	0.006***	-0.006***	0.010***	0.002	0.004
		(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
Panel B. Student achievement						
Math score	-0.039	-0.004	-0.006	0.014**	-0.002	-0.014
		(0.004)	(0.005)	(0.005)	(0.005)	(0.010)
Math score residuals	-0.017	0.001	0.002	-0.001	0.000	-0.000
		(0.005)	(0.006)	(0.004)	(0.004)	(0.010)
ELA score	-0.035	0.000	-0.010**	0.009***	0.006	0.007
		(0.004)	(0.004)	(0.003)	(0.004)	(0.008)
ELA score residuals	-0.009	-0.004	-0.004	-0.004	0.006*	0.009
		(0.004)	(0.005)	(0.003)	(0.003)	(0.008)
N (# school/grade/departure/years)	4,887	105,829	86,102	123,706	107,812	423,449

Notes: This table displays RD and RD-DiD estimates of the effects of the TExES reform on student composition (Panel A) and student achievement (Panel B). Column (A) shows the mean of each outcome in the year prior to the teacher departure ($\tau_{ty} = -1$) in school/grades that experienced a departure in the post-reform years (2011–2016). Columns (B)–(E) show RD coefficients β from equation (7) estimated separately for grades 3–4 and 7–8, and for departures in 2011–2016 (post-reform) and 2005–2010 (pre-reform). Column (F) shows the RD-DiD coefficient θ from equation (8). Standard errors in parentheses are clustered at the school level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

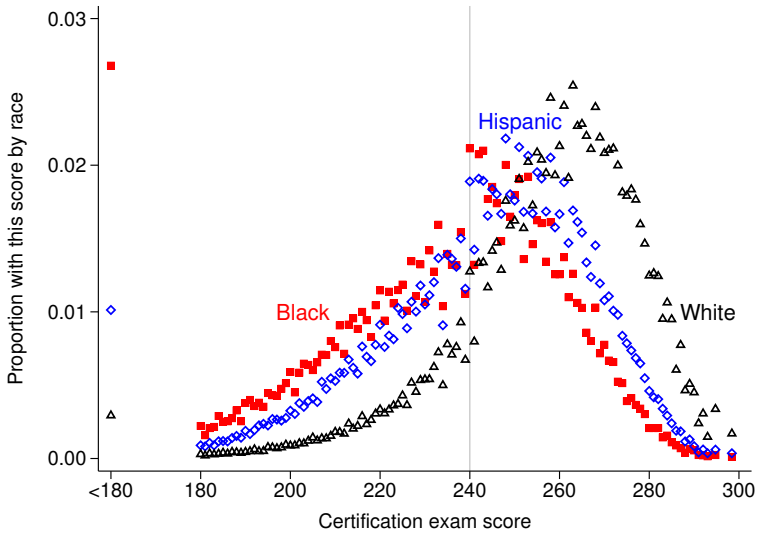
Table 7: RD-DiD effects of harder certification exams on student achievement — Grades 3–4 with various control groups

	(A)	(B)	(C)	(D)	(E)	(F)
	Departures in which teaching subject matches the exam subject			Departures in which subjects don't match		
	Post-reform mean at τ_{ty} = -1	Post-reform RD	Pre-reform RD	Post-reform RD	Pre-reform RD	RD-DiD
Panel A. Math vs. ELA scores when Math teacher departs (within school)						
Demographic index (Test score)	0.021	0.009***	-0.000	0.010**	0.002	0.001
		(0.004)	(0.005)	(0.004)	(0.005)	(0.002)
Test score	0.031	-0.012	0.022*	0.009	0.012	-0.030**
		(0.011)	(0.013)	(0.008)	(0.009)	(0.015)
Test score residuals	0.001	-0.021	0.049***	0.007	0.022*	-0.055***
		(0.013)	(0.015)	(0.009)	(0.011)	(0.018)
N (# *sty* observations)	740	14,776	9,578	14,775	9,577	48,706
Panel B. ELA vs. Math scores when ELA teacher departs (within school)						
Demographic index (Test score)	0.043	0.003	-0.002	0.004	-0.003	-0.002
		(0.003)	(0.004)	(0.003)	(0.004)	(0.002)
Test score	0.020	0.005	0.012*	0.004	0.011	-0.001
		(0.006)	(0.007)	(0.008)	(0.009)	(0.011)
Test score residuals	-0.015	0.015*	0.010	0.013	0.012	0.003
		(0.008)	(0.009)	(0.011)	(0.009)	(0.013)
N (# *sty* observations)	1,143	23,074	16,325	23,071	16,326	78,796
Panel C. Math teacher departs vs. ELA teacher departs (across school)						
Demographic index (Math score)	0.021	0.009***	-0.000	0.004	-0.003	0.002
		(0.004)	(0.005)	(0.003)	(0.004)	(0.008)
Math score	0.031	-0.012	0.022*	0.004	0.011	-0.027
		(0.011)	(0.013)	(0.008)	(0.009)	(0.022)
Math score residuals	0.001	-0.021	0.049***	0.013	0.012	-0.071***
		(0.013)	(0.015)	(0.011)	(0.009)	(0.025)
N (# *sty* observations)	740	14,776	9,578	23,071	16,326	63,751
Panel D. ELA teacher departs vs. Math teacher departs (across school)						
Demographic index (ELA score)	0.043	0.003	-0.002	0.010**	0.002	-0.003
		(0.003)	(0.004)	(0.004)	(0.005)	(0.008)
ELA score	0.020	0.005	0.012*	0.009	0.012	-0.004
		(0.006)	(0.007)	(0.008)	(0.009)	(0.017)
ELA score residuals	-0.015	0.015*	0.010	0.007	0.022*	0.020
		(0.008)	(0.009)	(0.009)	(0.011)	(0.020)
N (# *sty* observations)	1,143	23,074	16,325	14,775	9,577	63,751

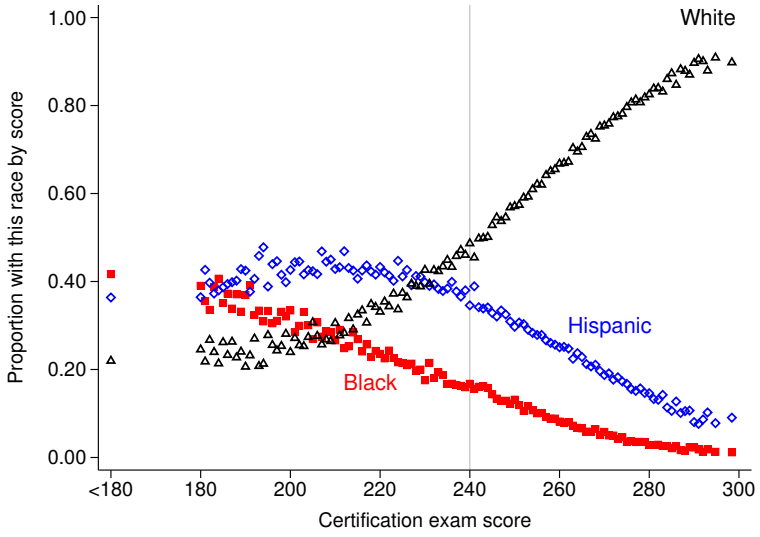
Notes: This table displays RD and RD-DiD estimates of the effects of the TExES reform on student composition and student achievement. The structure of this table is similar to Table 6, but in this table we restrict the sample to students in grades 3–4 who have a departing teacher that is identified in the data as specifically teaching math or ELA, and we use test scores or students in unaffected subjects as the control group. In Panels A–B, the regression sample includes school/grades with a departing math (Panel A) and ELA (Panel B) teacher, and the RD coefficients show how teacher departures affect exam scores in the same subject (columns B–C) relative to exam scores in the other subject (columns D–E). In Panels C–D, the outcome variables are math scores (Panel C) and ELA scores (Panel D), and the RD coefficients show how teacher departures affect test scores when the departing teacher teaches the same subject (columns B–C) relative to when the departing teacher teaches the other subject (columns D–E). In all panels, column (F) shows RD-DiD coefficient that equal column (B) − column (C) − (column D − column E). See Appendix B.3.3 for details on the sample and these RD-DiD specifications. Standard errors in parentheses are clustered at the school level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix — For Online Publication Only

A Appendix Figures and Tables



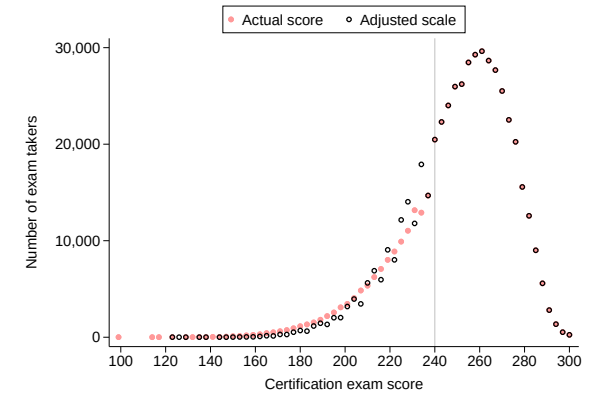
Panel A. Score distribution for each racial group



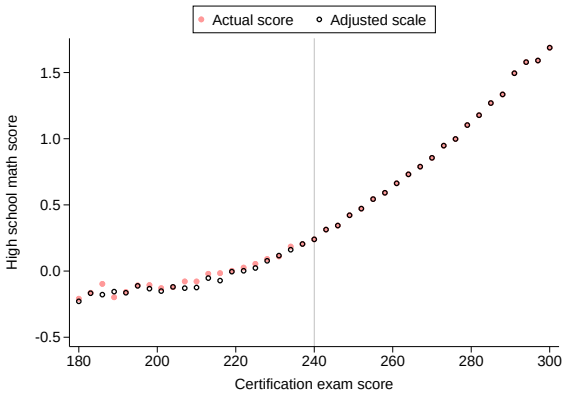
Panel B. Racial distribution for each score value

Figure A1: First-attempt certification exam scores by race

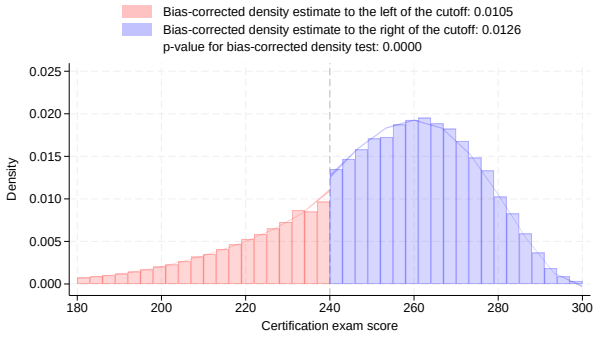
Notes: This figure shows the distribution of first-attempt certification exam scores by race. Panel A shows the density of exam scores for Black, Hispanic, and White exam takers. Panel B shows the proportion of exam takers who are Black, Hispanic, and White at each score value.



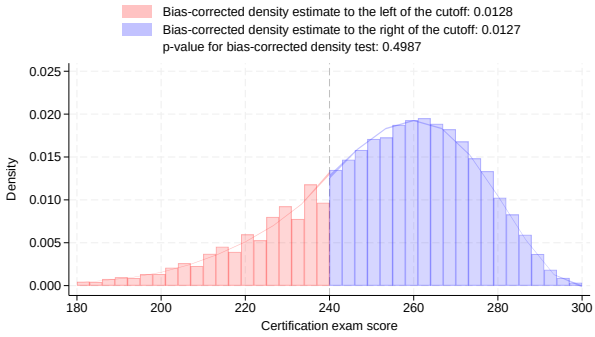
Panel A. Number of exam takers



Panel B. High school math score



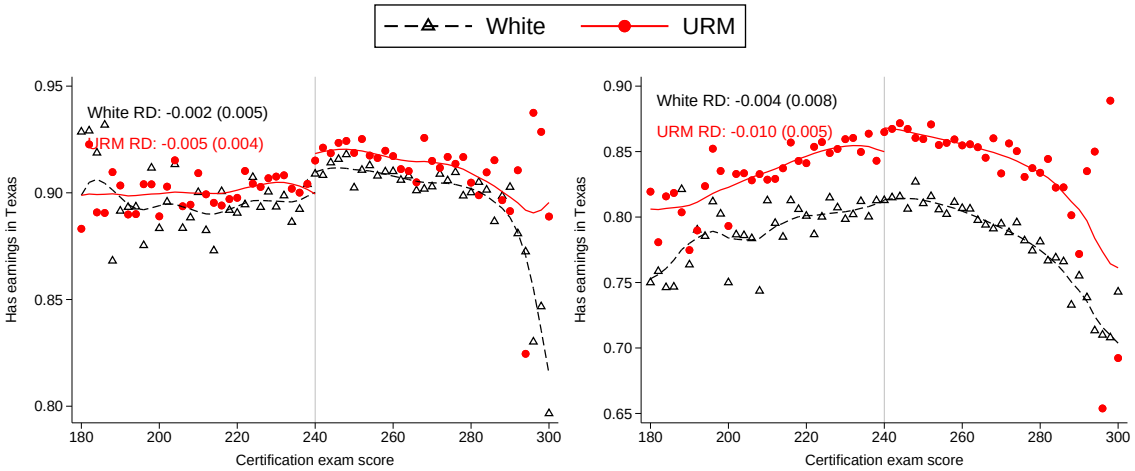
Panel C. rddensity test: Actual score



Panel D. rddensity test: Adjusted scale

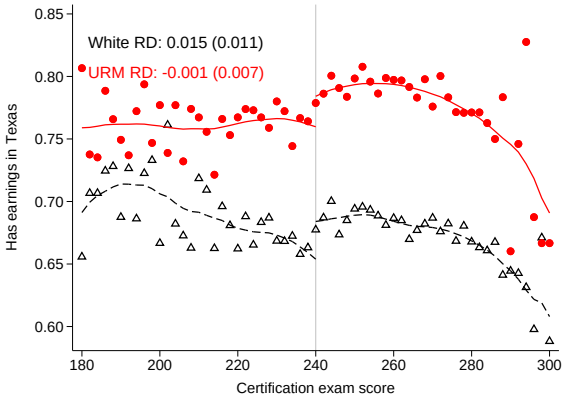
Figure A2: Density tests

Notes: The figure presents density tests of certification exam scores using both actual and adjusted scores. Panel A shows scatter plots of the number of exam takers within score bins, while Panel B illustrates the relationship between high school math scores and certification exam scores. Panels C and D display the regression discontinuity density tests using the `rddensity` package of Cattaneo et al. (2020), based on the actual and adjusted scores, respectively. The actual TExES certification exam scores are calculated using a scaled formula, see details in Section 3.2. We rescale failing scores so that they follow the same linear relationship with high school math scores as passing scores in the vicinity of the threshold, which constructs the adjusted scores.



Panel A. Has earnings 1 year later

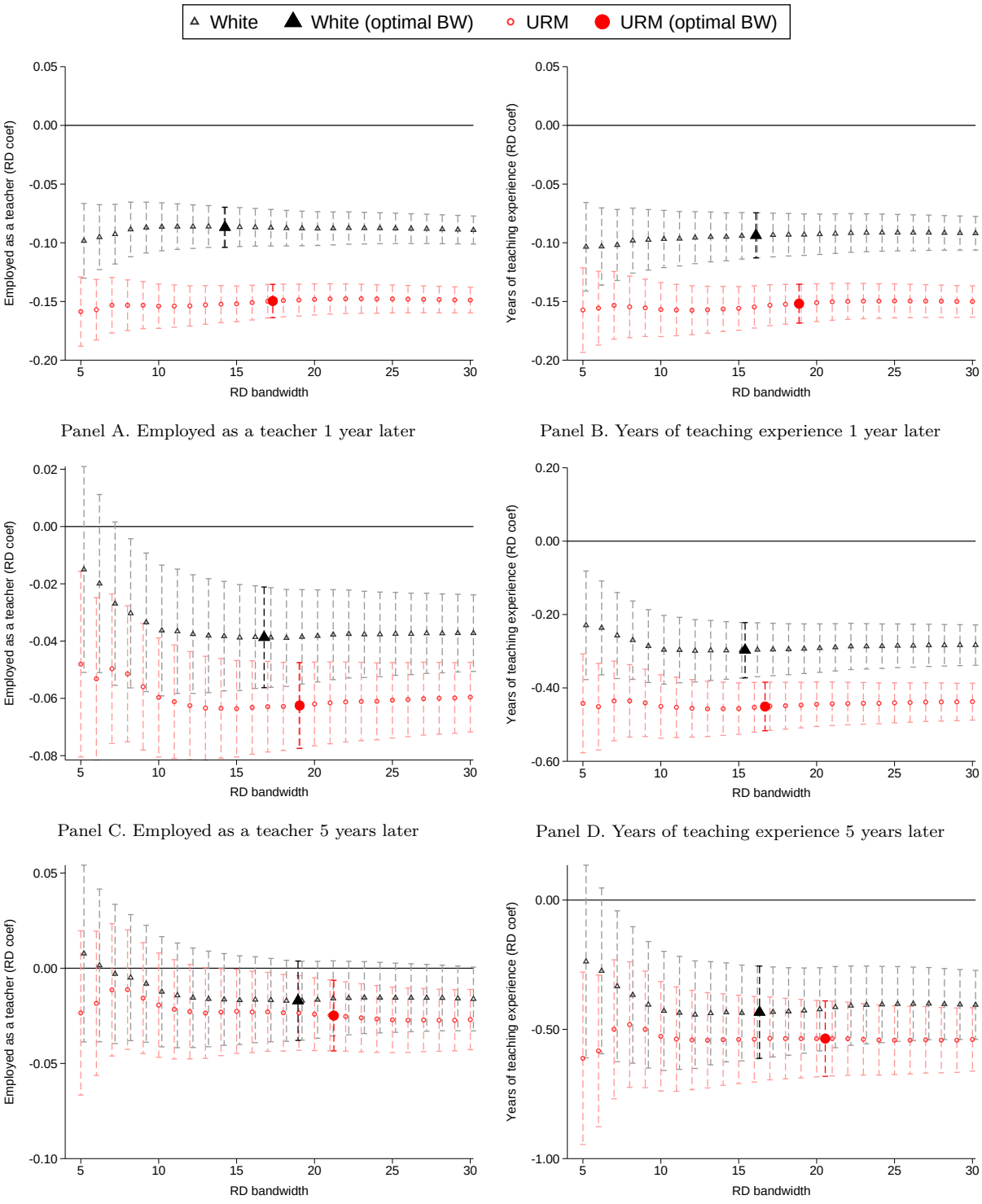
Panel B. Has earnings 5 years later



Panel C. Has earnings 10 years later

Figure A3: RD plots: Has earnings in Texas by race

Notes: The figure presents the RD effects of failing a certification exam on the likelihood of appearing in the TWC earnings data. Panels A-C display the RD effects on appearing the TWC data one, five, and ten years after the exam, respectively. The RD estimates are shown at the upper-left corner, with standard errors included in the parentheses.



Panel A. Employed as a teacher 1 year later

Panel B. Years of teaching experience 1 year later

Panel C. Employed as a teacher 5 years later

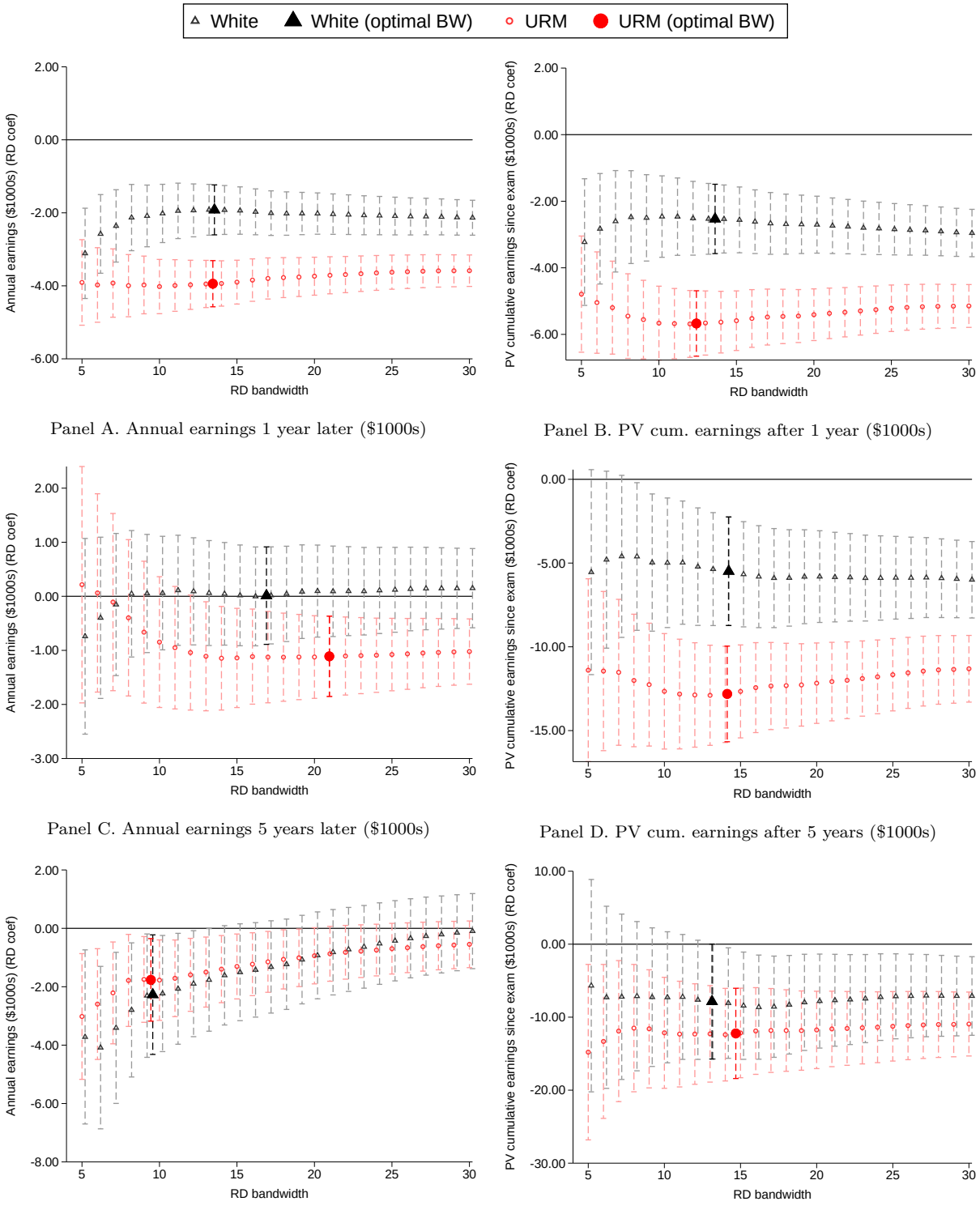
Panel D. Years of teaching experience 5 years later

Panel E. Employed as a teacher 10 years later

Panel F. Years of teaching experience 10 years later

Figure A4: Robustness to RD bandwidth: Teaching outcomes by race

Notes: The figure presents RD effects of failing a certification exam on teaching outcomes by race using different RD bandwidths (x-axis). Panels A-F display the RD effects on teacher employment and years of teaching experience one, five, and ten years after the exam, respectively. Markers depict RD coefficients. Dashed lines depict 95 percent confidence intervals using standard errors clustered at the individual level.



Panel A. Annual earnings 1 year later ($1000s)

Panel B. PV cum. earnings after 1 year ($1000s)

Panel C. Annual earnings 5 years later ($1000s)

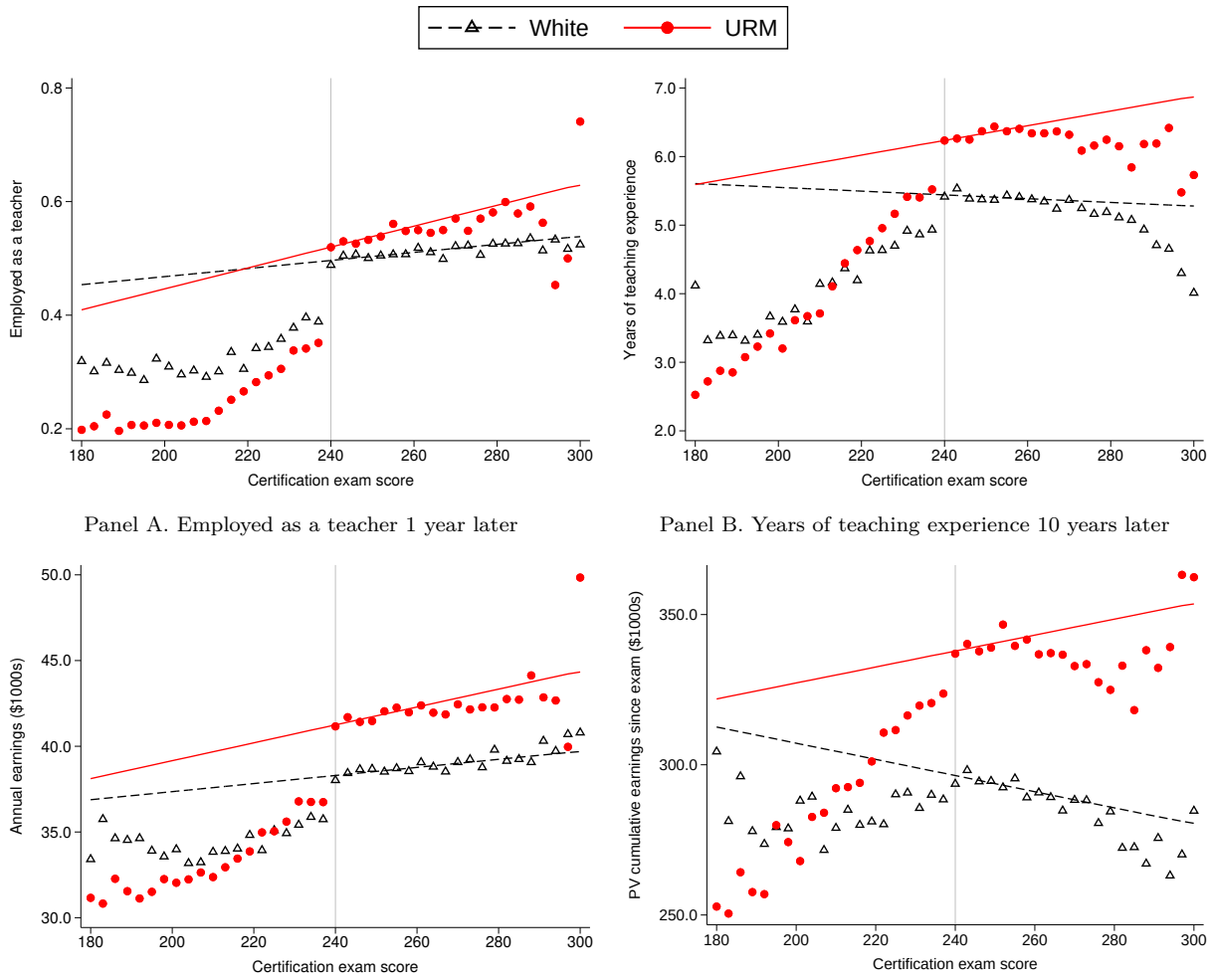
Panel D. PV cum. earnings after 5 years ($1000s)

Panel E. Annual earnings 10 year later ($1000s)

Panel F. PV cum. earnings after 10 years ($1000s)

Figure A5: Robustness to RD bandwidth: Earnings outcomes by race

Notes: The figure presents RD effects of failing a certification exam on earnings outcomes by race using different RD bandwidths (x-axis). Panels A-F display the RD effects on annual earnings and the present value of cumulative earnings one, five, and ten years after the exam, respectively. Markers depict RD coefficients. Dashed lines depict 95 percent confidence intervals using standard errors clustered at the individual level.



Panel A. Employed as a teacher 1 year later

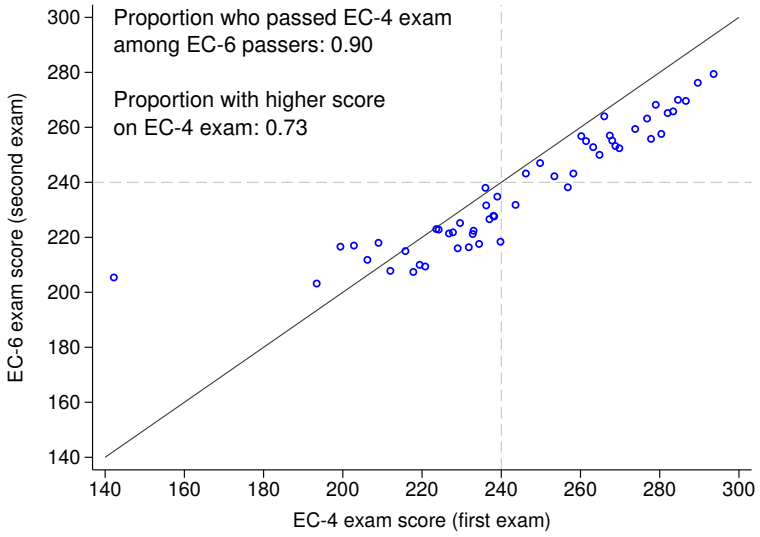
Panel B. Years of teaching experience 10 years later

Panel C. Annual earnings 1 year later ($1000s)

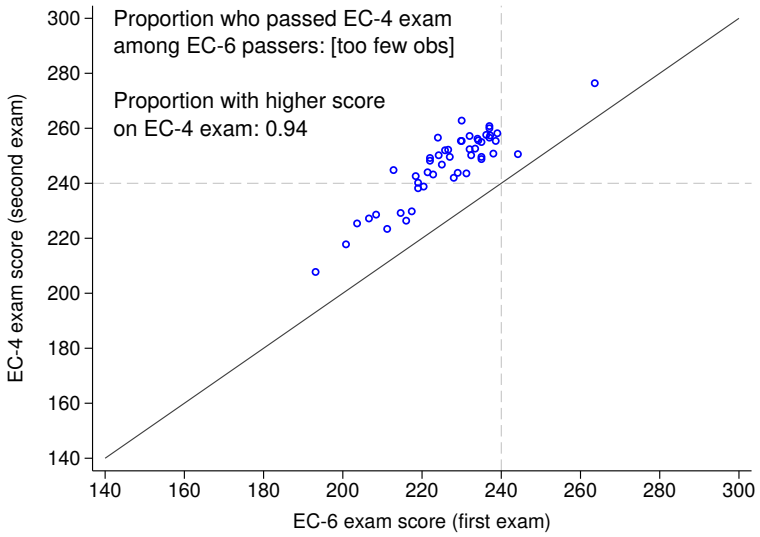
Panel D. PV cumulative earnings after 10 years ($1000s)

Figure A6: Extrapolated effects of failing certification exam on outcomes

Notes: The figure presents the extrapolated effects of failing a certification exam on baseline outcomes by white and URM groups. Panels A-D display the extrapolated effects on teacher employment (1 year later), years of teaching experience (10 years later), annual earnings (1 year later), and the present value of cumulative earnings (10 years later), respectively. While the scatter plots indicate the observed relationship between exam scores and outcomes, the projected linear lines are extrapolated using the relationship for exam takers with a score range of 240 and 260.



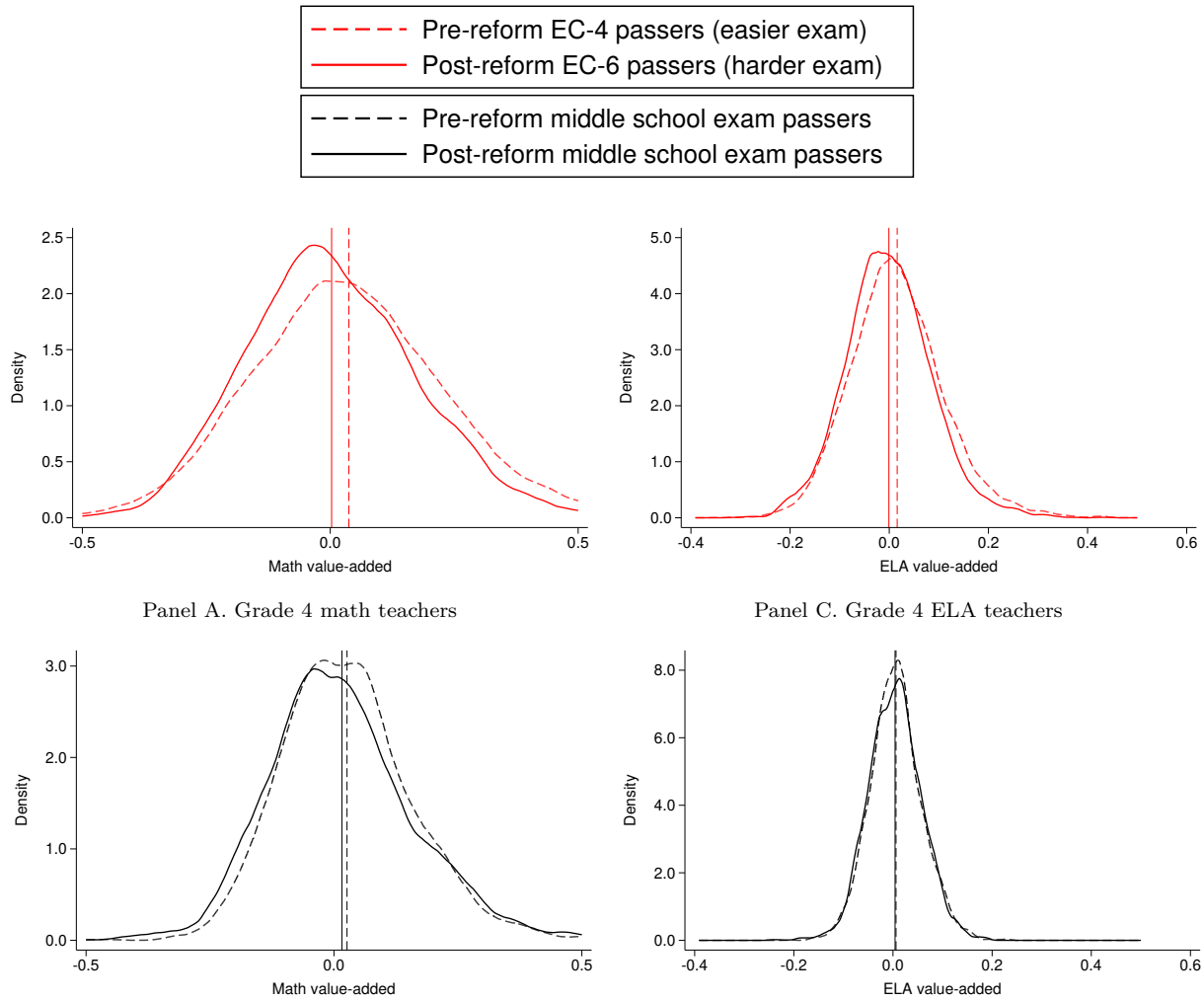
Panel A. Exam takers who took the EC-4 exam first



Panel B. Exam takers who took the EC-6 exam first

Figure A7: EC-4 and EC-6 exam performance for individuals who took both exams in the same year

Notes: This figure plots scores on the EC-4 and EC-6 exams for individuals who took both exams in 2010 (when both exams were offered). Panel A includes individuals who took the EC-4 exam first and the EC-6 exam second. Panel B includes individuals who took the EC-6 exam first and the EC-4 exam second. We show scores only first individuals' first-attempt at each exam. Due to FERPA requirements, dots represent average scores for groups of five exam takers. Dashed lines represent the passing score (240), and the solid line is a 45 degree line.



Panel A. Grade 4 math teachers

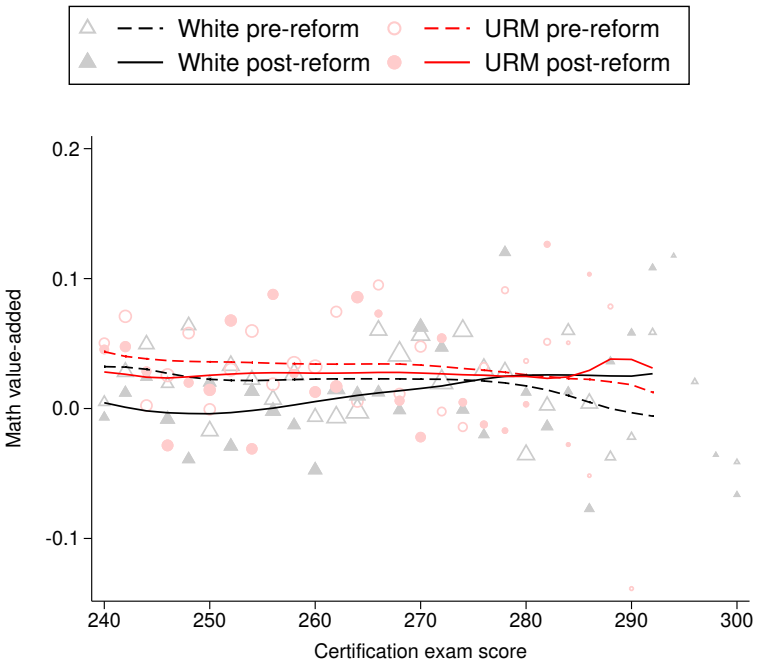
Panel C. Grade 4 ELA teachers

Panel B. Grade 7–8 math teachers

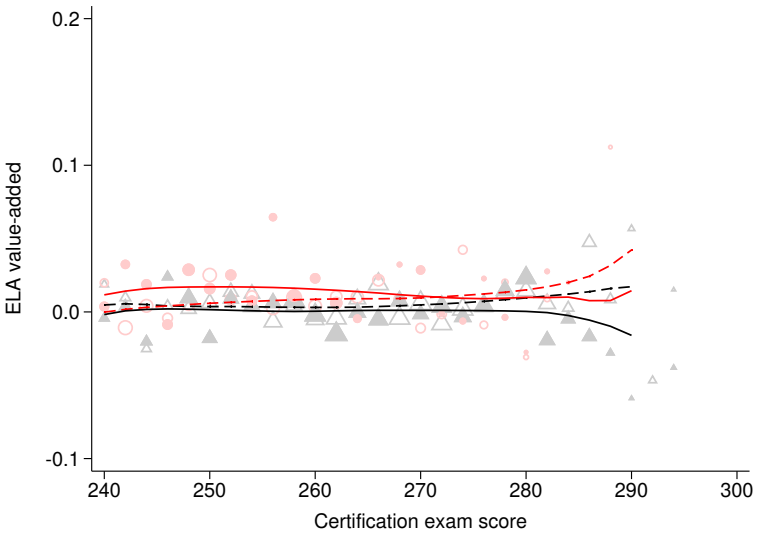
Panel D. Grade 7–8 ELA teachers

Figure A8: Value-added distributions for grade 4 and grade 7–8 teachers measured 3–8 years after the certification exam

Notes: This figure shows kernel densities of math (Panels A–B) and ELA (Panels C–D) value-added for individuals who passed different TExES certification exams. In Panels A and C, the sample includes individuals who took the EC-4 or EC-6 exams (Panel A of Table 3) as their first certification exam, who passed the exam, and who subsequently became grade 4 teachers. These graphs plot distributions of grade 4 value-added for individuals who took the EC-4 exam in 2003–2009 (dashed line) or the EC-6 exam in 2011–2015 (solid line). In Panels B and D, the sample includes individuals who took middle school Generalist or math/ELA exams (Panel B of Table 3) as their first certification exam, who passed the exam, and who subsequently became grade 7–8 teachers. The graph plots distributions of grade 7–8 value-added for individuals who took the middle school exams in 2003–2009 (dashed line) or in 2011–2015 (solid line). We measure value-added 3–8 years after taking the certification exam ("potential experience"), and weight observations so that the distribution of potential experience is the same for the two exams in each graph (using the exam with the minimum number of observations for each level of potential experience). Vertical lines denote the mean of each distribution.



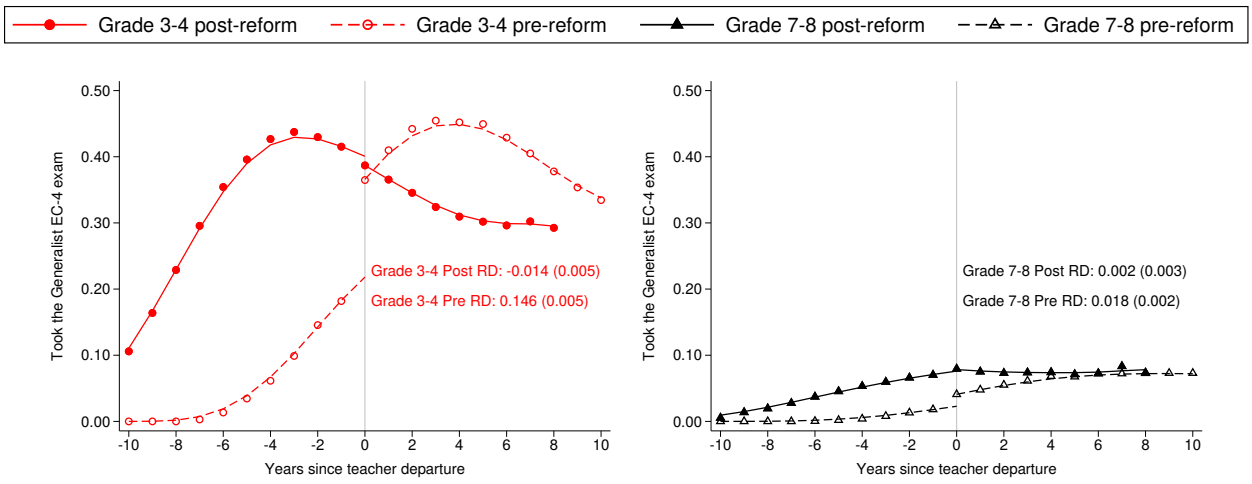
Panel A. Math value-added



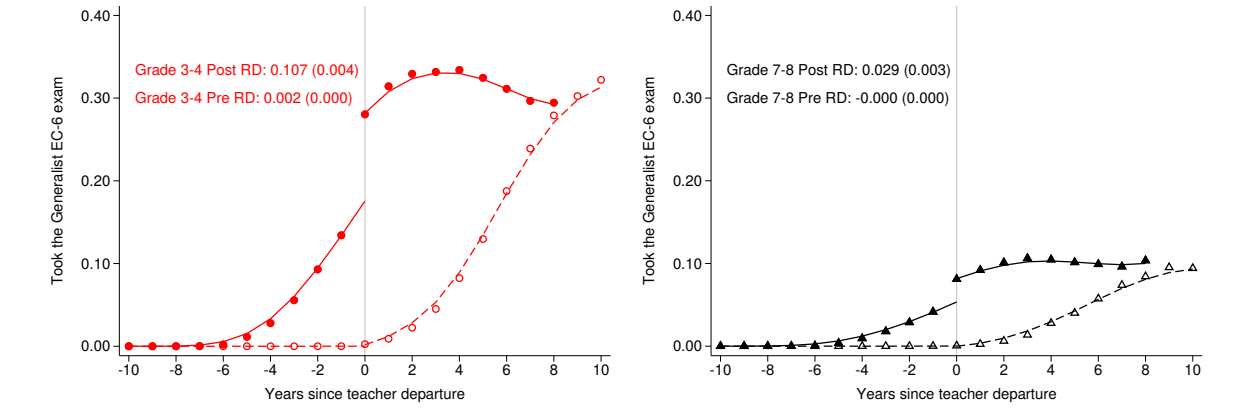
Panel B. ELA value-added

Figure A9: Value-added of grade 7–8 teachers by certification exam score

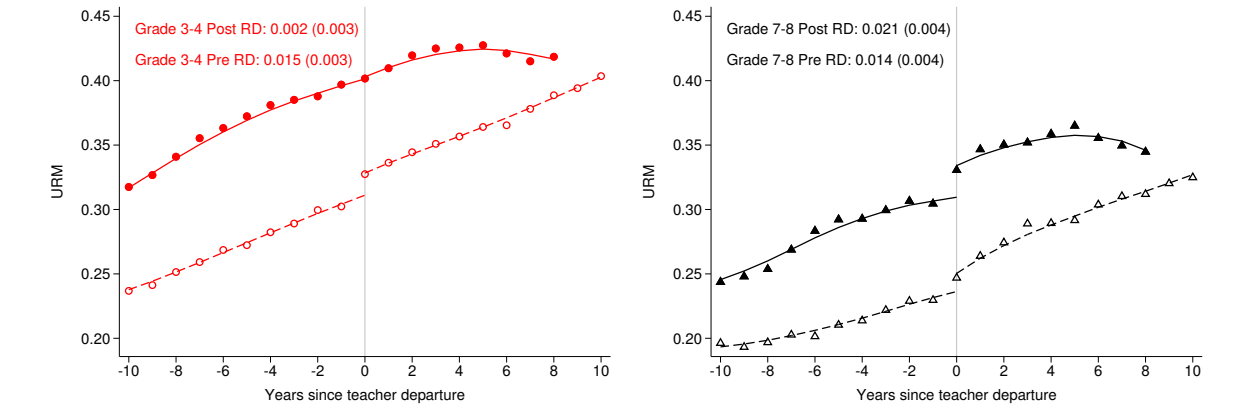
Notes: This figure shows the relationship between teacher value-added (y-axis) and certification exam scores (x-axis) for grade 7–8 math (Panel A) and ELA (Panel B) teachers. The sample includes individuals who took the middle school certification exams (Panel B of Table 3) as their first certification exam, who passed the exam, and who subsequently became grade 7 or 8 teachers. Markers depict average value-added within bins of certification exam scores, computed separately by race (white and URM) and for individuals who took the certification exams in 2003–2009 or in 2011–2015, as indicated by the legend. We measure value-added 3–8 years after taking the certification exam ("potential experience"), and weight observations so that the distribution of potential experience is the same for the two exams in each graph (using the exam with the minimum number of observations for each level of potential experience). Lines depict predicted values from local linear regressions.



Panel A. Took the Generalist EC-4 exam



Panel B. Took the Generalist EC-6 exam



Panel C. URM

Figure A10: Effects of teacher departures on teacher composition

Notes: This figure contains RD graphs that show how teacher departures affect the composition of teachers in a school/grade. Red circles represent grade 3–4 teachers. Black triangles represent grade 7–8 teachers. Hollow symbols depict teacher departures that occurred in $y \in$ 2005–2010 (pre-reform). Solid symbols depict teacher departures that occurred in $y \in$ 2011–2016 (post-reform). The x-axis is years relative to the teacher departure, τ_{ty}. The y-axis depicts the average outcome at the school/-grade/year level. Each graph displays RD coefficients β from equation (7) with standard errors clustered at the school level in parentheses.

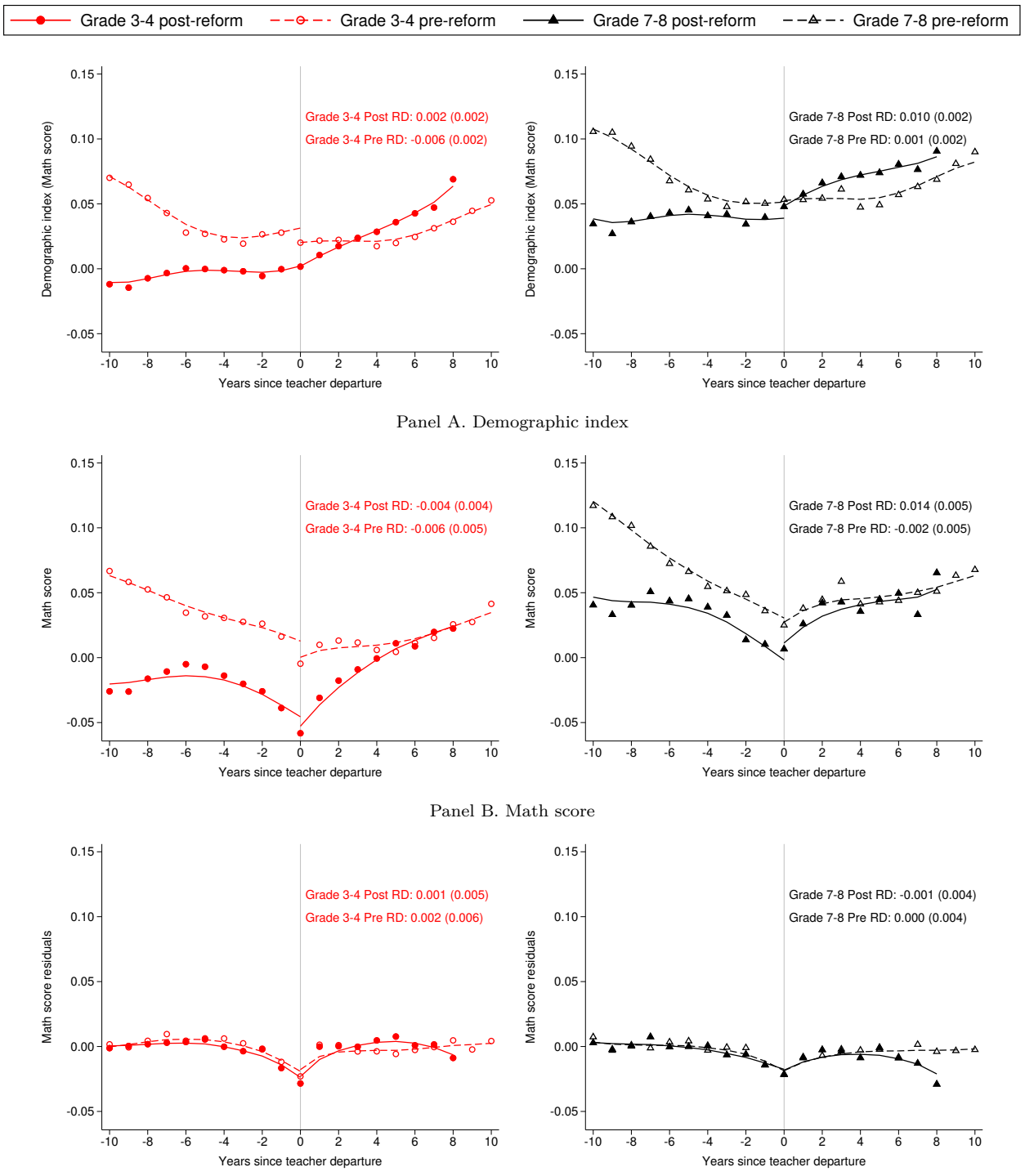


Figure A11: Effects of teacher departures on student math achievement

Notes: This figure contains RD graphs that show how teacher departures affect the math achievement of students in a school/-grade. Red circles represent grade 3–4 teachers. Black triangles represent grade 7–8 teachers. Hollow symbols depict teacher departures that occurred in $y \in$ 2005–2010 (pre-reform). Solid symbols depict teacher departures that occurred in $y \in$ 2011–2016 (post-reform). The x-axis is years relative to the teacher departure, τ_{ty}. The y-axis depicts the average outcome at the school/grade/year level. Each graph displays RD coefficients β from equation (7) with standard errors clustered at the school level in parentheses.

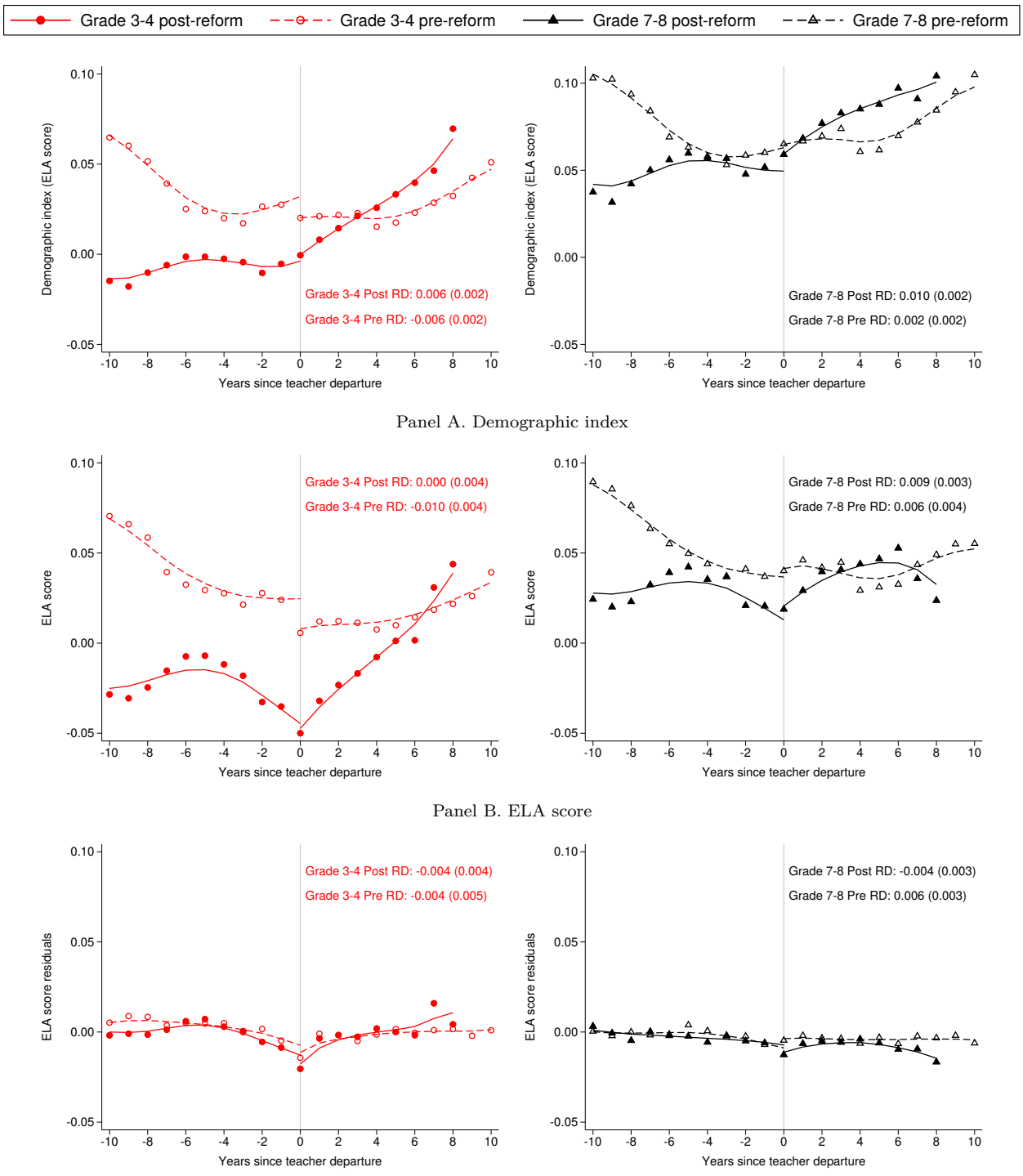
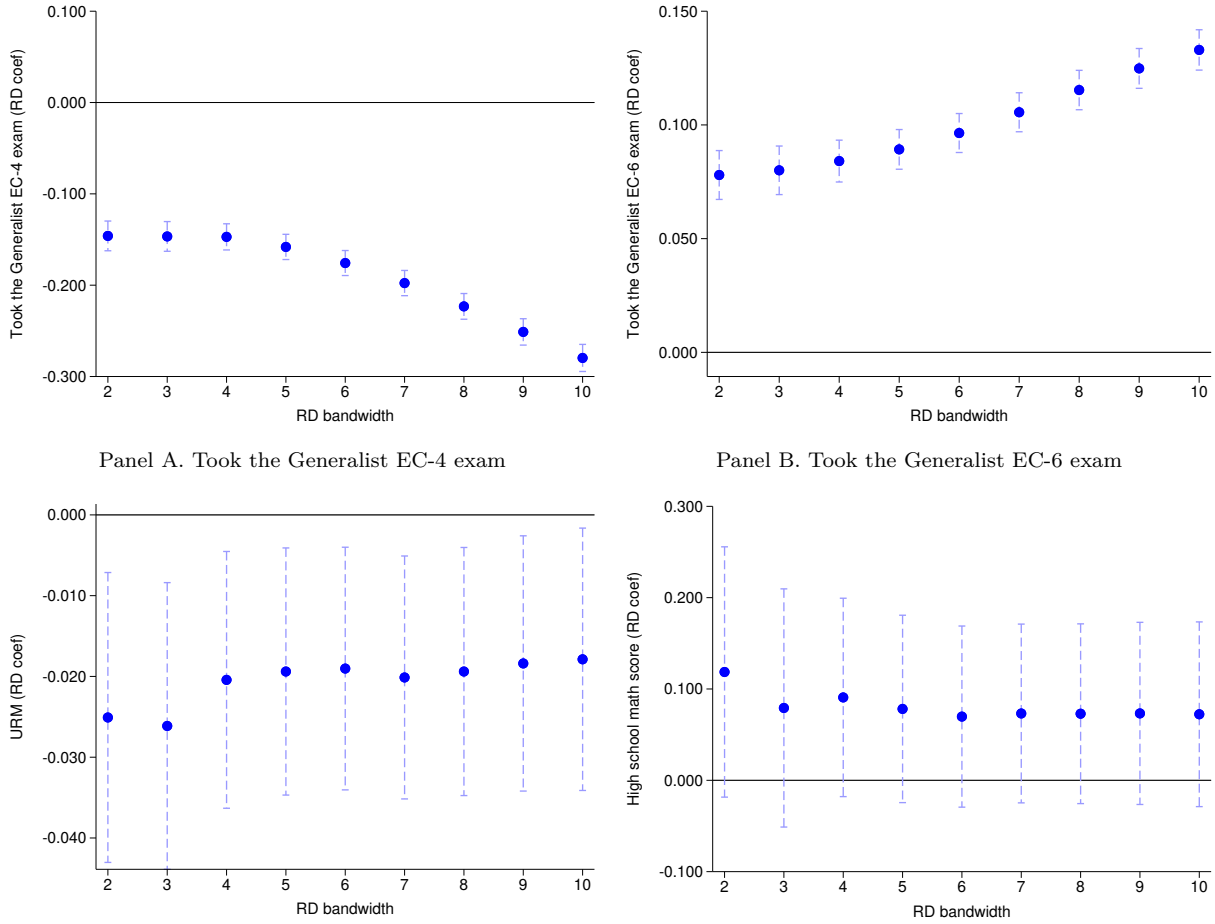


Figure A12: Effects of teacher departures on student ELA achievement

Notes: This figure contains RD graphs that show how teacher departures affect the ELA achievement of students in a school/-grade. Red circles represent grade 3–4 teachers. Black triangles represent grade 7–8 teachers. Hollow symbols depict teacher departures that occurred in $y \in$ 2005–2010 (pre-reform). Solid symbols depict teacher departures that occurred in $y \in$ 2011–2016 (post-reform). The x-axis is years relative to the teacher departure, τ_{ty}. The y-axis depicts the average outcome at the school/grade/year level. Each graph displays RD coefficients β from equation (7) with standard errors clustered at the school level in parentheses.



Panel A. Took the Generalist EC-4 exam

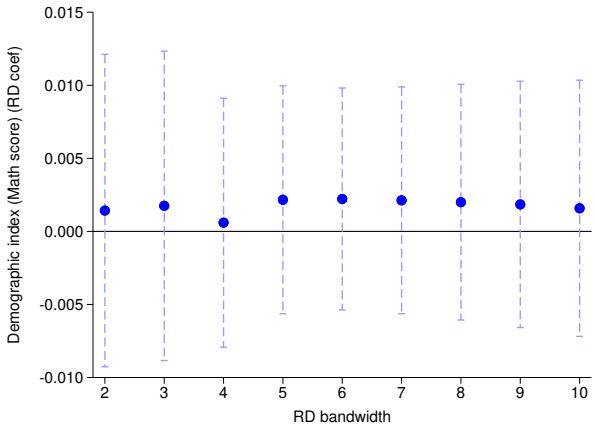
Panel B. Took the Generalist EC-6 exam

Panel C. URM

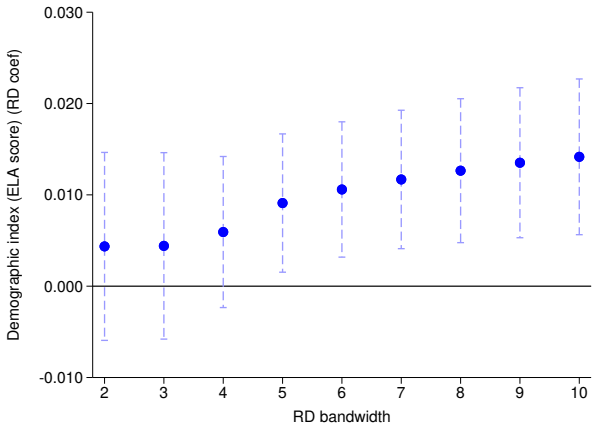
Panel D. HS math score

Figure A13: RD-DiD effects of harder certification exams on teacher composition — Robustness to RD bandwidth

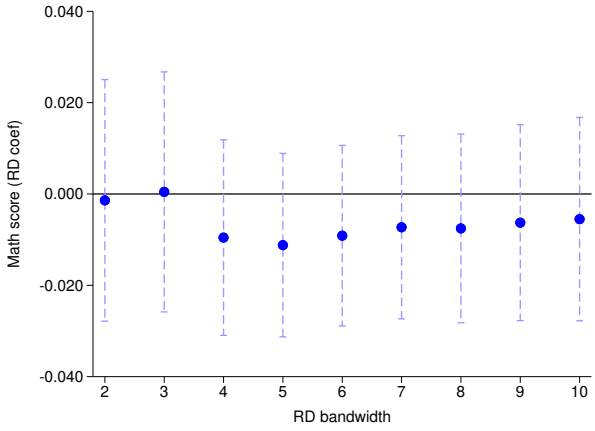
Notes: This figure displays the robustness of our RD-DiD estimates for teacher composition outcomes to the choice of RD bandwidth. The y-axis in each graph displays the θ from equation (8). The x-axis displays the bandwidth used for the RD regression (7), which we vary from $h^Y \in 2$–10 years relative to the year of the teacher departure. Dashed lines represent 95 percent confidence intervals using standard errors clustered at the school level.



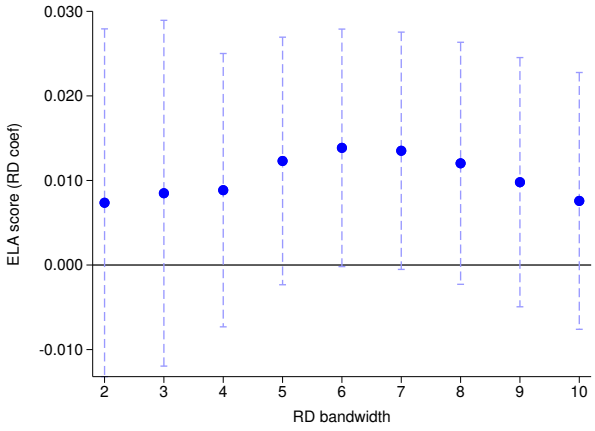
Panel A. Demographic index (Math score)



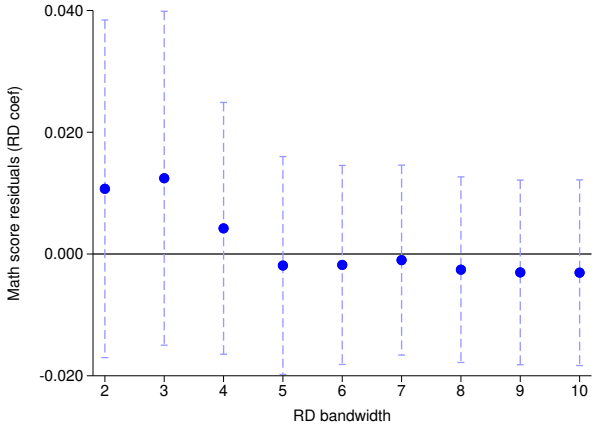
Panel B. Demographic index (ELA score)



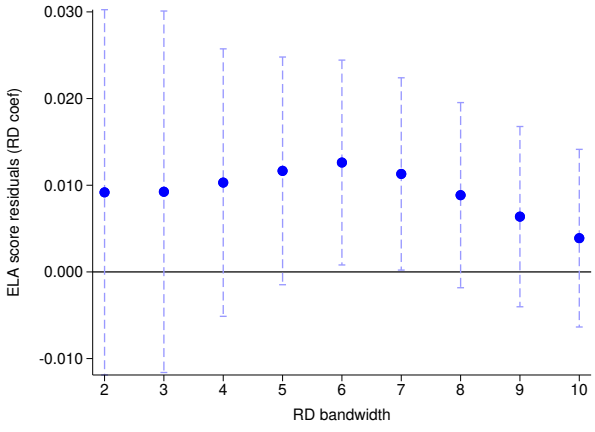
Panel C. Math score



Panel D. ELA score



Panel E. Math score residuals



Panel F. ELA score residuals

Figure A14: RD-DiD effects of harder certification exams on student achievement — Robustness to RD bandwidth

Notes: This figure displays the robustness of our RD-DiD estimates for student achievement outcomes to the choice of RD bandwidth. The y-axis in each graph displays the θ from equation (8). The x-axis displays the bandwidth used for the RD regression (7), which we vary from $h^Y \in 2$–10 years relative to the year of the teacher departure. Dashed lines represent 95 percent confidence intervals using standard errors clustered at the school level.

Table A1: Sources of the URM/White annual earnings gap

	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
	Dependent variable							
Covariate	TWC earnings excl. zeroes	TWC earnings	TEA earnings	TEA earnings from teaching	TEA earnings from teaching	TEA earnings from teaching	TEA base teaching salary	TEA base teaching salary
Panel A. 1 year after certification exam								
URM	3,003***	3,442***	2,790***	2,570***	1,144***	1,121***	1,089***	79***
	(76)	(79)	(97)	(99)	(28)	(28)	(26)	(24)
Employed as a teacher in this year					46,791***	46,633***	45,511***	25,657***
					(26)	(27)	(26)	(568)
Years of teaching experience						369***	359***	362***
						(17)	(16)	(14)
N (# individuals/exams)	259,641	285,352	285,352	285,352	285,352	285,352	285,352	285,314
Panel B. 5 years after certification exam								
URM	3,487***	5,978***	5,414***	5,379***	1,251***	1,194***	1,054***	100***
	(89)	(103)	(109)	(110)	(34)	(34)	(32)	(31)
Employed as a teacher in this year					49,259***	46,992***	45,937***	13,810***
					(30)	(68)	(64)	(721)
Years of teaching experience						612***	550***	394***
						(16)	(15)	(11)
N (# individuals/exams)	205,302	251,653	251,653	251,653	251,653	251,653	251,653	251,616
Panel C. 10 years after certification exam								
URM	2,815***	7,604***	6,291***	6,259***	1,047***	955***	825***	66
	(154)	(158)	(141)	(141)	(50)	(50)	(47)	(46)
Employed as a teacher in this year					49,853***	43,847***	43,161***	6,895***
					(45)	(167)	(160)	(771)
Years of teaching experience						739***	666***	357***
						(19)	(18)	(14)
N (# individuals/exams)	129,811	182,031	182,031	182,031	182,031	182,031	182,031	181,973
Missing values set to zero		Y	Y	Y	Y	Y	Y	Y
School district fixed effects								Y

Notes: This table reports estimates from regressions of annual earnings on an indicator for URM individuals. The sample includes TExES certification exam takers who passed on their first attempt. Column (A) excludes individuals with missing earnings, while these individuals are included in columns (B)–(H) with earnings set to zero. The dependent variables in each column are: (A)–(B) annual earnings in the TWC data; (C) annual earnings in the TEA data; (D)–(F) annual earnings associated with teaching jobs in the TEA data; (G)–(H) base salary (i.e., excluding other support) associated with teaching jobs in the TEA data. Column (E) adds a covariate that measures the number of full-time equivalent years the individual worked as a teacher in that year. Column (F) adds a covariate that measures the number of years of teaching experience in the TEA data. Column (H) adds school district fixed effects. All regressions include fixed effects for certification exam score values (240–300) and calendar years. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: RD balance tests

	(A)	(B)	(C)	(D)	(E)	(F)
	Mean above threshold		RD coefficients			*P*-value: W=URM
	White	URM	All	White	URM	
Male	0.235	0.264	0.004	0.004	0.004	0.993
			(0.005)	(0.006)	(0.007)	
White	1.000	0.000	0.009			
			(0.006)			
Black	0.000	0.314	-0.006		-0.010	
			(0.004)		(0.007)	
Hispanic	0.000	0.686	0.001		0.010	
			(0.005)		(0.007)	
Mean proportion white at school during K-12	0.741	0.319	0.008*	0.003	-0.001	0.485
			(0.005)	(0.003)	(0.006)	
High school graduation cohort	2004.491	2004.088	0.115**	0.035	0.132*	0.379
			(0.058)	(0.078)	(0.075)	
Ever identified as an immigrant during K-12	0.007	0.038	-0.003	-0.002	-0.003	0.744
			(0.002)	(0.002)	(0.004)	
Economically disadvantaged during K-12	0.077	0.533	0.003	0.007	0.008	0.904
			(0.005)	(0.005)	(0.009)	
Identified as at risk of dropping out during K-12	0.167	0.357	-0.004	0.002	-0.006	0.352
			(0.005)	(0.005)	(0.007)	
Gifted education during K-12	0.117	0.134	-0.003	-0.001	-0.005	0.643
			(0.004)	(0.005)	(0.006)	
In special education during K-12	0.039	0.018	0.002	-0.002	0.004*	0.149
			(0.002)	(0.003)	(0.002)	
Ever in bilingual education during K-12	0.001	0.146	0.004	-0.001	0.010	0.144
			(0.004)	(0.001)	(0.007)	
Ever in English as a Second Language during K-12	0.003	0.100	-0.003	0.000	-0.005	0.406
			(0.003)	(0.001)	(0.006)	
Ever had limited English proficiency during K-12	0.004	0.234	0.001	-0.000	0.006	0.501
			(0.005)	(0.001)	(0.009)	
High school math score	0.356	0.186	0.015	0.016	0.013	0.890
			(0.011)	(0.015)	(0.015)	
High school ELA score	0.477	0.273	-0.002	-0.003	-0.007	0.812
			(0.010)	(0.015)	(0.014)	
Had earnings in Texas in prior year	0.797	0.836	-0.001	0.009	-0.006	0.141
			(0.004)	(0.007)	(0.007)	
Prior year annual earnings ($1000s)	18.082	19.931	-0.322	0.009	-0.251	0.554
			(0.233)	(0.318)	(0.294)	
N (Male)	16,681	16,894	400,726	227,691	156,455	384,146
P-value: All coefs zero			0.242	0.737	0.524	

Notes: The table presents the RD estimates of the impacts of failing a certification exam on students' pre-determined characteristics. The dependent variables include students' gender, race, participation in K-12 initiatives and support systems, high school achievement, and pre-exam earnings in Texas. Columns (A) and (B) display the dependent variable means of candidates with exam scores ten points above the threshold. Columns (C)-(E) show the RD coefficients using equation (1). Column (F) indicates the p-value testing the statistical differences of RD effects between white and URM groups in Columns (D) and (E), respectively. "N (Male)" indicates the sample size for the outcome "Male", serving as a proxy sample size for other outcomes. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: RD effects of failing a certification exam on exam outcomes

	(A)	(B)	(C)	(D)	(E)	(F)
	Mean above threshold		RD coefficients			*P*-value: W=URM
	White	URM	All	White	URM	
Panel A. 1 year after certification exam						
Passed this certification exam by now	1.000	1.000	-0.294***	-0.258***	-0.295***	0.000
			(0.005)	(0.008)	(0.007)	
Passed any certification exam by now	1.000	1.000	-0.191***	-0.152***	-0.198***	0.000
			(0.005)	(0.007)	(0.006)	
# days to pass this exam (if passed)	0.000	0.000	119.648***	111.715***	128.507***	0.000
			(1.304)	(1.967)	(1.948)	
# days to pass any exam (if passed)	0.000	0.000	113.215***	103.849***	124.719***	0.000
			(1.203)	(1.684)	(1.869)	
N (Passed this certification exam by now)	16,683	16,896	506,345	227,740	156,499	384,239
Panel B. 5 years after certification exam						
Passed this certification exam by now	1.000	1.000	-0.260***	-0.232***	-0.251***	0.093
			(0.005)	(0.009)	(0.007)	
Passed any certification exam by now	1.000	1.000	-0.142***	-0.124***	-0.131***	0.421
			(0.004)	(0.007)	(0.006)	
# days to pass this exam (if passed)	0.000	0.000	146.828***	133.259***	161.650***	0.000
			(2.745)	(3.726)	(4.524)	
# days to pass any exam (if passed)	0.000	0.000	152.514***	127.439***	179.677***	0.000
			(2.998)	(3.696)	(5.164)	
N (Passed this certification exam by now)	14,528	14,425	442,510	201,864	133,272	335,136
Panel B. 10 years after certification exam						
Passed this certification exam by now	1.000	1.000	-0.241***	-0.214***	-0.221***	0.582
			(0.006)	(0.010)	(0.008)	
Passed any certification exam by now	1.000	1.000	-0.143***	-0.123***	-0.122***	0.930
			(0.006)	(0.009)	(0.007)	
# days to pass this exam (if passed)	0.000	0.000	160.183***	145.728***	170.863***	0.006
			(4.377)	(6.099)	(6.655)	
# days to pass any exam (if passed)	0.000	0.000	170.482***	150.324***	191.850***	0.000
			(5.461)	(7.231)	(8.244)	
N (Passed this certification exam by now)	9,848	9,734	310,404	145,790	88,718	234,508

Notes: The table presents the RD estimates of the impacts of failing a certification exam on exam outcomes. Panels A-C indicate the RD effects in one, five, and ten years after the exam, respectively. The dependent variables include four outcomes: an indicator of whether passed the exam by now, an indicator of whether passed any exam by now, the number of days to pass the exam (if passed), and the number of days to pass any exam (if passed). Columns (A) and (B) display the dependent variable means of candidates with exam scores ten points above the threshold. Columns (C)-(E) show the RD coefficients using equation (1). Column (F) indicates the *p*-value testing the statistical differences of RD effects between white and URM groups in Columns (D) and (E), respectively. "N (Passed this certification exam by now)" indicates the sample size for the outcome "Passed this certification exam by now", serving as a proxy sample size for other outcomes. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: RD effects of failing a certification exam on earnings outcomes

	(A)	(B)	(C)	(D)	(E)	(F)
	Mean above threshold		RD coefficients			*P*-value: W=URM
	White	URM	All	White	URM	
Panel A. 1 year after certification exam						
Has earnings in Texas	0.911	0.918	-0.010***	-0.002	-0.005	0.581
			(0.003)	(0.005)	(0.004)	
Annual earnings ($1000s)	38.239	41.432	-2.831***	-1.919***	-3.942***	0.000
			(0.213)	(0.349)	(0.323)	
Annual earnings, including zeroes ($1000s)	34.817	38.053	-2.884***	-1.818***	-3.669***	0.000
			(0.221)	(0.376)	(0.269)	
Earnings from teaching ($1000s)	45.204	46.444	-1.904***	-1.278***	-2.730***	0.001
			(0.174)	(0.287)	(0.292)	
Earnings from teaching, including zeroes ($1000s)	23.153	25.685	-5.229***	-3.640***	-6.589***	0.000
			(0.251)	(0.400)	(0.341)	
N (Annual earnings)	15,190	15,518	450,861	205,999	142,577	348,576
Panel B. 5 years after certification exam						
Has earnings in Texas	0.815	0.868	-0.010**	-0.004	-0.010**	0.508
			(0.004)	(0.008)	(0.005)	
Annual earnings ($1000s)	48.275	51.687	-0.552*	0.012	-1.110***	0.058
			(0.290)	(0.460)	(0.379)	
Annual earnings, including zeroes ($1000s)	39.323	44.864	-0.898***	-0.248	-1.441***	0.071
			(0.322)	(0.516)	(0.414)	
Earnings from teaching ($1000s)	50.416	52.253	-0.636***	-0.569***	-0.815***	0.410
			(0.142)	(0.201)	(0.213)	
Earnings from teaching, including zeroes ($1000s)	29.750	34.550	-2.452***	-2.143***	-3.149***	0.114
			(0.291)	(0.458)	(0.429)	
N (Annual earnings)	11,834	12,521	342,975	160,902	113,374	274,276
Panel C. 10 years after certification exam						
Has earnings in Texas	0.689	0.789	0.001	0.015	-0.001	0.207
			(0.006)	(0.011)	(0.007)	
Annual earnings ($1000s)	54.772	56.658	-1.595***	-2.271**	-1.765**	0.697
			(0.559)	(1.045)	(0.721)	
Annual earnings, including zeroes ($1000s)	37.736	44.679	-0.841**	-0.557	-1.338**	0.410
			(0.416)	(0.653)	(0.677)	
Earnings from teaching ($1000s)	53.565	55.158	-0.625***	-0.589*	-0.549*	0.927
			(0.207)	(0.308)	(0.294)	
Earnings from teaching, including zeroes ($1000s)	22.159	28.270	-1.144***	-1.012*	-1.439***	0.595
			(0.334)	(0.570)	(0.556)	
N (Annual earnings)	6,785	7,676	208,615	99,031	69,173	168,204

Notes: The table presents the RD estimates of the impacts of failing a certification exam on earning outcomes. Panels A-C indicate the RD effects in one, five, and ten years after the exam, respectively. The dependent variables include four outcomes: annual earnings, annual earnings (including zeros), earnings from teaching, and earnings from teaching (including zeros). Columns (A) and (B) display the dependent variable means of candidates with exam scores ten points above the threshold. Columns (C)-(E) show the RD coefficients using equation (1). Column (F) indicates the p-value testing the statistical differences of RD effects between white and URM groups in Columns (D) and (E), respectively. "N (Annual earnings)" indicates the sample size for the outcome "Annual earnings ($1000s)", serving as a proxy sample size for other outcomes. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: RD effects of failing a certification exam on outcomes: Running variable with adjusted scale

	(A)	(B)	(C)	(D)	(E)	(F)
	Mean above threshold		RD coefficients			*P*-value: W=URM
	White	URM	All	White	URM	
Panel A. 1 year after certification exam						
Employed as a teacher	0.497	0.525	-0.116***	-0.084***	-0.149***	0.000
			(0.005)	(0.008)	(0.007)	
Years of teaching experience	0.565	0.597	-0.120***	-0.092***	-0.153***	0.000
			(0.006)	(0.009)	(0.009)	
Annual earnings ($1000s)	38.239	41.432	-2.776***	-1.850***	-3.808***	0.000
			(0.206)	(0.336)	(0.290)	
PV cumulative earnings since exam ($1000s)	49.084	54.467	-4.007***	-2.489***	-5.587***	0.000
			(0.304)	(0.507)	(0.493)	
N (Employed as a teacher)	16,683	16,896	506,345	227,740	156,499	384,239
Panel B. 5 years after certification exam						
Employed as a teacher	0.608	0.677	-0.048***	-0.039***	-0.063***	0.053
			(0.006)	(0.009)	(0.008)	
Years of teaching experience	3.076	3.342	-0.359***	-0.291***	-0.453***	0.002
			(0.026)	(0.038)	(0.035)	
Annual earnings ($1000s)	48.275	51.687	-0.564*	0.037	-1.062***	0.057
			(0.292)	(0.450)	(0.369)	
PV cumulative earnings since exam ($1000s)	173.604	194.717	-8.856***	-5.514***	-12.843***	0.001
			(1.037)	(1.578)	(1.489)	
N (Employed as a teacher)	14,528	14,425	442,510	201,864	133,272	335,136
Panel C. 10 years after certification exam						
Employed as a teacher	0.459	0.553	-0.019***	-0.017*	-0.023**	0.691
			(0.007)	(0.010)	(0.010)	
Years of teaching experience	5.477	6.248	-0.457***	-0.432***	-0.524***	0.450
			(0.060)	(0.089)	(0.080)	
Annual earnings ($1000s)	54.772	56.658	-1.517***	-2.103**	-1.777***	0.791
			(0.512)	(0.992)	(0.677)	
PV cumulative earnings since exam ($1000s)	296.041	338.626	-9.275***	-7.736**	-12.088***	0.377
			(2.351)	(3.784)	(3.039)	
N (Employed as a teacher)	9,848	9,734	310,404	145,790	88,718	234,508

Notes: The table presents the RD estimates of the impacts of failing a certification exam on teaching and earning outcomes, using running variables with adjusted scale. Panels A-C indicate the RD effects in one, five, and ten years after the exam, respectively. The dependent variables include four outcomes: an indicator of whether employed as a teacher, years of teaching experience, annual earnings, and the present value of cumulative earnings since the exam. Columns (A) and (B) display the dependent variable means of candidates with exam scores ten points above the threshold. Columns (C)-(E) show the RD coefficients using equation (1). Column (F) indicates the p-value testing the statistical differences of RD effects between white and URM groups in Columns (D) and (E), respectively. "N (Employed as a teacher)" indicates the sample size for the outcome "Employed as a teacher", serving as a proxy sample size for other outcomes. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: RD effects of failing a certification exam on outcomes — Heterogeneity within URM groups

	(A)	(B)	(C)	(D)	(E)	(F)
	Mean above threshold		RD coefficients			*P*-value: B=H
	Black	Hispanic	All URM	Black	Hispanic	
Panel A. 1 year after certification exam						
Employed as a teacher	0.561	0.508	-0.150***	-0.189***	-0.137***	0.002
			(0.007)	(0.014)	(0.009)	
Years of teaching experience	0.654	0.571	-0.152***	-0.183***	-0.144***	0.073
			(0.008)	(0.018)	(0.011)	
Annual earnings ($1000s)	43.243	40.584	-3.942***	-4.561***	-3.593***	0.160
			(0.323)	(0.558)	(0.366)	
PV cumulative earnings since exam ($1000s)	59.873	51.988	-5.672***	-6.634***	-5.093***	0.149
			(0.502)	(0.853)	(0.576)	
N (Employed as a teacher)	5,312	11,584	156,499	47,292	109,207	156,499
Panel B. 5 years after certification exam						
Employed as a teacher	0.650	0.690	-0.062***	-0.085***	-0.055***	0.098
			(0.008)	(0.015)	(0.010)	
Years of teaching experience	3.302	3.361	-0.451***	-0.573***	-0.412***	0.043
			(0.034)	(0.065)	(0.042)	
Annual earnings ($1000s)	52.941	51.110	-1.109***	-2.188***	-0.846*	0.100
			(0.377)	(0.653)	(0.469)	
PV cumulative earnings since exam ($1000s)	202.532	191.085	-12.817***	-16.851***	-10.813***	0.056
			(1.458)	(2.600)	(1.635)	
N (Employed as a teacher)	4,577	9,848	133,272	40,236	93,036	133,272
Panel C. 10 years after certification exam						
Employed as a teacher	0.501	0.577	-0.025***	-0.045**	-0.017	0.199
			(0.009)	(0.018)	(0.012)	
Years of teaching experience	5.911	6.404	-0.536***	-0.729***	-0.483***	0.197
			(0.075)	(0.155)	(0.100)	
Annual earnings ($1000s)	58.556	55.759	-1.764**	-3.501***	-0.724	0.064
			(0.721)	(1.232)	(0.739)	
PV cumulative earnings since exam ($1000s)	347.442	334.551	-12.274***	-17.281***	-10.249***	0.308
			(3.178)	(5.644)	(3.603)	
N (Employed as a teacher)	3,077	6,657	88,718	26,451	62,267	88,718

Notes: The table presents the RD estimates of the impacts of failing a certification exam on teaching and earning outcomes, focusing on heterogeneity within the URM groups. Panels A-C indicate the RD effects in one, five, and ten years after the exam, respectively. The dependent variables include four outcomes: an indicator of whether employed as a teacher, years of teaching experience, annual earnings, and the present value of cumulative earnings since the exam. Columns (A) and (B) display the dependent variable means of candidates with exam scores ten points above the threshold. Columns (C)-(E) show the RD coefficients using equation (1). Column (F) indicates the p-value testing the statistical differences of RD effects between Black and Hispanic groups in Columns (D) and (E), respectively. "N (Employed as a teacher)" indicates the sample size for the outcome "Employed as a teacher", serving as a proxy sample size for other outcomes. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: RD effects of failing a certification exam on outcomes — Heterogeneity by SES

	(A)	(B)	(C)	(D)	(E)	(F)
	Mean above threshold		RD coefficients			*P*-value: H=L
	High SES	Low SES	All	High SES	Low SES	
Panel A. 1 year after certification exam						
Employed as a teacher	0.522	0.518	-0.114***	-0.099***	-0.135***	0.007
			(0.007)	(0.010)	(0.009)	
Years of teaching experience	0.590	0.574	-0.116***	-0.100***	-0.138***	0.012
			(0.008)	(0.011)	(0.011)	
Annual earnings ($1000s)	39.687	40.208	-3.031***	-2.769***	-3.541***	0.148
			(0.268)	(0.378)	(0.363)	
PV cumulative earnings since exam ($1000s)	51.306	52.551	-4.031***	-3.128***	-5.302***	0.006
			(0.398)	(0.532)	(0.574)	
N (Employed as a teacher)	14,177	10,204	270,856	173,378	97,478	270,856
Panel B. 5 years after certification exam						
Employed as a teacher	0.639	0.705	-0.045***	-0.045***	-0.049***	0.773
			(0.007)	(0.010)	(0.010)	
Years of teaching experience	3.231	3.408	-0.334***	-0.311***	-0.387***	0.235
			(0.031)	(0.044)	(0.045)	
Annual earnings ($1000s)	49.865	51.478	-0.621	-0.015	-1.509***	0.043
			(0.398)	(0.566)	(0.486)	
PV cumulative earnings since exam ($1000s)	183.231	193.917	-8.664***	-7.518***	-11.209***	0.147
			(1.257)	(1.804)	(1.743)	
N (Employed as a teacher)	12,265	8,192	229,901	152,031	77,870	229,901
Panel C. 10 years after certification exam						
Employed as a teacher	0.483	0.571	-0.018*	-0.020*	-0.013	0.716
			(0.010)	(0.012)	(0.016)	
Years of teaching experience	5.780	6.379	-0.407***	-0.402***	-0.462***	0.685
			(0.077)	(0.101)	(0.106)	
Annual earnings ($1000s)	56.897	57.124	-1.949**	-2.509**	-1.587*	0.531
			(0.805)	(1.132)	(0.878)	
PV cumulative earnings since exam ($1000s)	315.340	339.155	-10.119***	-12.429***	-12.213***	0.971
			(3.040)	(4.385)	(3.888)	
N (Employed as a teacher)	8,232	4,725	150,934	106,544	44,390	150,934

Notes: The table presents the RD estimates of the impacts of failing a certification exam on teaching and earning outcomes, focusing on heterogeneity by teacher candidates' socioeconomic status (SES). A teacher candidate is defined as having a low SES if the candidate has ever enrolled in Free or Reduced-Price Lunch during K-12. Panels A-C indicate the RD effects in one, five, and ten years after the exam, respectively. The dependent variables include four outcomes: an indicator of whether employed as a teacher, years of teaching experience, annual earnings, and the present value of cumulative earnings since the exam. Columns (A) and (B) display the dependent variable means of candidates with exam scores ten points above the threshold. Columns (C)-(E) show the RD coefficients using equation (1). Column (F) indicates the p-value testing the statistical differences of RD effects between low and high SES groups in Columns (D) and (E), respectively. "N (Employed as a teacher)" indicates the sample size for the outcome "Employed as a teacher", serving as a proxy sample size for other outcomes. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: RD effects of failing a certification exam on outcomes — Heterogeneity by Race & SES

	(A)	(B)	(C)	(D)	(E)	(F)
	High SES			Low SES		
	RD coefficients		*P*-value:	RD coefficients		*P*-value:
	White	URM	W=URM	White	URM	W=URM
Panel A. 1 year after certification exam						
Employed as a teacher	-0.075***	-0.181***	0.000	-0.105***	-0.147***	0.009
	(0.011)	(0.018)		(0.013)	(0.009)	
Years of teaching experience	-0.080***	-0.177***	0.000	-0.113***	-0.151***	0.030
	(0.013)	(0.022)		(0.014)	(0.010)	
Annual earnings ($1000s)	-2.242***	-3.930***	0.032	-1.534***	-3.899***	0.000
	(0.443)	(0.616)		(0.534)	(0.354)	
PV cumulative earnings since exam ($1000s)	-2.517***	-5.409***	0.023	-2.934***	-5.805***	0.004
	(0.644)	(1.029)		(0.813)	(0.534)	
N (Employed as a teacher)	135,775	33,987	169,762	91,965	122,512	214,477
Panel B. 5 years after certification exam						
Employed as a teacher	-0.041***	-0.067***	0.263	-0.035**	-0.062***	0.114
	(0.011)	(0.020)		(0.014)	(0.009)	
Years of teaching experience	-0.263***	-0.471***	0.034	-0.340***	-0.453***	0.119
	(0.047)	(0.082)		(0.060)	(0.038)	
Annual earnings ($1000s)	-0.287	0.104	0.729	0.674	-1.462***	0.011
	(0.559)	(1.057)		(0.713)	(0.384)	
PV cumulative earnings since exam ($1000s)	-6.620***	-11.333***	0.234	-4.273	-13.306***	0.005
	(1.990)	(3.301)		(2.692)	(1.534)	
N (Employed as a teacher)	119,628	29,395	149,023	82,236	103,877	186,113
Panel C. 10 years after certification exam						
Employed as a teacher	-0.021	-0.027	0.829	-0.012	-0.022**	0.602
	(0.014)	(0.022)		(0.017)	(0.011)	
Years of teaching experience	-0.424***	-0.480**	0.807	-0.448***	-0.559***	0.525
	(0.110)	(0.190)		(0.143)	(0.092)	
Annual earnings ($1000s)	-2.453*	-1.915	0.800	-0.666	-1.735**	0.451
	(1.353)	(1.477)		(1.152)	(0.727)	
PV cumulative earnings since exam ($1000s)	-12.684**	-12.021*	0.939	-2.793	-13.343***	0.151
	(5.123)	(6.595)		(6.114)	(3.676)	
N (Employed as a teacher)	84,261	20,281	104,542	61,529	68,437	129,966

Notes: The table presents the RD estimates of the impacts of failing a certification exam on teaching and earning outcomes, focusing on heterogeneity by teacher candidates' race and SES. A teacher candidate is defined as having a low SES if the candidate has enrolled in Free or Reduced-Price Lunch during K-12. Panels A-C indicate the RD effects in one, five, and ten years after the exam, respectively. The dependent variables include four outcomes: an indicator of whether employed as a teacher, years of teaching experience, annual earnings, and the present value of cumulative earnings since the exam. Columns (A)–(C) report results for high-SES candidates, while Columns (D)–(F) report results for low-SES candidates. Columns (A) and (B) show RD coefficients using equation (1), and Column (C) indicates the *p*-value testing for differences between white and URM groups. Similarly, Columns (D) and (E) show RD coefficients for low-SES candidates, with Column (F) reporting the corresponding *p*-value. "N (Employed as a teacher)" indicates the sample size for the outcome "Employed as a teacher", serving as a proxy sample size for other outcomes. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Extrapolated effects of failing certification exam on outcomes

	(A)	(B)	(C)	(D)	(E)	(F)
	Exam takers within 10 pts of passing		All exam takers who failed		All exam takers (passed + failed)	
	White	URM	White	URM	White	URM
Panel A. 1 year after certification exam						
Employed as a teacher	-0.107***	-0.168***	-0.134***	-0.203***	-0.021***	-0.082***
	(0.004)	(0.003)	(0.003)	(0.002)	(0.000)	(0.001)
Years of teaching experience	-0.107***	-0.169***	-0.127***	-0.196***	-0.020***	-0.079***
	(0.005)	(0.004)	(0.003)	(0.002)	(0.001)	(0.001)
Annual earnings ($1000s)	-2.585***	-4.301***	-3.013***	-5.646***	-0.463***	-2.268***
	(0.155)	(0.135)	(0.109)	(0.077)	(0.017)	(0.032)
PV cumulative earnings since exam ($1000s)	-3.526***	-6.137***	-4.083***	-8.230***	-0.635***	-3.339***
	(0.236)	(0.204)	(0.166)	(0.115)	(0.026)	(0.048)
N (Employed as a teacher)	15,430	19,688	35,391	63,496	227,740	156,499
Panel B. 5 years after certification exam						
Employed as a teacher	-0.056***	-0.090***	-0.112***	-0.199***	-0.017***	-0.080***
	(0.004)	(0.004)	(0.003)	(0.002)	(0.000)	(0.001)
Years of teaching experience	-0.378***	-0.589***	-0.619***	-1.013***	-0.092***	-0.406***
	(0.018)	(0.016)	(0.012)	(0.009)	(0.002)	(0.004)
Annual earnings ($1000s)	-0.146	-1.800***	-0.564***	-4.130***	-0.084***	-1.638***
	(0.251)	(0.172)	(0.152)	(0.098)	(0.023)	(0.039)
PV cumulative earnings since exam ($1000s)	-7.420***	-14.665***	-11.452***	-26.659***	-1.705***	-10.687***
	(0.749)	(0.621)	(0.506)	(0.372)	(0.076)	(0.153)
N (Employed as a teacher)	13,208	16,484	30,058	53,425	201,864	133,272
Panel C. 10 years after certification exam						
Employed as a teacher	-0.033***	-0.045***	-0.077***	-0.127***	-0.010***	-0.048***
	(0.005)	(0.005)	(0.004)	(0.003)	(0.000)	(0.001)
Years of teaching experience	-0.563***	-0.749***	-0.984***	-1.509***	-0.129***	-0.568***
	(0.044)	(0.039)	(0.029)	(0.022)	(0.004)	(0.009)
Annual earnings ($1000s)	-0.445	-0.621**	-1.133***	-2.330***	-0.148***	-0.857***
	(0.438)	(0.257)	(0.290)	(0.147)	(0.038)	(0.054)
PV cumulative earnings since exam ($1000s)	-10.083***	-15.263***	-15.727***	-31.497***	-2.062***	-11.844***
	(1.785)	(1.362)	(1.178)	(0.802)	(0.155)	(0.306)
N (Employed as a teacher)	8,365	10,185	19,115	33,362	145,790	88,718

Notes: The table presents the extrapolated effects of failing a certification exam on teaching and earning outcomes. The extrapolation uses the observed relationship between the scores and outcomes of samples with an exam score range of 240 and 260, and employs a linear extrapolation to project the outcomes for candidates who are below the passing score. Panels A-C indicate the extrapolated effects in one, five, and ten years after the exam, respectively. The dependent variables include four outcomes: an indicator of whether being employed as a teacher, years of teaching experience, annual earnings, and the present value of cumulative earnings. Columns (A)-(B), (C)-(D), and (E)-(F) use exam takers with a score within ten points of the passing score, all exam takers with a score below the 240 cutoff, and all exam takers, respectively. "N (Employed as a teacher)" indicates the sample size for the outcome "Employed as a teacher", serving as a proxy sample size for other outcomes. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: Extrapolated effects of failing certification exam on outcomes
Lower bounds on White/URM gap

	(A)	(B)	(C)	(D)	(E)	(F)
	Exam takers within 10 pts of passing		All exam takers who failed		All exam takers (passed + failed)	
	White	URM	White	URM	White	URM
Panel A. 1 year after certification exam						
Employed as a teacher	-0.107***	-0.168***	-0.134***	-0.203***	-0.021***	-0.082***
	(0.004)	(0.003)	(0.003)	(0.002)	(0.000)	(0.001)
Years of teaching experience	-0.107***	-0.169***	-0.127***	-0.196***	-0.020***	-0.079***
	(0.005)	(0.004)	(0.003)	(0.002)	(0.001)	(0.001)
Annual earnings ($1000s)	-2.824***	-4.395***	-3.013***	-5.646***	-0.426***	-2.152***
	(0.155)	(0.135)	(0.109)	(0.077)	(0.016)	(0.030)
PV cumulative earnings since exam ($1000s)	-3.025***	-5.095***	-3.177***	-6.583***	-0.494***	-2.671***
	(0.236)	(0.204)	(0.166)	(0.115)	(0.026)	(0.047)
N (Employed as a teacher)	15,430	19,688	35,391	63,496	227,740	156,499
Panel B. 5 years after certification exam						
Employed as a teacher	-0.055***	-0.075***	-0.104***	-0.186***	-0.015***	-0.074***
	(0.004)	(0.004)	(0.003)	(0.002)	(0.000)	(0.001)
Years of teaching experience	-0.383***	-0.573***	-0.619***	-1.013***	-0.092***	-0.406***
	(0.018)	(0.016)	(0.012)	(0.009)	(0.002)	(0.004)
Annual earnings ($1000s)	-0.257	-1.466***	-0.899***	-4.420***	-0.013	-1.368***
	(0.251)	(0.172)	(0.152)	(0.098)	(0.019)	(0.036)
PV cumulative earnings since exam ($1000s)	-4.986***	-10.296***	-7.918***	-22.447***	-1.179***	-8.998***
	(0.749)	(0.621)	(0.505)	(0.372)	(0.076)	(0.152)
N (Employed as a teacher)	13,208	16,484	30,058	53,425	201,864	133,272
Panel C. 10 years after certification exam						
Employed as a teacher	-0.043***	-0.046***	-0.098***	-0.132***	-0.013***	-0.050***
	(0.005)	(0.005)	(0.004)	(0.003)	(0.000)	(0.001)
Years of teaching experience	-0.613***	-0.783***	-0.984***	-1.509***	-0.129***	-0.568***
	(0.044)	(0.039)	(0.029)	(0.022)	(0.004)	(0.009)
Annual earnings ($1000s)	-0.445	-0.621**	-1.133***	-2.330***	-0.105***	-0.733***
	(0.438)	(0.257)	(0.290)	(0.147)	(0.027)	(0.046)
PV cumulative earnings since exam ($1000s)	-10.083***	-15.263***	-15.727***	-31.497***	-2.062***	-11.844***
	(1.785)	(1.362)	(1.178)	(0.802)	(0.155)	(0.306)
N (Employed as a teacher)	8,365	10,185	19,115	33,362	145,790	88,718

Notes: The table presents the extrapolated effects of failing a certification exam on teaching and earning outcomes. The extrapolation employs a linear extrapolation to project the outcomes for candidates who are below the passing score, using a sample that generates the lower bounds on the white-URM gap. Panels A-C indicate the extrapolated effects in one, five, and ten years after the exam, respectively. The dependent variables include four outcomes: an indicator of whether being employed as a teacher, years of teaching experience, annual earnings, and the present value of cumulative earnings. Columns (A)-(B), (C)-(D), and (E)-(F) use exam takers with a score within ten points of the passing score, all exam takers with a score below the 240 cutoff, and all exam takers, respectively. "N (Employed as a teacher)" indicates the sample size for the outcome "Employed as a teacher", serving as a proxy sample size for other outcomes. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Proportion of first-year teachers who took each TExES certification exam by grade

(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)
	First-year teachers in 2005–2010						First-year teachers in 2011–2016					
TExES Exam	3rd grade	4th grade	5th grade	6th grade	7th grade	8th grade	3rd grade	4th grade	5th grade	6th grade	7th grade	8th grade
Panel A. Elementary school exams												
Generalist EC-4	0.753	0.659	0.249	0.107	0.076	0.065	0.184	0.157	0.071	0.045	0.041	0.038
Generalist EC-6	0.004	0.006	0.019	0.015	0.003	0.003	0.630	0.555	0.573	0.309	0.127	0.109
Bilingual Generalist EC-4	0.116	0.099	0.026	0.004	-	0.002	0.024	0.024	0.005	0.003	0.002	-
Bilingual Generalist EC-6	-	-	-	-	-	-	0.054	0.058	0.059	0.011	0.003	-
English as a Second Language (ESL)/Generalist EC-4	0.045	0.042	0.010	0.003	-	0.002	0.008	0.009	0.002	0.002	0.001	-
English as a Second Language (ESL)/Generalist EC-6	-	-	-	-	-	-	0.066	0.058	0.054	0.020	0.004	0.003
Panel B. Middle school exams												
Generalist 4-8	0.040	0.132	0.544	0.403	0.326	0.277	0.052	0.153	0.260	0.473	0.455	0.407
Bilingual Generalist 4-8	0.001	0.008	0.033	0.007	0.004	0.002	0.001	0.004	0.005	0.002	-	-
English as a Second Language (ESL)/Generalist 4-8	0.002	0.004	0.024	0.017	0.010	0.010	0.002	0.004	0.005	0.008	0.005	0.007
Mathematics 4-8	0.005	0.014	0.047	0.182	0.239	0.218	0.010	0.021	0.038	0.156	0.217	0.209
English Language Arts and Reading 4-8	0.009	0.030	0.063	0.191	0.214	0.179	0.012	0.021	0.025	0.114	0.190	0.158
Science 4-8	0.003	0.006	0.044	0.010	0.007	0.005	0.004	0.008	0.023	0.012	0.010	0.006
Social Studies 4-8	0.002	0.006	0.023	0.013	0.009	0.008	0.004	0.005	0.012	0.013	0.018	0.014
Mathematics/Science 4-8	0.001	0.005	0.024	0.032	0.034	0.032	0.005	0.009	0.015	0.031	0.041	0.038
English Language Arts and Reading/Social Studies 4-8	0.004	0.017	0.039	0.059	0.053	0.040	0.007	0.013	0.018	0.042	0.040	0.040
# first-year teachers	7,058	7,593	5,118	3,367	3,999	3,407	6,773	7,148	6,262	3,848	4,036	3,048

Notes: This table shows the proportion of first-year teachers who took each of the TExES certification exams listed in column (A) by grade. Columns (B)–(G) show statistics for first-year teachers in 2005–2010. Columns (H)–(M) show statistics for first-year teachers in 2011–2016. "-" represents values that are censored due to a small number of observations.

Table A12: Characteristics of math and ELA teachers by grade and exam timing

	(A)	(B)	(C)	(D)	(E)	(F)
		Grade 4		Grades 7–8		DiD
Characteristic	N	Pre-reform	Post-reform	Pre-reform	Post-reform	Coef. (SE)
Panel A. Math teachers						
Cert exam score	21,696	265.079	257.678	263.134	262.047	-6.750 (0.736)***
URM	21,024	0.473	0.329	0.339	0.383	-0.153 (0.028)***
Female	21,691	0.906	0.917	0.767	0.750	0.026 (0.022)
High school math score	12,128	0.539	0.579	1.000	0.954	0.109 (0.057)*
High school ELA score	12,077	0.546	0.685	0.678	0.677	0.117 (0.055)**
Math value-added	21,696	0.038	0.003	0.026	0.016	-0.019 (0.009)**
Panel B. ELA teachers						
Cert exam score	18,954	265.291	257.397	263.956	262.928	-7.346 (0.788)***
URM	18,517	0.459	0.321	0.357	0.339	-0.072 (0.030)**
Female	18,954	0.941	0.946	0.902	0.875	0.020 (0.019)
High school math score	10,023	0.479	0.520	0.506	0.459	0.089 (0.065)
High school ELA score	9,967	0.574	0.723	0.709	0.769	0.086 (0.067)
ELA value-added	18,954	0.016	-0.001	0.006	0.004	-0.014 (0.004)***

Notes: This table reports characteristics of math and ELA teachers who took the pre-reform (2003–2009) and post-reform (2011–2015) certification exams. The sample includes individuals who took the elementary or middle school certification exams (Table 3) as their first certification exam, who passed the exam, and who subsequently became grade 4 or grade 7–8 teachers. Panel A shows statistics for Math teachers, and Panel B shows statistics for ELA teachers. Column (A) shows the total number of teacher × year observations. Columns (B)–(C) show statistics for grade 4 teachers who took the pre- and post-reform exams. Columns (D)–(E) show statistics for grade 7–8 teachers who took the pre- and post-reform exams. Column (F) reports a difference-in-differences (DiD) coefficient that equals column (C) − column (B) − [column (E) − column (D)]. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A13: Disparate impact robustness checks — Math teachers

Statistic	Race	(A) Benchmark	(B) 20 quantiles of VA	(C) VA kernel densities	(D) Binary VA (median)	(E) Binary VA (bottom 10%)	(F) Leave-out VA	(G) Non-cognitive VA	(H) Averaged to overall VA	(I) HS exams as control
Panel A. Number of observations										
Observations (grade 4)	URM	5,096	5,096	5,096	5,096	5,096	5,096	5,096	5,096	4,667
	White	6,979	6,979	6,979	6,979	6,979	6,979	6,979	6,979	6,203
Observations (placebo)	URM	3,215	3,215	3,215	3,215	3,215	3,215	3,135	3,215	800
	White	5,734	5,734	5,734	5,734	5,734	5,734	5,536	5,734	1,677
Panel B. Elementary school exams and grade 4 teachers										
Ratio of pass rates	URM	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.673
	White	0.822	0.822	0.822	0.822	0.822	0.822	0.822	0.822	0.822
Ratio of VA distributions	URM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.031	1.000
	White	0.988	0.994	0.977	0.985	1.005	0.998	0.960	0.994	0.988
PassRate(r, y, e)	URM	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.687	0.673
	White	0.813	0.818	0.803	0.810	0.827	0.821	0.790	0.818	0.812
Disparate impact		-0.146***	-0.151***	-0.137***	-0.143***	-0.160***	-0.154***	-0.123***	-0.130***	-0.139***
		(0.015)	(0.020)	(0.012)	(0.010)	(0.008)	(0.014)	(0.022)	(0.014)	(0.017)
Panel C. Placebo (middle school exams and grade 7–8 teachers, except column I)										
Ratio of pass rates	URM	0.971	0.971	0.971	0.971	0.971	0.971	0.971	0.971	1.082
	White	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	1.005
Ratio of VA distributions	URM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.023	1.000
	White	0.999	1.025	0.991	0.993	0.997	1.003	0.997	1.000	1.027
PassRate(r, y, e)	URM	0.971	0.971	0.971	0.971	0.971	0.971	0.971	0.993	1.082
	White	0.979	1.004	0.971	0.973	0.977	0.983	0.977	0.979	1.032
Disparate impact		-0.008	-0.032	0.000	-0.002	-0.006	-0.012	-0.006	0.014	0.050
		(0.024)	(0.035)	(0.017)	(0.016)	(0.015)	(0.024)	(0.039)	(0.028)	(0.089)
Panel D. Mean ratio-of-ratios estimator										
Ratio of pass rates	URM	0.687	0.687	0.687	0.687	0.687	0.687	0.687	0.687	0.622
	White	0.839	0.839	0.839	0.839	0.839	0.839	0.839	0.839	0.818
Ratio of VA distributions	URM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.008	1.000
	White	0.989	0.971	0.986	0.992	1.008	0.995	0.963	0.995	0.962
PassRate(r, y, e)	URM	0.687	0.687	0.687	0.687	0.687	0.687	0.687	0.692	0.622
	White	0.830	0.815	0.827	0.832	0.846	0.835	0.808	0.835	0.787
Disparate impact		-0.144***	-0.128***	-0.141***	-0.146***	-0.159***	-0.148***	-0.122***	-0.143***	-0.164***
		(0.025)	(0.033)	(0.018)	(0.015)	(0.014)	(0.023)	(0.036)	(0.023)	(0.058)
Panel E. Ratio-of-ratios estimator for each y and e										
Ratio of pass rates	URM	0.662	0.662	0.687	0.687	0.662	0.662	0.687	0.661	0.600
	White	0.810	0.810	0.839	0.839	0.810	0.810	0.839	0.808	0.789
Ratio of VA distributions	URM	1.043	1.137	1.136	1.000	0.976	1.059	1.133	1.059	1.084
	White	1.054	1.124	1.012	1.031	0.977	1.045	1.070	1.048	1.024
PassRate(r, y, e)	URM	0.716	0.780	0.780	0.687	0.670	0.727	0.778	0.727	0.675
	White	0.884	0.943	0.850	0.865	0.820	0.877	0.898	0.880	0.838
Disparate impact		-0.168	-0.163	-0.070	-0.178***	-0.150***	-0.151	-0.120	-0.153	-0.163
		(0.136)	(0.123)	(294.567)	(0.038)	(0.028)	(0.131)	(0.137)	(0.128)	(0.163)

Notes: See Table A15.

Table A14: Disparate impact robustness checks — ELA teachers

Statistic	Race	(A) Benchmark	(B) 20 quantiles of VA	(C) VA kernel densities	(D) Binary VA (median)	(E) Binary VA (bottom 10%)	(F) Leave-out VA	(G) Non-cognitive VA	(H) Averaged to overall VA	(I) HS exams as control
Panel A. Number of observations										
Observations (grade 4)	URM	4,903	4,903	4,903	4,903	4,903	4,903	4,903	4,903	4,485
	White	7,087	7,087	7,087	7,087	7,087	7,087	7,087	7,087	6,277
Observations (placebo)	URM	2,280	2,280	2,280	2,280	2,280	2,280	2,280	2,280	985
	White	4,247	4,247	4,247	4,247	4,247	4,247	4,244	4,247	2,648
Panel B. Elementary school exams and grade 4 teachers										
Ratio of pass rates	URM	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.673
	White	0.822	0.822	0.822	0.822	0.822	0.822	0.822	0.822	0.822
Ratio of VA distributions	URM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.032	1.000
	White	0.953	0.976	0.978	0.985	1.004	0.961	1.021	0.979	0.954
PassRate(r, y, e)	URM	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.688	0.673
	White	0.784	0.802	0.804	0.810	0.826	0.790	0.839	0.805	0.785
Disparate impact		-0.117***	-0.136***	-0.137***	-0.143***	-0.159***	-0.123***	-0.173***	-0.116***	-0.111***
		(0.018)	(0.024)	(0.022)	(0.011)	(0.009)	(0.018)	(0.032)	(0.015)	(0.019)
Panel C. Placebo (middle school exams and grade 7–8 teachers, except column I)										
Ratio of pass rates	URM	0.930	0.930	0.930	0.930	0.930	0.930	0.930	0.930	1.061
	White	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	1.004
Ratio of VA distributions	URM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.030	1.000
	White	1.004	1.013	1.003	1.005	0.996	0.998	1.003	1.001	0.983
PassRate(r, y, e)	URM	0.930	0.930	0.930	0.930	0.930	0.930	0.930	0.958	1.061
	White	0.971	0.980	0.970	0.971	0.963	0.965	0.970	0.968	0.987
Disparate impact		-0.041	-0.049	-0.040**	-0.041***	-0.033**	-0.035	-0.040	-0.011	0.074
		(0.028)	(0.047)	(0.017)	(0.015)	(0.014)	(0.028)	(0.035)	(0.031)	(0.057)
Panel D. Mean ratio-of-ratios estimator										
Ratio of pass rates	URM	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.635
	White	0.851	0.851	0.851	0.851	0.851	0.851	0.851	0.851	0.819
Ratio of VA distributions	URM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.003	1.000
	White	0.949	0.963	0.974	0.981	1.007	0.963	1.017	0.977	0.971
PassRate(r, y, e)	URM	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.719	0.635
	White	0.808	0.819	0.829	0.834	0.857	0.819	0.865	0.831	0.795
Disparate impact		-0.091***	-0.102***	-0.112***	-0.117***	-0.140***	-0.102***	-0.148***	-0.112***	-0.160***
		(0.028)	(0.039)	(0.026)	(0.016)	(0.015)	(0.028)	(0.041)	(0.027)	(0.044)
Panel E. Ratio-of-ratios estimator for each y and e										
Ratio of pass rates	URM	0.717	0.692	0.717	0.717	0.717	0.717	0.717	0.717	0.635
	White	0.851	0.821	0.851	0.851	0.851	0.851	0.851	0.851	0.819
Ratio of VA distributions	URM	1.180	1.053	1.020	1.021	1.032	1.117	1.062	1.242	1.085
	White	1.201	1.104	0.978	0.989	1.006	1.028	1.113	1.223	1.069
PassRate(r, y, e)	URM	0.846	0.755	0.731	0.732	0.740	0.801	0.762	0.890	0.688
	White	1.021	0.939	0.832	0.841	0.855	0.874	0.947	1.040	0.875
Disparate impact		-0.175	-0.184*	-0.100	-0.109***	-0.116***	-0.073	-0.185	-0.150	-0.187
		(0.148)	(0.112)	(4.490)	(0.040)	(0.042)	(0.149)	(0.140)	(0.151)	(0.141)

Notes: See Table A15.

Table A15: Notes for Tables A13 and A14

Notes: Tables A13 and A14 present our estimates of the policy-relevant disparate impact of the EC-6 exam relative to the EC-4 exam. Table A13 presents results for math teachers using math value-added as a measure of teaching quality. Table A14 presents results for ELA teachers using ELA value-added as a measure of teaching quality. In each table, Panel A shows the number of teacher × year observations for the grade 4 and placebo analyses. Panel B presents our main disparate impact estimates based on equation (6) using the elementary school exams and grade 4 teachers. Panel C presents our placebo disparate impact estimates based on equation (6) using the middle school exams and grade 7–8 teachers (except for column I, which uses high school certification exams and grade 9 teachers). Panel D present estimates from our main ratio-of-ratios estimator defined by equation (C9) in Appendix C.2. Panel E present estimates from our alternative ratio-of-ratios estimator defined by equation (C10) in Appendix C.2.

For each of Panels B–E, we show the following statistics separately for white and URM individuals:

- The ratio of pass rates on the post-reform (2011–2015) and pre-reform (2003–2009) exams.
- The average (across levels of value-added y and potential experience e) ratio of the value-added distributions between teachers who took the post- and pre-reform exams.
- The average value of $\text{PassRate}(r, y, e)$ defined by equation (5), which is the product of the ratio of pass rates and the average ratio of the value-added distributions.
- Our estimate of policy-relevant disparate impact, which is the different between the average values of $\text{PassRate}(r, y, e)$ for URM and white teachers.

Parentheses contain standard errors computed from an individual-level bootstrap that accounts for variation in both teacher value-added estimates and the sample used to compute disparate impact with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Columns show results from different specification that test the robustness of our results to methodological decisions for computing disparate impact. The specifications for each column are as follows:

A. Our benchmark specification, in which values of teacher quality y are defined by deciles of the value-added distribution, and
B. We group value-added into 20 quantiles rather than into deciles.
C. We estimate kernel densities of value-added and then group the estimated density values into 100 percentiles.
D. We define two value-added groups: above and below the median.
E. We define two value-added groups: the bottom 10% and the top 90%.
F. We estimate a leave-out measure of value-added where the scale for test scores is defined only by experienced teachers who were certified before *either* the EC-4 or EC-6 exams were in place.
G. We follow Jackson (2018) in defining a measure of each teacher's non-cognitive using an index of grade retention, attendance, and disciplinary incidents.
H. We average across levels of teacher value-added y and potential experience e using the distribution of these values for *all* teachers who passed the EC-4 exam (rather the distribution for URM teachers who passed the EC-4 exam, as in our benchmark specification).
I. We use grade 9 teachers as a control group in our ratio-of-ratios estimator using high school math/ELA certification exams and teacher value-added for end-of-course exams in Algebra I and English I (rather than grade 7–8 teachers and middle school certification exams, as in our benchmark specification). Since the high school exams were modified in 2014, this specification restricts the post-reform period to 2011–2013.

Table A16: RD-DiD effects of harder certification exams on other teacher composition outcomes

	(A)	(B)	(C)	(D)	(E)	(F)
	Grade 3–4 departures			Grade 7–8 departures		
	Post-reform mean at τ_{ty} = -1	Post-reform RD	Pre-reform RD	Post-reform RD	Pre-reform RD	RD-DiD
Panel A. Teachers' cert exam outcomes						
Average pass rate of teacher's exam(s)	0.807	-0.020***	-0.006***	-0.002*	-0.005***	-0.018***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Score on first attempt at TExES exam(s)	257.671	-1.266***	0.512	0.009	0.095	-1.692***
		(0.188)	(0.424)	(0.200)	(0.399)	(0.652)
Panel B. Teacher demographics						
White	0.584	-0.003	-0.017***	-0.022***	-0.019***	0.017**
		(0.003)	(0.003)	(0.004)	(0.005)	(0.008)
Hispanic	0.278	0.006**	0.012***	0.012***	0.005	-0.013**
		(0.002)	(0.003)	(0.003)	(0.003)	(0.006)
Black	0.119	-0.004*	0.003	0.007**	0.009***	-0.005
		(0.002)	(0.002)	(0.003)	(0.003)	(0.006)
Male	0.091	-0.003	0.007***	0.018***	0.012***	-0.016**
		(0.002)	(0.003)	(0.004)	(0.004)	(0.006)
High school math score	0.467	-0.034***	-0.013	-0.016	0.089**	0.084
		(0.012)	(0.023)	(0.023)	(0.040)	(0.053)
High school ELA score	0.526	0.004	0.005	0.021	0.023	0.001
		(0.012)	(0.022)	(0.020)	(0.036)	(0.048)
Panel C. Teacher pay						
Earnings growth from last year	3,277	3,213***	4,119***	3,769***	4,517***	-158
		(187)	(281)	(235)	(319)	(520)
Total earnings growth for all teachers	12,976	12,956***	15,635***	7,010***	7,876***	-1,813
		(773)	(1,097)	(454)	(507)	(1,478)
Total earnings growth for URM teachers	5,051	4,448***	6,453***	2,485***	2,092***	-2,397***
		(498)	(647)	(251)	(239)	(868)
Total earnings growth for White teachers	7,577	8,027***	8,308***	4,164***	5,549***	1,105
		(516)	(803)	(321)	(417)	(1,083)
N (# school/grade/departure/years)	4,887	105,829	86,102	123,706	107,812	423,449

Notes: This table displays RD and RD-DiD estimates of the effects of the TExES reform on teacher composition as defined by their certification exam outcomes (Panel A), demographics (Panel B), and pay (Panel C). Column (A) shows the mean of each outcome in the year prior to the teacher departure ($\tau_{ty} = -1$) in school/grades that experienced a departure in the post-reform years (2011–2016). Columns (B)–(E) show RD coefficients β from equation (7) estimated separately for grades 3–4 and 7–8, and for departures in 2011–2016 (post-reform) and 2005–2010 (pre-reform). Column (F) shows the RD-DiD coefficient θ from equation (8). Standard errors in parentheses are clustered at the school level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A17: RD-DiD effects of harder certification exams on student achievement with τ_{ty} = -3 to +3 RD window

	(A)	(B)	(C)	(D)	(E)	(F)
	Grade 3–4 departures			Grade 7–8 departures		
	Post-reform mean at τ_{ty} = -1	Post-reform RD	Pre-reform RD	Post-reform RD	Pre-reform RD	RD-DiD
Panel A. Student demographics (balance tests)						
Number of students with test scores	74.197	-0.306	0.972***	-6.529***	-2.280***	2.971**
		(0.263)	(0.326)	(0.778)	(0.836)	(1.214)
Male	0.497	-0.001	0.003***	0.004***	0.006***	-0.002
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
White	0.270	-0.022***	-0.040***	-0.037***	-0.050***	0.005**
		(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Hispanic	0.530	0.018***	0.038***	0.031***	0.043***	-0.008***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Black	0.149	-0.004***	-0.004***	-0.002*	-0.000	0.001
		(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Economically disadvantaged	0.664	0.007***	0.046***	0.026***	0.054***	-0.011***
		(0.001)	(0.002)	(0.002)	(0.002)	(0.003)
In gifted education	0.094	0.001	-0.002**	-0.015***	-0.004***	0.014***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
At risk of dropping out	0.471	0.033***	0.049***	0.087***	0.031***	-0.072***
		(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
Demographic index (Math score)	-0.003	0.015***	0.001	0.021***	0.001	-0.006
		(0.001)	(0.002)	(0.002)	(0.002)	(0.004)
Demographic index (ELA score)	-0.007	0.015***	0.002	0.016***	0.007***	0.006
		(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
Panel B. Student achievement						
Math score	-0.028	-0.002	-0.014***	0.007	-0.012**	-0.008
		(0.004)	(0.005)	(0.006)	(0.006)	(0.011)
Math score residuals	-0.007	0.000	-0.003	-0.001	-0.005	-0.000
		(0.003)	(0.003)	(0.004)	(0.003)	(0.007)
ELA score	-0.029	-0.004	-0.012***	0.002	-0.000	0.006
		(0.003)	(0.004)	(0.004)	(0.004)	(0.007)
ELA score residuals	-0.005	-0.002	-0.005**	-0.003*	0.001	0.006
		(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
N (# *sty* observations)	14,661	34,209	26,551	39,779	33,341	133,880

Notes: This table displays estimates of the effects of the TExES reform on student composition (Panels A) and student achievement (Panel B) from an RD-DiD specification with a larger RD window. Column (A) shows the mean of each outcome in the year prior to the teacher departure ($\tau_{ty} = -1$) in school/grades that experienced a departure in the post-reform years (2011–2016). Columns (B)–(E) show coefficients β from a modified version of equation (7) in which we omit the running variables terms (τ_{ty} and $\mathbf{1}\{\tau_{ty} \geq 0\}\tau_{ty}$) and include only observations from $\tau_{ty} = -3$ to $+3$ in the regression sample. We estimate these β coefficients separately for grades 3–4 and 7–8, and for departures in 2011–2016 (post-reform) and 2005–2010 (pre-reform). Column (F) shows θ coefficients from a version of our RD-DiD specification (9) with the same modifications as the RD regression (see equation B5 in Appendix B.3.2) . Standard errors in parentheses are clustered at the school level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A18: RD-DiD effects of harder certification exams on student achievement — Heterogeneity by race

	(A)	(B)	(C)	(D)	(E)	(F)
	Grade 3–4 departures			Grade 7–8 departures		
	Post-reform mean at τ_{ty} = -1	Post-reform RD	Pre-reform RD	Post-reform RD	Pre-reform RD	RD-DiD
Panel A. Math scores						
Math score (All students)	-0.039	-0.004	-0.006	0.014**	-0.002	-0.014
		(0.004)	(0.005)	(0.005)	(0.005)	(0.010)
Math score (White students)	0.292	-0.008	-0.001	0.005	-0.016***	-0.028*
		(0.007)	(0.009)	(0.007)	(0.006)	(0.015)
Math score (URM students)	-0.210	-0.004	0.001	0.015***	0.006	-0.014
		(0.005)	(0.007)	(0.006)	(0.006)	(0.012)
Math score (Hispanic students)	-0.146	-0.005	-0.007	0.015**	0.005	-0.008
		(0.006)	(0.008)	(0.006)	(0.007)	(0.013)
Math score (Black students)	-0.440	0.004	0.016	0.018**	0.007	-0.022
		(0.010)	(0.012)	(0.008)	(0.008)	(0.019)
Panel B. ELA scores						
ELA score (All students)	-0.035	0.000	-0.010**	0.009***	0.006	0.007
		(0.004)	(0.004)	(0.003)	(0.004)	(0.008)
ELA score (White students)	0.377	-0.013*	-0.013*	-0.001	0.007	0.007
		(0.007)	(0.007)	(0.005)	(0.006)	(0.012)
ELA score (URM students)	-0.229	0.003	-0.011*	0.012***	-0.002	-0.000
		(0.004)	(0.006)	(0.004)	(0.007)	(0.011)
ELA score (Hispanic students)	-0.199	0.006	-0.014**	0.013***	-0.008	-0.002
		(0.005)	(0.007)	(0.004)	(0.008)	(0.012)
ELA score (Black students)	-0.336	-0.009	-0.009	0.006	0.009	0.003
		(0.008)	(0.010)	(0.006)	(0.009)	(0.017)
N (# school/grade/departure/years)	4,887	105,829	86,102	123,706	107,812	423,449

Notes: This table displays RD and RD-DiD estimates of the effects of the TExES reform on student math (Panel A) and ELA (Panel B) test scores among different race/ethnicity groups. Column (A) shows the mean of each outcome in the year prior to the teacher departure ($\tau_{ty} = -1$) in school/grades that experienced a departure in the post-reform years (2011–2016). Columns (B)–(E) show RD coefficients β from equation (7) estimated separately for grades 3–4 and 7–8, and for departures in 2011–2016 (post-reform) and 2005–2010 (pre-reform). Column (F) shows the RD-DiD coefficient θ from equation (8). Standard errors in parentheses are clustered at the school level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Empirical Appendix

B.1 Variable definitions.

- **Annual earnings.** Annual earnings from the TWC data in 2019 dollars. We sum any reported earnings across all four quarters, and convert to 2019 dollars using CPI data from the Bureau of Labor Statistics.
- **At risk of dropping out.** The proportion of students who are reported as being "At Risk of Dropping Out."
- **Certification exam score.** The individual's score on their first attempt at their first TExES content exam. Scores range from 100–300 with 240 representing a passing score.
- **Content exam.** Throughout the paper, we focus on an individual's first TExES content exam. We define content exams to include all TExES exams that lead to classroom teacher certification excluding the the Pedagogy and Professional Responsibilities (PPR) exams. We exclude TExES exams for Principals, School Counselors, School Librarians, Educational Diagnosticians, and the Performance Assessment for School Leaders (PASL). We exclude TExES exams that do not have score ranges in 100–300 (Art, French, German, Latin, Spanish, and AAFCS: Human Development & Family Studies). We also exclude EXCET exams (the precursor to the TExES exams) as well as all other non-TExES exams (ACTFL, PACT, TASC, TASC-ASL, TEXMAT, and edTPA). We identify the earliest date at which an individual took one of the TExES content exams and focus on outcomes related to that exam. If an individual took multiple TExES exams on their earliest date, we include observations for each separate exam.
- **Days between first try and passing.** The number of days between the date at which the individual first took the content exam and the date at which they first passed the exam. This variable is defined only for individuals who ever passed the exam, and is equal to zero for individuals who passed on their first try.
- **Demographic index (math/ELA score).** An index of predicted math/ELA scores (in SD units) based on a large vector of student covariates. The covariates are dummies for the full interaction of sex, race/ethnicity, economic disadvantage, at-risk of dropping out, special education, gifted education, grade, and year. The demographic indices are predicted values from regressions of math/ELA scores (standardized to the exam/year level) on these dummies, estimated separately for each test regime (TAAS, TAKS, and STAAR).
- **Economically disadvantaged.** The proportion of students who are eligible for free or reduced-price meals or who are reported as having "Other economic disadvantages".
- **Employed as a teacher.** The total paid full-time equivalent (FTE) years that the individual worked in any Texas public school in a given year since taking their first TExES content exam.

- **Ever passed any exam.** An indicator equal to one if the individual ever passed any TExES content exam (regardless of whether or not it was the first TExES content exam that they attempted).
- **Ever passed exam.** An indicator equal to one if the individual ever passed the first TExES content exam that they attempted.
- **Experienced URM teacher.** The proportion of teachers who are both URM and received their initial teaching certification more than five years ago relative to the year in which they are teaching
- **Experienced white teacher.** The proportion of teachers who are both white and received their initial teaching certification more than five years ago relative to the year in which they are teaching
- **First-year teacher.** A teacher that meets two criteria: 1) They are in the first academic year in which we observe the teacher in the TEA staff files; and 2) the TEA employment files report that the teacher has 0 years of experience.
- **High school (math/ELA) score.** The math/ELA score we have for a teacher's high school math/ELA exams, standardized at the grade-year level. For the TAAS exam (1994–2002), we use the (mostly 10th grade) math and reading exit scores. For the TAKS exam (2003–2011), we use the grade 10 end-of-grade math and reading/ELA scores. For the STAAR exam (2012–2022), we use the end-of-course Algebra I score and the average of the end-of-course English 1 and English 2 scores. We use the individual's first score from any of these math/ELA exams. See the definition of math/ELA scores below for details on the test scores.
- **In gifted education.** The proportion of students who are reported as being in Gifted and Talented programs.
- **Math/ELA score (SD units).** Scale scores on standardized grade 3–8 math and English Language Arts (ELA) achievement tests, standardized to mean zero and SD one within the population of individuals who took the same exam in the same year. For each student $\times$ year $\times$ subject exam, we use the first non-missing score. Scores are from the three testing regimes that existed in Texas during our time period: TAAS (1994–2002), TAKS (2003–2011), and STAAR (2012–2022). We use Texas Learning Index (TLI) and Normal Curve Equivalent (NCE) scores for the TAAS regime, and scale scores for the TAKS and STAAR regimes. Scores for some exam takers in grades 3–5 are from Spanish language versions of these tests.
- **Math/ELA score residuals (SD units).** Residuals from a regression of math/ELA scores in SD units (defined above) on a large vector of student, school $\times$ grade, and school level controls. The controls variables mirror those in Chetty et al. (2014a)'s teacher value-added specification, except we use school/grade level controls rather than classroom controls because we only observe students' classrooms from 2012 onward. Student-level controls include cubics in lagged math and ELA scores, sex, economic

disadvantage, race/ethnicity dummies, at-risk of dropping out, gifted education, and indicators for missing values of the demographics variables. We also include school × grade and school averages of each of these variables. We interact all covariates with grade dummies, and include grade × year dummies in the regression. For this residual variable we include only students who have lagged test scores in both subjects, with the exception of grade 3 students since there are no grade 2 tests. Thus controls for grade 3 students include only those based on demographic variables, not lagged test scores.

- **New teacher (within 5 years of cert).** The proportion of teachers who received their initial teaching certification within five years relative to the year in which they are teaching.
- **New URM teacher.** The proportion of teachers who are both URM and received their initial teaching certification within five years relative to the year in which they are teaching.
- **New white teacher.** The proportion of teachers who are both white and received their initial teaching certification within five years relative to the year in which they are teaching
- **Number of students with test scores.** The total number of students in the school/-grade with non-missing math scores.
- **Number of times taking this exam.** The number of times the individual attempted the content exam.
- **Passed exam on first try.** An indicator equal to one if the individual passed their first TExES content exam on their first attempt.
- **Pedagogy and Professional Responsibilities (PPR) exam.** In Table 1, we include summary statistics for TExES PPR exams. These include the Pedagogy and Professional Responsibilities EC-4/EC-6/EC-12/4-8/8-12 exams and the Pedagogy & Prof Responsibilities Trade & Industrial Ed 6-12/8-12 exams.
- **Present value (PV) cumulative earnings since exam.** We deflate nominal earnings to the year in which the individual took their first TExES content exam using a 5 percent discount rate, and then sum across all years in the reported time range since taking the exam.
- **Total salary for all teachers.** The sum of total teaching salary from the TEA data for all teachers in a given school × grade × subject × year cell. For this we sum the base pay and other supplementary pay to get the teacher's total annual salary, multiply the annual salary by the proportion of full-time equivalent (FTE) years associated with each teaching assignment, and then sum across all teachers in the school × grade × subject × year.

- **Total salary for URM teachers.** The sum of total teaching salary from the TEA data for all URM teachers in a given school × grade × subject × year cell.
- **Total salary for white teachers.** The sum of total teaching salary from the TEA data for all white teachers in a given school × grade × subject × year cell.
- **Took the Generalist EC-4 exam.** The proportion of teachers who ever took and Generalist EC-4, Bilingual Generalist EC-4, or English as a Second Language (ESL)/Generalist EC-4 exam.
- **Took the Generalist EC-6 exam.** The proportion of teachers who ever took and Generalist EC-6, Bilingual Generalist EC-6, or English as a Second Language (ESL)/Generalist EC-6 exam.
- **Underrepresented minority (URM).** Individuals whose race/ethnicity is reported in the data as Black or Hispanic. Throughout the paper, when we compare outcomes for white and URM individuals, we exclude individuals whose race ethnicity is reported in the data as Native American, Asian, or Two or More Races (although these race/ethnicity groups are included when we examine outcomes for all individuals).
- **Years of teaching experience.** The sum of all paid full-time equivalent (FTE) years that the individual worked in any Texas public school in the reported time range since taking the exam.

B.2 Calculating Value-Added.

B.2.1 Teacher Value-Added Estimation. We obtain the academic and non-cognitive value-added estimate for each teacher j using the following equations (B1) and (B2) separately for each subject (sub)/non-cognitive measure:

$$A_{ijkgst}^{sub} = \alpha_1 A_{it-1}^{sub} + \alpha_2 A_{it-1}^{-sub} + \gamma X_{it} + \lambda C_{kgst} + \zeta S_{st} + \nu_{gt} + \epsilon_{ikgst} \tag{B1}$$

$$NC_{ijkgst} = \alpha_0 NC_{it-1} + \alpha_1 A_{it-1}^{sub} + \alpha_2 A_{it-1}^{-sub} + \gamma X_{it} + \lambda C_{kgst} + \zeta S_{st} + \nu_{gt} + \epsilon_{ikgst} \tag{B2}$$

For academic value-added estimation in equation (B1), A_{ijkgst}^{sub} is student i's standardized math or ELA score in year t, grade g, classroom k, and taught by teacher j in school s. We include a set of controls following Chetty et al. (2014a). Student i's A_{it-1}^{sub} and A_{it-1}^{-sub} represent lagged standardized math and ELA scores and their squares and cubes. X_{it} are student characteristics (economic disadvantage, ethnicity/race, sex, whether they are in special education, whether they are at risk, and whether they are gifted). Classroom characteristics, C_{kgst}, and school characteristics, S_{st}, include the mean individual characteristics, mean lagged standardized test scores in math and ELA and their squares and cubes for all students in classroom k and school s, respectively.[46] We also interact all student, class, and

[46]We implement a similar procedure as of Chetty et al. (2014a) to process the missing values when estimating the model: (i) we generate indicators for students with missing values in the other subject and include the interaction terms between the indicators and lagged own-subject test scores; (ii) we do the same thing for class-level lagged test scores across the two subjects; (iii) we include indicators for students with missing values in demographic controls.

school-level controls with grade-level to allow for differences in effect across grades. Lastly, to control for grade-year specific factors affecting all students, we include ν_{gt}.[47]

We implement a similar procedure as of Chetty et al. (2014a) for sampling. We exclude students with missing values in current or lagged test scores in the subject for which we are estimating VA. We restrict students who have only one teacher per subject for the whole year. We restrict to teachers who taught the school-grade-subject cell for the whole year ($\geq$ 275 days), as well as restrict to students who enrolled in the class of a school-grade-subject cell for the whole year ($\geq$ 275 days). We keep classrooms which have at least 10 but no more than 50 observations with current and lagged scores in the relevant subject. We also keep the observation associated with the most recent classrooms.

When estimating non-cognitive value-added, we employ the equation (B2). NC_{ijkgst} is an index that measures student i's non-cognitive outcomes in year t, grade g, classroom k, and taught by teacher j in school s. The index is constructed using students' behavioral outcomes: attendance (percentage of days that student was present), grade retention (indicator), and suspension (indicator) (Jackson, 2018).[48] In addition to the same set of controls used in equation (B1), equation (B2) also controls for lagged values of the non-cognitive index, including its linear, squared, and cubic terms, along with their interactions with grade fixed effects.[49]

Lastly, we obtain teacher-year estimates of academic ($\bar{A}_{jt}$) and non-cognitive value-added ($\bar{NC}_{jt}$) by estimating equations (B1) and (B2) and performing the following exercises, separately for each subject:

1. Obtain residuals by taking the difference between the student test score and the estimates from the previous regression with all controls except the teacher dummy:
   $a_{it} = A_{it} - [\hat{\alpha}_1 A^{sub}_{it-1} + \hat{\alpha}_2 A^{-sub}_{it-1} + \hat{\gamma} X_{it} + \hat{\lambda} C_{kgst} + \hat{\zeta} S_{st} + \nu_{gt}]$
   $nc_{it} = NC_{it} - [\hat{\alpha}_0 NC_{it-1} + \hat{\alpha}_1 A^{sub}_{it-1} + \hat{\alpha}_2 A^{-sub}_{it-1} + \hat{\gamma} X_{it} + \hat{\lambda} C_{kgst} + \hat{\zeta} S_{st} + \nu_{gt}]$

2. Then average the residuals among teacher j's students at the teacher-year:
   $\bar{A}_{jt} = \text{Mean}[a_{it} | i \in \{i : j(i,t) = j\}]$
   $\bar{NC}_{jt} = \text{Mean}[nc_{it} | i \in \{i : j(i,t) = j\}]$

$\bar{A}_{jt}$ and $\bar{NC}_{jt}$ are our estimates of teacher-year academic and non-cognitive value-added, respectively. The estimated value-added is essentially the classroom-year residual where $\hat{\beta}$ is estimated using within teacher variation. This follows Chetty et al. (2014a) by accounting for non-random sorting of students across teachers and its effect on the estimation of student, class, and school characteristics. We differ from Chetty et al. (2014a) by not completing their final step which calculates each $\bar{A}_{jt}$ (or $\bar{NC}_{jt}$) as a function of other years' estimates

[47] We also include fixed-effects for served student population type (e.g., regular, honors, gifted and talented, bilingual, and special education) defined at the classroom-level interacted with grade.

[48] Similar to Jackson (2018), we first estimate an iterated principal factors method on the behavioral outcomes. There is only one principal component (the first eigenvalue is 0.49 and the second and third are 0.0002 and -0.0002). We then computed the unbiased prediction of this sole component using the Bartlett method. The predicted non-cognitive index equals 1.21 (attendance) - 0.53 (grade retention) - 0.56(log of 1 + suspension).

[49] The non-cognitive value-added estimation thus excludes students with missing current or lagged values in grade retention, attendance, or suspension.

(drift). We deviate because we prefer to have an estimate for teachers with only one year of teaching, which is not possible in their method. In practice, the estimates on value-added between the two approaches are highly correlated.

B.2.2 Differences in Teacher Effectiveness across Groups. Rather than directly estimating teacher value added, we can alternatively assess how teacher effectiveness very across groups following Kane et al. (2008). To do so, we add an additional indicator, URM, into the VA equations (B1) and (B2), which gives the following specification for academic and non-cognitive value added:

$$A^{sub}_{ijkgst} = \alpha_1 A^{sub}_{it-1} + \alpha_2 A^{-sub}_{it-1} + \gamma X_{it} + \lambda C_{kgst} + \zeta S_{st} + \nu_{gt} + URM_{jt} + \epsilon_{ikgst} \quad \text{(B3)}$$

$$NC_{ijkgst} = \alpha_0 NC_{it-1} + \alpha_1 A^{sub}_{it-1} + \alpha_2 A^{-sub}_{it-1} + \gamma X_{it} + \lambda C_{kgst} + \zeta S_{st} + \nu_{gt} + URM_{jt} + \epsilon_{ikgst} \quad \text{(B4)}$$

Given that we assess the teacher effectiveness among exam takers in different race groups, we include exam-year FEs as in the baseline RD equation (1). In some models, we further include terms that measure exam performance — an indicator for passing the exam and certification exam scores — to assess how teacher effectiveness vary by performance and race for exam takers who end up teaching. In these models, we include alternative exam-year-race, exam-year-passed exam, or exam-year-race-passed exam FEs. Finally, we cluster standard errors at the school-level.

B.3 RD-DiD specification. This section provides details on our regression discontinuity difference-in-differences (RD-DiD) specification that we use in Section 5

B.3.1 Definition of teacher departures. We define a *teacher departure* as an instance in which a teacher with five or more years of experience leaves a given school. In other words, a departure occurs when a teacher with 5+ years of experience appears in a school one year but does not appear at that same school in the next year. Teachers that change subjects or grades within the same school are not counted as departures. We let t denote calendar years, y denote the year of the teacher departure, and τ_{ty} denote years relative to the departure. We define the departure year y as the first year that the teacher is no longer at the school, or, equivalently, $\tau_{ty} = 0$.

The sample of teacher departures that we include in our RD-DiD analysis includes two main restrictions. First, we consider only the departure of grade 3–4 and grade 7–8 teacher in subjects for which we have test scores. In most of our analyses, we define focus on grade 3–4 *generic* subject teacher, where generic means a teacher that taught any of the following subjects (with associated `subject` codes in the ERC data): General Science (8), Mathematics (10), English (22), Reading (27), Social Studies (38), and Generic (98). We combine all of these subjects into a single generic subject because many elementary teachers teach all of these core subjects, and in many cases the TEA data codes the teaching subject as Generic (98). For Table 7, however, we focus on grade 3–4 teachers who are specifically reported as teaching either Mathematics (`subject` = 10) or English/Reading (`subject` = 22 or 27). For grade 7–8 teachers, we require that the departing teacher taught either Mathematics (`subject` = 10) or English/Reading (`subject` = 22 or 27), and we treat the two subjects separately for our analysis.

Second, we include only teachers who taught one third or more of the students in a given

school, grade, and subject in the year prior to their departure year. We sum the teacher's full-time equivalent (FTE) years in the school/grade/subject, divide it by the total FTE in that school/grade/subject, and keep only departures in which the teacher's FTE is one-third or more of the total FTE. This restriction allows us to focus on instances in which the departing teacher causes a significant change in teacher composition at the school/grade/subject level, which is the level at which we can connect teachers to student test scores across all years of our data.

Teacher departures that meet both the subject and the FTE requirements are included in our RD-DiD analysis. We let s denote the school/grade/subject triplets that are associated with each teacher departure, and stack our dataset to include observations associated with each departure event as described in Section 5.1.[50]

B.3.2 Benchmark RD-DiD specifications. This subsection provides details on our RD-DiD regression specifications.

In Tables 5–6 and Appendix Tables A16 and A18), column (F) displays RD-DiD coefficients θ from equation (9) with Treated_g defined as an indicator for grades 3–4 and Post_p defined as an indicator for teacher departures in $y \in$ 2011–2016. As described in Sections 5.1–5.2, the intuition for equation (9) comes from a two-step specification. First, estimate the RD regression (7) separately for each pairwise combination of school exposure group $g \in$ {treated, control} and departure period $p \in$ {2005–2010, 2011–2016}, which gives four RD coefficients β_{gp}. Second, use the resulting RD coefficients β_{gp} as dependent variables in the simple DiD regression (8).

For Appendix Table A17, we estimate a modified version of equation (9) in which we omit the running variables terms (τ_{ty} and $\mathbf{1}\{\tau_{ty} \geq 0\}\tau_{ty}$) and include only observations from $\tau_{ty} = -3$ to $+3$ in the regression sample:

$$Y_{st} = \Big(\phi \text{Treated}_g + \delta \text{Post}_p + \theta \text{Treated}_g \text{Post}_p\Big)\mathbf{1}\{\tau_{ty} \geq 0\} + \gamma_{sy} + \varepsilon_{sty} \quad \text{if } |\tau_{ty}| \leq 3. \tag{B5}$$

In this specification, the intuition for the θ coefficient comes from the two step procedure: 1) Let β_{gp} represent the change in average outcomes in the three years after the teacher departure relative to the three years before the departure (rather than an RD coefficient, as in equation 7) for each pairwise combination of school exposure group $g \in$ {treated, control} and departure period $p \in$ {2005–2010, 2011–2016}; and 2) Use the resulting β_{gp} coefficients as dependent variables in the simple DiD regression (8).

B.3.3 RD-DiD specifications for Table 7. In Table 7, we restrict our sample to students in grades 3–4 who have a departing teacher that is identified in the data as specifically teaching math or ELA (see Appendix B.3.1) and use test scores or students in subjects that are unaffected by the departure as the control group. For this analysis, we also restrict our sample to school/grades that have teachers who are identified as teaching math or ELA specifically in *every* year from 2005–2016.

In Panels A–B of Table 7, we exploit within-school variation by using exam scores in the subject the teacher did *not* teach as a control group. In Panel A, we restrict our regression sample to school/grades with a departing math teacher. In Panel B, we restrict our regression

[50]Note that a teacher can depart from multiple grades or subjects in the same year if they teach more than one grade or subject.

sample to school/grades with a departing ELA teacher. In both cases we reshape our data so that there are two test score observations for each student/year (one in math and the other in ELA), and then estimate the following RD-DiD specification:

$$Y_{st} = \Big(\phi \text{SameSubject}_e + \delta \text{Post}_p + \theta \text{SameSubject}_g \text{Post}_p\Big)\mathbf{1}\{\tau_{ty} \geq 0\} + \\ \alpha_{gp}\tau_{ty} + \psi_{gp}\mathbf{1}\{\tau_{ty} \geq 0\}\tau_{ty} + \gamma_{sy} + \varepsilon_{sty} \quad \text{if } |\tau_{ty}| \leq h^Y_{gp}. \tag{B6}$$

where s now denotes school/grade/*exam-subject* triplets and SameSubject_e is an indicator equal to one if the exam subject is the same as the subject of the departing teacher (e.g., the math score when the math teacher departs, and the ELA score when the ELA teacher departs). All other variables in equation (B6) are defined as in our benchmark RD-DiD specification (9). Thus in Panel A, the θ coefficient shows how the effect of a math teacher departures on math scores changes relative to the effect on ELA scores. In Panel B, the θ coefficient shows how the effect of an ELA teacher departures on ELA scores changes relative to the effect on math scores.

In Panels C–D of Table 7, we exploit across-school variation by using school/grades with a departing teacher who taught the *other* subject as a control group. We again restrict our regression sample to schools with a departing grade 3–4 math or ELA teacher. We estimate the same RD-DiD specification (B6) as above, but s now denotes school/grade/*teaching-subject* triplets. We continue to define SameSubject_e as an indicator equal to one if the exam subject is the same as the subject of the departing teacher. In Panel C, we use math scores as the outcome variable, and thus the θ coefficient shows how the effect of a math teacher departure on math scores changes relative to the effect of an ELA teacher departure on math scores. In Panel D, we use ELA scores as the outcome variable, and thus the θ coefficient shows how the effect of an ELA teacher departure on ELA scores changes relative to the effect of a math teacher departure on ELA scores.

C Theoretical Appendix

C.1 Policy-relevant disparate impact. This appendix develops a potential outcomes framework that defines policy-relevant disparate impact and shows the identification assumptions that allow us to estimate it.

We consider a population of prospective elementary school teachers characterized by their race, potential exam performance, and potential teaching quality. Let $R_i \in \{U, W\}$ denote the race of individual i, where U is URM and W is white. Let $T_i \in \{E, H\}$ denote whether individual i took an easy or hard test as their first certification exam. In our context, $T_i = E$ represents the easier EC-4 TExES exam and $T_i = H$ represents the harder EC-6 exam. Let D_i^E and D_i^H be potential outcomes indicating whether individual i would pass the easy and hard exams on their first try. We let $D_i = \mathbb{1}\{T_i = E\}D_i^E + \mathbb{1}\{T_i = H\}D_i^H$ be the observed indicator of whether individual i passed their exam. Lastly, we let Y_{it}^* be individual i's potential teaching quality in year t if they were to become a teacher. Potential teaching quality Y_{it}^* exists for everyone, but we only observe it for individuals who pass their exam and enter the teaching profession. Our analysis uses teacher value-added as a measure of quality, and we compare teachers at the same level of potential experience $e(t)$, defined as the number of years since individuals took their first certification exam. We denote observed value-added by Y_{it}.

Given this notation, we define the *policy-relevant disparate impact* (PRDI) at a particular value of teacher value-added $Y_{it}^* = y$ and level of potential experience $e(t) = e$ as:

$$\begin{aligned} \text{PRDI}(y, e) \equiv \quad & \Pr(D_i^H = 1 | R_i = U, T_i = H, D_i^E = 1, Y_{it}^* = y, e(t) = e) \\ & - \Pr(D_i^H = 1 | R_i = W, T_i = H, D_i^E = 1, Y_{it}^* = y, e(t) = e). \end{aligned} \tag{C1}$$

The object to be estimated in equation (C1) is $\text{PassRate}(r, y, e) \equiv \Pr(D_i^H = 1 | R_i = r, T_i = H, D_i^E = 1, Y_{it}^* = y, e(t) = e)$, i.e., the hard exam pass rate conditional on race $R_i = r$, a particular level of value-added $Y_{it}^* = y$, and potential experience $e(t) = e$. By Bayes' Rule, we have:[51]

$$\text{PassRate}(r, y, e) = \Pr(D_i^H = 1 | R_i = r, T_i = H, D_i^E = 1) \times \frac{\Pr(Y_{it}^* = y | R_i = r, T_i = H, D_i^E = 1, D_i^H = 1, e(t) = e)}{\Pr(Y_{it}^* = y | R_i = r, T_i = H, D_i^E = 1, e(t) = e)}. \tag{C2}$$

To identify these unobservable quantities, we make the following three assumptions.

Assumption 1. *Any prospective fourth grade teacher who passed the hard exam would also have passed the easy exam:* $D_i^H = 1 \implies D_i^E = 1$.

Assumption 2A. *Individuals' potential pass rates and value-added are independent of whether they took the easy exam or hard exam:* $T_i \perp\!\!\!\perp D_i^E, D_i^H, Y_{it}^*$.

[51] Equation (4) assumes $\Pr(D_i^H = 1 | T_i = H, D_i^E = 1, e(t) = e) = \Pr(D_i^H = 1 | T_i = H, D_i^E = 1)$ for all values of potential experience e. Heterogeneity in hard exam pass rates by potential experience is driven only by the timing of our data, which determines the years in which we observe teaching outcomes for both EC-4 and EC-6 exam takers. There is minimal heterogeneity on this dimension in our data, so we ignore it to simplify our analysis.

Assumption 3. *Among exam passers, the ratio of potential value-added distributions on the hard/easy exams is equal to that for observed value-added distributions (for all r, y, and e):*

$$\frac{Pr(Y_{it}^* = y|R_i = r, T_i = H, D_i = 1, e(t) = e)}{Pr(Y_{it}^* = y|R_i = r, T_i = E, D_i = 1, e(t) = e)} = \frac{Pr(Y_{it} = y|R_i = r, T_i = H, D_i = 1, e(t) = e)}{Pr(Y_{it} = y|R_i = r, T_i = E, D_i = 1, e(t) = e)}$$

Under these three assumptions, the pass rate for race $R_i = r$ at $Y_{it}^* = y$ and $e(t) = e$ is equal to:[52]

$$\text{PassRate}(r, y, e) = \frac{\Pr(D_i = 1|R_i = r, T_i = H)}{\Pr(D_i = 1|R_i = r, T_i = E)} \times \frac{\Pr(Y_{it} = y|R_i = r, T_i = H, D_i = 1, e(t) = e)}{\Pr(Y_{it} = y|R_i = r, T_i = E, D_i = 1, e(t) = e)}. \tag{C3}$$

Equation (C3) shows that we can estimate $\text{PassRate}(r, y, e)$ with two ratios of observable quantities: 1) the ratio of race-specific pass rates, $\Pr(D_i = 1)$, for the hard and easy exams; and 2) the ratio of race-specific distribution of value-added, Y_{it}, for exam takers who passed each exam.

We compute policy-relevant disparate impact by estimating $\text{PassRate}(r, y, e)$ using equation (C3), plugging these values into equation (C1) to compute $\text{PRDI}(y, e)$, and then averaging across values of value-added y and levels of potential experience e. In all of our analyses, we group values of teacher value-added into bins (e.g., deciles, quantiles, percentiles, or binary), and so we can define the average policy-relevant disparate impact as:

$$\text{PRDI} = \sum_e \sum_y [\text{PassRate}(U, y, e) - \text{PassRate}(W, y, e)] g(y|e) f(e), \tag{C4}$$

where $g(y|e)$ is the probability mass function (PMF) of teacher value-added y conditional on potential experience e, and $f(e)$ is the PMF of potential experience e. As a benchmark, we compute this average using the observed PMF of teacher value-added for URM teachers who passed the easy (EC-4) exam, i.e., $g(y|e) = \Pr(Y_{it} = y|R_i = U, T_i = E, D_i = 1, e(t) = e)$. Plugging this into the above expression and simplifying gives:

$$\text{PRDI} = \frac{\Pr(D_i = 1|R_i = U, T_i = H)}{\Pr(D_i = 1|R_i = U, T_i = E)} - \frac{\Pr(D_i = 1|R_i = W, T_i = H)}{\Pr(D_i = 1|R_i = W, T_i = E)} \times \text{AdjustmentFactor}, \tag{C5}$$

where AdjustmentFactor is equal to:

$$\sum_e \sum_y \frac{\Pr(Y_{it} = y|R_i = U, T_i = E, D_i = 1, e(t) = e)}{\Pr(Y_{it} = y|R_i = W, T_i = E, D_i = 1, e(t) = e)} \times \Pr(Y_{it} = y|R_i = W, T_i = H, D_i = 1, e(t) = e) \times f(e).$$

[52] The derivation of the ratio of pass rates in equation (C3) uses Bayes' rule plus Assumptions 1 and 2A:

$$\begin{aligned}\Pr(D_i^H = 1|R_i = r, T_i = H, D_i^E = 1) &= \frac{\Pr(D_i^E = 1|R_i = r, T_i = H, D_i^H = 1) \times \Pr(D_i^H = 1|R_i = r, T_i = H)}{\Pr(D_i^E = 1|R_i = r, T_i = H)} \\ &= \frac{\Pr(D_i^H = 1|R_i = r, T_i = H)}{\Pr(D_i^E = 1|R_i = r, T_i = E)}.\end{aligned}$$

C.2 Policy-relevant disparate impact with control group. This subsection shows how we identify policy-relevant disparate impact using grade 7–8 teachers as a control for time trends in certification exam pass rates and teacher quality. This allows us to relax Assumption 2A to a weaker assumption that is analogous to a "parallel trends" assumption in difference-in-differences analyses.

Our goal is to identify the race-specific pass-rates conditional on teaching quality $Y_{it}^* = y$ and potential experience $e(t) = e$ for prospective fourth grade teachers, i.e., $\text{PassRate}(r, y, e)$ as defined by equation (C2). For this derivation, we change notation for the exam indicator so that the possible values are $T_i \in \{\text{Pre}, \text{Post}\}$. Among prospective fourth grade teachers, $T_i = \text{Pre}$ denotes individuals who took the certification exam prior to 2010 when the EC-4 exam was in place ($T_i = E$ in the original notation), and $T_i = \text{Post}$ denotes individuals who took the certification exam after 2010 when the EC-6 exam was in place ($T_i = H$ in the original notation). We also use the indicator G_i to denote the teaching grade of individual i to make the conditioning on grade explicit. Given this modified notation, we can use equation (C2) to express $\text{PassRate}(r, y, e)$ as the product of two terms: the overall race-specific hard exam pass rate in our policy-relevant population, $\text{PassRate}(r)$, and the race-specific ratio of value-added for hard exam passers relative to all individuals in our policy-relevant population, $\text{VARatio}(r, y, e)$, as defined by the following equations:

$$\begin{aligned}
\text{PassRate}(r, y, e) &= \text{PassRate}(r) \times \text{VARatio}(r, y, e) \\
\text{PassRate}(r) &\equiv \Pr(D_i^H = 1 | R_i = r, G_i = 4, T_i = \text{Post}, D_i^E = 1) \\
\text{VARatio}(r, y, e) &\equiv \frac{\Pr(Y_{it}^* = y | R_i = r, G_i = 4, T_i = \text{Post}, D_i^E = 1, D_i^H = 1, e(t) = e)}{\Pr(Y_{it}^* = y | R_i = r, G_i = 4, T_i = \text{Post}, D_i^E = 1, e(t) = e)}.
\end{aligned}$$

We first describe how we identify each of these two terms, $\text{PassRate}(r)$ and $\text{VARatio}(r, y, e)$. Finally, we show how we use the estimators of these two terms to compute our average measure of policy-relevant disparate impact.

C.2.1 Identification assumptions. To identify the two terms that comprise $\text{PassRate}(r, y, e)$, we maintain Assumptions 1 and 3 as described in Appendix C.1. But instead of making Assumption 2A, we now make the following assumption.

Assumption 2B. *For both URM and white exam takers, $R_i \in \{U, W\}$, the following three conditions hold:*

(i) The ratio of pass rates for prospective grade 4 teachers on the post- and pre-reform exams would have been the same as that for prospective grade 7–8 teachers in the absence of the TExES reform:

$$\frac{Pr(D_i^E = 1 | R_i = r, G_i = 4, T_i = Post)}{Pr(D_i^E = 1 | R_i = r, G_i = 4, T_i = Pre)} = \frac{Pr(D_i^M = 1 | R_i = r, G_i = 7/8, T_i = Post)}{Pr(D_i^M = 1 | R_i = r, G_i = 7/8, T_i = Pre)}.$$

(ii) The average (across y and e) ratio of the value-added distributions for prospective grade 4 teachers who passed the post- and pre-reform exams would have been the same as that for prospective grade 7–8 teachers in the absence of the TExES reform:

$$E\left[\frac{Pr(Y_{it}^* = y|R_i = r, G_i = 4, T_i = Post, D_i^E = 1, e(t) = e)}{Pr(Y_{it}^* = y|R_i = r, G_i = 4, T_i = Pre, D_i^E = 1, e(t) = e)}\right] = E\left[\frac{Pr(Y_{it}^* = y|R_i = r, G_i = 7/8, T_i = Post, D_i^M = 1, e(t) = e)}{Pr(Y_{it}^* = y|R_i = r, G_i = 7/8, T_i = Pre, D_i^M = 1, e(t) = e)}\right].$$

(iii) Among prospective grade 4 teachers, the effect of the increase in exam difficulty from the TExES reform on the distribution of teacher value-added is independent (across values of y and e) of trends in teacher value-added in the absence of the reform:

$$\frac{Pr(Y_{it}^* = y|R_i = r, G_i = 4, T_i = Post, D_i^H = 1, e(t) = e)}{Pr(Y_{it}^* = y|R_i = r, G_i = 4, T_i = Post, D_i^E = 1, e(t) = e)} \perp\!\!\!\perp \frac{Pr(Y_{it}^* = y|R_i = r, G_i = 4, T_i = Post, D_i^E = 1, e(t) = e)}{Pr(Y_{it}^* = y|R_i = r, G_i = 4, T_i = Pre, D_i^E = 1, e(t) = e)}.$$

C.2.2 Identifying PassRate(r). Under Assumption (1), we can write PassRate(r) as:

$$\begin{aligned}
\text{PassRate} &= \frac{\Pr(D_i^H = 1|R_i = r, G_i = 4, T_i = \text{Post})}{\Pr(D_i^E = 1|R_i = r, G_i = 4, T_i = \text{Post})} \\
&= \frac{\Pr(D_i^H = 1|R_i = r, G_i = 4, T_i = \text{Post})}{\Pr(D_i^E = 1|R_i = r, G_i = 4, T_i = \text{Post})} \times \frac{\Pr(D_i^E = 1|R_i = r, G_i = 4, T_i = \text{Pre})}{\Pr(D_i^E = 1|R_i = r, G_i = 4, T_i = \text{Pre})} \\
&= \frac{\Pr(D_i = 1|R_i = r, G_i = 4, T_i = \text{Post})}{\Pr(D_i = 1|R_i = r, G_i = 4, T_i = \text{Pre})} \times \frac{\Pr(D_i^E = 1|R_i = r, G_i = 4, T_i = \text{Pre})}{\Pr(D_i^E = 1|R_i = r, G_i = 4, T_i = \text{Post})}.
\end{aligned}$$

The first line of this derivation uses Bayes' rule and Assumption 1. The second line multiplies by an identity. The third line rearranges terms and substitutes the observed outcome D_i for the potential outcomes D_i^H and D_i^E. The final result expresses PassRate(r) as a product of two terms. The first term on the righthand side is the ratio of pass rates for post- and pre-reform exam takers, which is observable in the data. The second term is a selection bias term that reflects the difference in easy exam pass rates between pre- and post-reform exam takers.

To eliminate this selection bias term, we bring in grades 7–8 teachers. We let $g = 7/8$ denote prospective seventh or eighth grade teachers, and we let D_i^M be an indicator for whether these individuals would pass the middle school exams required to become a certified teacher in these grades. $T_i \in \{\text{Pre}, \text{Post}\}$ still denotes individuals who took the exam in the pre-2010 or post-2010 periods, but importantly, the middle school exam did not change over this time period. In other words, the ratio of observable race-specific pass rates in the post- and pre-reform periods is equal to:

$$\frac{\Pr(D_i = 1|R_i = r, G_i = 7/8, T_i = \text{Post})}{\Pr(D_i = 1|R_i = r, G_i = 7/8, T_i = \text{Pre})} = \frac{\Pr(D_i^M = 1|R_i = r, G_i = 7/8, T_i = \text{Post})}{\Pr(D_i^M = 1|R_i = r, G_i = 7/8, T_i = \text{Pre})}$$

Our estimator, which we call a ratio-of-ratios (RoR) estimator, divides the ratio of observable post-/pre-reform pass rates for prospective grade 4 teachers by the ratio of observable post-/pre-reform pass rates for prospective grade 7–8 teachers. Under the "equal ratios"

assumption 2B(i), this estimator identifies PassRate(r):

$$\text{PassRateRoR}(r) \equiv \frac{\Pr(D_i = 1|R_i = r, G_i = 4, T_i = \text{Post})}{\Pr(D_i = 1|R_i = r, G_i = 4, T_i = \text{Pre})} \div \frac{\Pr(D_i = 1|R_i = r, G_i = 7/8, T_i = \text{Post})}{\Pr(D_i = 1|R_i = r, G_i = 7/8, T_i = \text{Pre})} \tag{C6}$$

$$= \text{PassRate}(r) \times \frac{\Pr(D_i^E = 1|R_i = r, G_i = 4, T_i = \text{Post})}{\Pr(D_i^E = 1|R_i = r, G_i = 4, T_i = \text{Pre})} \div \frac{\Pr(D_i^M = 1|R_i = r, G_i = 7/8, T_i = \text{Post})}{\Pr(D_i^M = 1|R_i = r, G_i = 7/8, T_i = \text{Pre})}$$

$$= \text{PassRate}(r). \quad \text{(Assumption 2B(i))}$$

C.2.3 Identifying VARatio(r, y, e). We can identify the VARatio(r, y, e) term following a similar procedure. Under Assumption 1 and 3, we can write VARatio(r, y, e) as:

$$\text{VARatio}(r, y, e) = \frac{\Pr(Y_{it}^* = y|R_i = r, G_i = 4, T_i = \text{Post}, D_i^H = 1, e(t) = e)}{\Pr(Y_{it}^* = y|R_i = r, G_i = 4, T_i = \text{Post}, D_i^E = 1, e(t) = e)}$$

$$= \frac{\Pr(Y_{it} = y|R_i = r, G_i = 4, T_i = \text{Post}, D_i = 1, e(t) = e)}{\Pr(Y_{it} = y|R_i = r, G_i = 4, T_i = \text{Pre}, D_i = 1, e(t) = e)} \times \frac{\Pr(Y_{it}^* = y|R_i = r, G_i = 4, T_i = \text{Pre}, D_i^E = 1, e(t) = e)}{\Pr(Y_{it}^* = y|R_i = r, G_i = 4, T_i = \text{Post}, D_i^E = 1, e(t) = e)}.$$

VARatio(r, y, e) is the product of: 1) the observable ratio of the teacher value-added distributions between post- and pre-reform exam passers; and 2) a selection bias term that reflects the difference in value-added distributions between pre- and post-reform exam takers who would have passed the easy exam.

For prospective grade 7–8 teachers, under an assumption that is analogous to Assumption 3, the observed ratio of the post- and pre-reform value added distributions is equal to:

$$\frac{\Pr(Y_{it} = y|R_i = r, G_i = 7/8, T_i = \text{Post}, D_i = 1, e(t) = e)}{\Pr(Y_{it} = y|R_i = r, G_i = 7/8, T_i = \text{Pre}, D_i = 1, e(t) = e)} = \frac{\Pr(Y_{it}^* = y|R_i = r, G_i = 7/8, T_i = \text{Post}, D_i^M = 1, e(t) = e)}{\Pr(Y_{it}^* = y|R_i = r, G_i = 7/8, T_i = \text{Pre}, D_i^M = 1, e(t) = e)}$$

Our main RoR estimator is based on the *average* race-specific ratio of observed value-added for grade 4 and grade 7–8 teachers, where the average is taken across values of value-added y and potential experience e. Letting $g(y|e)$ denote the PMF teacher value-added y conditional on potential experience e and $f(e)$ is the PMF of potential experience e (as in Appendix C.1), we define the expectation $E[\cdot]$ as the average across both y and e. Under Assumptions 2B(ii) and 2B(iii), our RoR estimator identifies the *average* race-specific value

of VARatio(r, y, e):

$$\begin{aligned}
\text{VARatioRoR}(r) \equiv & E\left[\frac{\Pr(Y_{it}=y|R_i=r, G_i=4, T_i=\text{Post}, D_i=1, e(t)=e)}{\Pr(Y_{it}=y|R_i=r, G_i=4, T_i=\text{Pre}, D_i=1, e(t)=e)}\right] \\
& \div E\left[\frac{\Pr(Y_{it}=y|R_i=r, G_i=7/8, T_i=\text{Post}, D_i=1, e(t)=e)}{\Pr(Y_{it}=y|R_i=r, G_i=7/8, T_i=\text{Pre}, D_i=1, e(t)=e)}\right] \qquad \text{(C7)} \\
= & E\left[\text{VARatio}(r, y, e) \times \frac{\Pr(Y^*_{it}=y|R_i=r, G_i=4, T_i=\text{Post}, D^E_i=1, e(t)=e)}{\Pr(Y^*_{it}=y|R_i=r, G_i=4, T_i=\text{Pre}, D^E_i=1, e(t)=e)}\right] \\
& \div E\left[\frac{\Pr(Y^*_{it}=y|R_i=r, G_i=7/8, T_i=\text{Post}, D^M_i=1, e(t)=e)}{\Pr(Y^*_{it}=y|R_i=r, G_i=7/8, T_i=\text{Pre}, D^M_i=1, e(t)=e)}\right] \\
= & E[\text{VARatio}(r, y, e)] \quad \text{(Assumptions 2B(ii) and 2B(iii))}
\end{aligned}$$

We can also identify VARatio(r, y, e) using an alternative assumption that assumes equal ratios of the teacher value-added distributions for each value of y (rather than equal ratios on average). Specifically, as an alternative to Assumption 2B, we can instead assume:

Assumption 2C. *For both URM and white exam takers, $R_i \in \{U, W\}$, the following two conditions hold:*

(i) Same as Assumption 2B(i).

(ii) For each value of y and e, the ratio of the value-added distributions for prospective grade 4 teachers who passed the post- and pre-reform exams would have been the same as that for prospective grade 7–8 teachers in the absence of the TExES reform:

$$\frac{Pr(Y^*_{it}=y|R_i=r, G_i=4, T_i=Post, D^E_i=1, e(t)=e)}{Pr(Y^*_{it}=y|R_i=r, G_i=4, T_i=Pre, D^E_i=1, e(t)=e)} = \frac{Pr(Y^*_{it}=y|R_i=r, G_i=7/8, T_i=Post, D^M_i=1, e(t)=e)}{Pr(Y^*_{it}=y|R_i=r, G_i=7/8, T_i=Pre, D^M_i=1, e(t)=e)}$$

When we maintain assumption 2C, we use an RoR estimator for each race $R_i = r$, each

value of y, and each value of potential experience e to identify the value of $\text{VARatio}(r, y, e)$:

$$\begin{aligned}\text{VARatioRoR}(r, y, e) \equiv & \frac{\Pr(Y_{it}^* = y | R_i = r, G_i = 4, T_i = \text{Post}, D_i = 1, e(t) = e)}{\Pr(Y_{it}^* = y | R_i = r, G_i = 4, T_i = \text{Pre}, D_i = 1, e(t) = e)} \\ & \div \frac{\Pr(Y_{it}^* = y | R_i = r, G_i = 7/8, T_i = \text{Post}, D_i = 1, e(t) = e)}{\Pr(Y_{it}^* = y | R_i = r, G_i = 7/8, T_i = \text{Pre}, D_i = 1, e(t) = e)} \qquad \text{(C8)} \\ = & \text{VARatio}(r, y, e) \\ & \times \frac{\Pr(Y_{it}^* = y | R_i = r, G_i = 4, T_i = \text{Post}, D_i^E = 1, e(t) = e)}{\Pr(Y_{it}^* = y | R_i = r, G_i = 4, T_i = \text{Pre}, D_i^E = 1, e(t) = e)} \\ & \div \frac{\Pr(Y_{it}^* = y | R_i = r, G_i = 7/8, T_i = \text{Post}, D_i^M = 1, e(t) = e)}{\Pr(Y_{it}^* = y | R_i = r, G_i = 7/8, T_i = \text{Pre}, D_i^M = 1, e(t) = e)} \\ = & \text{VARatio}(r, y, e) \quad \text{(Assumption 2C(ii))}.\end{aligned}$$

We find similar results using both: 1) the average RoR estimator $\text{VARatioRoR}(r)$ under Assumption 2B; and 2) the y- and e-specific RoR estimator $\text{VARatioRoR}(r, y, e)$ under Assumption 2C. However, in cases where y takes many different values, the $\text{VARatioRoR}(r, y, e)$ estimator has much larger standard errors due to noise from taking ratios with small sample sizes.

C.2.4 Estimating average policy-relevant disparate impact. As shown in Appendix C.1, our average measure of average policy-relevant disparate impact (PRDI) is equal to:

$$\begin{aligned}\text{PRDI} = & \sum_e \sum_y [\text{PassRate}(U, y, e) - \text{PassRate}(W, y, e)] g(y|e) f(e) \qquad \text{(C9)} \\ = & \text{PassRate}(U) \times \sum_e \sum_y [\text{VARatio}(U, y, e)] g(y|e) f(e) \\ & - \text{PassRate}(W) \times \sum_e \sum_y [\text{VARatio}(W, y, e)] g(y|e) f(e).\end{aligned}$$

Using our RoR estimators defined in equation (C6) and (C7), we can estimate PRDI as:

$$\text{PRDI} = \text{PassRateRoR}(U) \times \text{VARatioRoR}(U) - \text{PassRateRoR}(W) \times \text{VARatioRoR}(W).$$

Alternatively, using our RoR estimators defined in equation (C6) and (C8), we can estimate PRDI as:

$$\begin{aligned}\text{PRDI} = & \text{PassRateRoR}(U) \times \sum_e \sum_y \text{VARatioRoR}(U, y, e) g(y|e) f(e) \qquad \text{(C10)} \\ & - \text{PassRateRoR}(W) \times \sum_e \sum_y \text{VARatioRoR}(W, y, e) g(y|e) f(e).\end{aligned}$$