Disparate Impacts of Teacher Certification Exams

Christa Deneault*
Evan Riehl†
Jian Zou‡

February 10, 2026

Abstract

We use Texas administrative data to assess the long-standing claim that teacher certification exams discriminate against underrepresented minority (URM) candidates. In a regression discontinuity design, we find that failing a certification exam delays entry into teaching and costs the average candidate $10,000 in forgone earnings. These costs fall disproportionately on URM candidates both because they are more likely to fail and because their earnings losses from failing are 50 percent larger on average. To examine whether these disparities are justified by racial/ethnic differences in teaching quality, we develop a new measure of disparate impact and estimate it using a policy change that increased the difficulty of Texas' elementary certification exam. The harder exam reduced the URM share of new teachers but had no significant benefits for teaching quality or student achievement. Taken together, our findings show that certification exams have a disparate impact in the sense that they impose much larger economic costs on URM teaching candidates than on white candidates with similar potential 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 research reported in this article was made possible by a grant from the Spencer Foundation (#202600026). 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, the state of Texas, or the Spencer Foundation. For helpful comments we thank Barbara Biasi, Bobby W. Chung, Samuel Dodini, Caroline Hoxby, Peter Hull, Michael Lovenheim, Mindy Marks, Pia Orrenius, Jonah Rockoff, Tucker Smith, Reid Taylor, and participants at various seminars and conferences. 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 typically examine labor market discrimination from the perspective of individual agents, such as firms engaging in taste-based or statistical discrimination (Becker, 1959; Arrow, 1971). By contrast, the field has paid relatively little attention to how systemic or institutional features of labor markets may discriminate against certain groups of prospective workers (Small and Pager, 2020; Bohren et al., 2025b). A salient example is the requirement that workers pass licensing exams to work in many occupations, including medicine, law, accounting, and teaching. Licensing exams create racial disparities in access to these professions given differences in test-taking abilities between racial groups. Importantly, racial disparities in licensing exam pass rates may be considered illegal discrimination under Title VII of the Civil Rights Act even if the exams do not have discriminatory intent. The central legal question is whether the exams 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.

Identifying disparate impacts of licensing exams is empirically demanding because it requires credibly estimating both their costs for prospective workers and their potential benefits for worker quality. Few datasets contain labor market outcomes for all individuals who *attempt* to gain licenses, meaning that earnings are typically not observed for those who failed and were prevented from accessing an occupation. Instead research often estimates wage gaps between licensed and unlicensed workers (e.g., Kleiner and Krueger, 2013), which may be biased for the true costs of licensing barriers if licensed workers are positively selected. Further, in many occupations it is difficult to observe measures of worker performance, and even if good measures are available there is a fundamental censoring problem: job performance is not observed for individuals who do not pass exams. Thus it is unclear whether observed correlations between licensing exam scores and job performance (e.g., Clotfelter et al., 2007; Goldhaber, 2007) are informative about the potential productivity of workers who were prevented from obtaining licenses.

This paper addresses these challenges to provide new evidence on whether licensing exams have a disparate impact on prospective Black and Hispanic workers. Our analysis focuses on one of the most widely-taken licensing exams in the United States: certification tests for public school teachers. We use linked certification, education, and labor market data from Texas that contains samples and variables that are difficult to measure in the literature, including the population of certification exam takers, their race/ethnicity, and their earnings regardless of whether they end up teaching. Leveraging this comprehensive data and several quasi-experimental approaches, we provide the most compelling demonstration to date

that certification exams place a disproportionate burden on Black and Hispanic teaching candidates relative to white candidates of comparable teaching quality.

We begin by providing causal evidence on the economic costs of certification exams for prospective teachers. Using a regression discontinuity (RD) design, we find that narrowly failing a license exam delays entry into the teaching profession, with larger effects for candidates from underrepresented minority (URM) groups than for whites. In the year following their first certification exam attempt, whites are 9 percentage points (pp) less likely to be working in a public school if they narrowly fail, while URMs are 15pp less likely. Although these employment effects decline over time as candidates retake exams, individuals who are delayed or discouraged from teaching experience large earnings losses since teaching jobs give them a substantial pay increase. Our RD estimates show that marginally failing the exam costs whites $7,800 and URMs $12,200 in forgone earnings over a ten year period. Importantly, URM candidates are also much more likely to experience these earnings losses since they have lower first-attempt pass rates than white candidates (57 percent vs. 83 percent).

Next, we ask whether racial disparities in the costs of certification exams are justified by differences in teaching quality. For this we exploit a 2010 reform in which Texas replaced its early childhood through grade 4 (EC-4) certification exam with an early childhood through grade 6 (EC-6) exam. This reform made it much harder to become an EC–4 teacher because candidates also had to learn grade 5–6 content. Specifically, the change in exams reduced first-attempt pass rates for prospective elementary teachers from 94 percent to 77 percent for whites and from 68 percent to 45 percent for URMs.

In a potential outcomes framework, we show how this reform allows us to estimate the disparate impacts of certification exams for URM candidates. Following Arnold et al. (2022) and Baron et al. (2024), we define disparate impact as the difference in certification exam pass rates between URMs and whites with the same potential teaching quality. Yet our paper differs in that we are interested in the effects of a policy — licensing exam requirements — rather than the decisions of individuals. We therefore develop a new measure of *policy-relevant disparate impact* that estimates effects of the EC-6 exam *relative* to the EC-4 exam. Specifically, our measure estimates racial differences in pass rates on the EC-6 exam (conditional on teaching quality) among individuals who would have passed the EC-4 exam if it had remained in place. Under plausible assumptions, this focus allows us to overcome the fundamental difficulty in estimating disparate impact — that job performance is not observed for individuals who fail — by comparing the distributions of teaching quality among individuals who passed the easier and harder exams.[1]

[1]Specifically, we show that we can identify policy-relevant disparate impact under three assumptions: 1) individuals who passed the hard exam would have also passed the easy exam; 2) the timing of the exam

We find a large disparate impact in pass rates among white and URM candidates with the same potential teaching quality. We follow the literature in using test score value-added as a measure of teacher quality (Kane and Staiger, 2008; Chetty et al., 2014a). Among candidates who would have passed the EC-4 exam, URMs were 14.6pp (11.7pp) less likely to pass the EC-6 exam than whites with the same value-added for math (English) test scores. Notably, these estimates of policy-relevant disparate impact are similar to raw differences in pass rates between URMs and whites. This similarity arises because the value-added distributions for teachers who passed the easier and harder exams are similar for both URMs and whites, suggesting that the harder exams did little to screen out poor performing teachers. Our disparate impact estimates are robust to different ways of measuring and controlling for value-added and to using value-added for non-cognitive outcomes as an alternate measure of teacher quality (Jackson, 2018). In support of our identification assumptions, we find null effects on placebo measures of disparate impact for teachers at grade levels that were not affected by the reform, and we show that our results are robust to using these grade levels as a control group in a difference-in-differences (DiD) style analysis.

While these methods highlight the disparate impacts of certification exams for URM teaching candidates, they do not fully capture the potential supply-side effects of certification exam reforms for students. For example, harder certification exams can discourage prospective teachers from attempting the exams and reduce the number of teachers who pass, making it more difficult for schools to fill open teaching vacancies.

We therefore conclude the paper by examining the reduced-form effects of certification exam reforms on school staffing decisions and student achievement. For this we develop a new identification strategy that combines DiD variation in exposure to the EC-4/EC-6 exam reform with RD regressions that estimate sharp changes in outcomes following the departure of experienced teachers. Together, these two sources of variation allow us to examine how schools fill open teaching vacancies under easier or harder certification exam regimes and the resulting effects on student test scores.

We find that the increase in certification exam difficulty reduced the URM share of teachers without any measurable improvements in student achievement. Our RD-DiD estimates show that the adoption of the harder EC-6 exam reduced the share of newly-hired teachers who were URMs by 1.9pp, consistent with the effect of the reform on the racial/ethnic composition of exam passers. Yet these changes in teacher composition did not translate into higher test scores for students; we find small and statistically insignificant effects of the reform on math and English test scores. These null effects corroborate our disparate impact

reform is unrelated to trends in pass rates and teacher quality; and 3) the reform did not affect selection into teaching conditional on passing the exam.

results using alternative assumptions, and they additionally show that supply-side responses to certification exam difficulty do not meaningfully impact student achievement.

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 not justified by differences in teaching quality. A back-of-the-envelope calculation suggests that eliminating pass rate disparities on the harder EC-6 exam cost would increase the present value of future earnings for URM teaching candidates in Texas by $8.5 million in a typical year.

Our paper makes four main contributions to the literatures on occupational licensing, discrimination, and teacher certification. First, we provide the first credible evidence on the economic costs of licensing exams for prospective workers. Much of the existing work on the costs of occupational licensing compares wages of licensed and unlicensed workers or examines earnings changes before and after becoming licensed (Kleiner and Kudrle, 2000; Kleiner and Krueger, 2013; Thornton and Timmons, 2013; Gittleman et al., 2018; Kleiner and Xu, 2024; Farronato et al., 2024), approaches that are subject to selection bias. Our data and RD design circumvents these issues by comparing labor market outcomes for individuals who marginally pass or fail license exams.[2] Unlike prior work that studies changes in overall licensing stringency (Angrist and Guryan, 2008; Kleiner et al., 2016; Larsen et al., 2020; Law et al., 2023; Chung and Zou, 2025), we focus on individuals actively attempting to become licensed, which allows us to directly track workers and the costs they face. This distinction sheds light on why these studies often find weak or mixed effects of licensing stringency on labor supply and quality. In particular, we identify an important but previously overlooked mechanism: delayed entry. Failing the first licensing exam has only modest discouragement effects, implying a small supply response at the margin as most individuals eventually become licensed. Nevertheless, these delays lead to economically meaningful earnings losses.

Second, we contribute to the discrimination literature by examining how institutional features of labor markets can generate discriminatory outcomes (Small and Pager, 2020; Bohren et al., 2025b). In contrast, much of the existing work emphasizes the role of individual agents and their intentional or unintentional discriminatory behavior (see, e.g., Kline et al., 2022; Benson et al., 2024; Lepage, 2024; Bohren et al., 2025a, for recent examples). Within this broader literature, our paper contributes most directly to recent work that quantifies disparate impacts (Arnold et al., 2022; Baron et al., 2024) by focusing on a new use case—occupational licensing—and by introducing a novel measure that we argue is applicable across a wider range of settings. These papers exploit the presence of extremely lenient

[2]In a contemporary working paper, Tsao (2025) uses a regression kink design based on certification exam thresholds to estimate teacher pay premiums measured four years after the exam with the goal of identifying rents for Kentucky teachers. Our focus is different: we use an RD design to estimate racial/ethnic disparities in the causal effects of failing a Texas certification exams on career outcomes over the next ten years.

decision makers, e.g., judges who release nearly all bail defendants or screeners who place very few children in foster care. Combined with the random assignment of cases to decision makers, this allows the authors to observe average uncensored outcomes and thus overcome the data censoring issue inherent in estimating disparate impacts. While compelling in their respective settings, the approach of using lenient decision makers typically cannot be used to examine the disparate impacts of a policy. Our new measure shows how researchers can overcome this limitation by estimating disparate impacts for individuals affected by a policy change. This measure can be applied to other contexts in which policy changes alter screening rates, such as those commonly studied in the occupational licensing literature.

Third, our paper is unique in providing evidence on racial/ethnic disparities in the costs of licensing exams. Critics have long argued that licensing exams discriminate against URM candidates, and multiple lawsuits have been filed about teacher certification exams in particular.[3] Yet few datasets contain information on race/ethnicity and earnings for licensing exam takers without conditioning on employment outcomes. Thus existing research focuses on how exams impact the racial composition of the licensed occupation (e.g., Angrist and Guryan, 2008) but does not examine career consequences for individuals. Other work asks how licensing stringency more broadly affects workforce racial composition (Law and Marks, 2009) or racial wage gaps (Xia, 2021; Blair and Chung, 2025) but does not isolate the role of licensing exams. Our data and RD strategy allows us to uniquely show that not only are Black and Hispanic candidates more likely to fail certification exams than white candidates, but they also have more negative costs from failing.

Finally, we provide new and more convincing evidence that changes in certification exam difficulty have little overall impact on teacher quality or student achievement. Previous work has consistently found positive correlations between certification exam scores and teacher value-added (Clotfelter et al., 2007; Goldhaber, 2007; Clotfelter et al., 2010; Goldhaber and Hansen, 2010; Shuls and Trivitt, 2015; Hendricks, 2016; Goldhaber et al., 2017; Cowan et al., 2020), but it is unclear whether these correlations are informative about teaching quality among individuals who do not pass the exams.[4] Our approach uses a certification exam reform to overcome this challenge. Our findings suggest that correlations for the full

[3]These lawsuits have had mixed success. While New York City recently paid settlements to URM teachers for a long-running case on its certification tests (Green, 2023), similar lawsuits were withdrawn in Texas and Arkansas after plaintiffs failed to demonstrate that the exams imposed significant costs on individuals who failed (Rodman, 1987). In a California lawsuit, judges ruled that certification exams did not have disparate impact because they were a valid measure of teaching performance (https://clearinghouse.net/case/10826/).

[4]In a different approach using an RD design, Orellana and Winters (2023) find that failing a certification pushes out individuals who would have been *better* teachers. This approach only speaks to the quality of teachers on the margin of passing/failing, and it is difficult to know whether the findings are driven by selection into teaching or by changes in teacher experience due to delayed entry into the profession.

population of teachers may not be informative for the quality of teachers who are marginal to changes in certification exam difficulty. Further, the existing literature does not consider potential supply-side impacts of changes to exam stringency, such as changes in the number of certified teachers. Our RD-DiD analysis incorporates these potential supply-side adjustments and finds that harder certification exams did not improve student achievement yet resulted in a less diverse workforce.

2 Background and Data

2.1 Texas teacher certification exams. Certification is a requirement for public school teachers in Texas. To become certified, a candidate must hold a bachelor's degree, complete teacher training, and pass two exams: a content exam that is aligned with the candidate's intended teaching subject and grade level, and a general test on classroom management called the Pedagogy and Professional Responsibilities (PPR) exam. Since 2003, Texas administers these exams under the Texas Examinations of Educator Standards (TExES).[5] The exams consist of multiple choice questions and can take up to five hours to complete. Scores are scaled from 100–300 with a passing threshold set at 240. Exams are offered year round and cost $116 per attempt as of 2025. Candidates who do not pass may retake the exam up to four times with a mandatory 30-day waiting period between retakes.

We focus our analysis on the TExES content exams, which have lower pass rates and are most directly related to student standardized exams. There are many different content exams to cover all grades and all areas of teaching. Teachers who wish to teach early childhood education through grade 6 take a general exam that tests material in math, English language arts (ELA), science, and social studies. Grade 4–8 teachers can choose between general or subject-specific exams in these areas. The exams for grades 7–12 teachers cover a range of more specialized subjects (e.g., Chemistry or Computer Science). There are also exams for Bilingual Education, Special Education, and other non-core subjects at all grade levels (e.g., Art, Music, and Physical Education).[6]

2.2 Data. We use individual-level administrative data from four Texas agencies:

- **Texas State Board for Educator Certification (SBEC)**. SBEC provides data on all teacher certification exams in Texas between 2003 and 2022. This data includes

[5]The TExES exams replaced the previous testing regime, known as the Examination for the Certification of Educators in Texas (ExCET). The first TExES exams were offered in October 2002, although ExCET exams were offered in some subjects until 2006. Throughout the paper, we define years as academic years (e.g., 2003 represents July 2002 to June 2003).

[6]For details on the TExES exams, see: https://www.tx.nesinc.com/PageView.aspx?f=GEN_Tests.html.

exam dates, test type, and test scores. Our analyses focus on individuals' first teacher certification exam, which we identify using the exam dates.

- **Texas Education Agency (TEA)**. TEA provides data on all K–12 public school teachers and students in Texas from 1996–2022. For teachers, these data include demographics, salary, years of teaching experience, school of employment, and full-time equivalent (FTE) years associated with their teaching grade(s) and subject(s). Student data includes demographics, attendance, and disciplinary incidents. We also use TEA data on student standardized tests from 1994–2022, which include end-of-grade math and ELA exams in grades 3–8 and end-of-course high school exams in Algebra I and English I (typically taken in ninth grade). We can connect student test scores to groups of teachers at the school/grade/subject level for all years from 1996–2022. From 2012–2022, we also observe classroom identifiers that allow us to connect students to individual teachers and compute teacher value-added estimates.

- **Texas Higher Education Coordinating Board (THECB)**. We use THECB data on all bachelor's degree recipients from public (1992–2022) and private colleges (2003–2022) in Texas. We observe each graduate's college, major, graduation year, and demographics.

- **Texas Workforce Commission (TWC).** TWC provides earnings records for all individuals working at registered firms in Texas from 1992–2022. We use this data to measure annual earnings converted to 2019 dollars using the Consumer Price Index. We also calculate the present value (PV) of cumulative earnings for a given period of time since individuals took their first certification exam by discounting nominal earnings in each year using a 5% discount rate and summing across years.

We accessed the data via the University of Texas at Dallas Education Research Center (ERC), which contains unique personal identifiers that connect all datasets at the individual level. These linkages allow us to observe characteristics and outcomes that are often hard to measure in the teacher certification literature. The SBEC data, like most other certification exam datasets, does not include race/ethnicity, but we can identify race/ethnicity for any individual who attended a K–12 school or college in Texas.[7] Similarly, we observe a measure of ability for prospective teachers in the form of high school math and ELA test scores. Lastly, we observe labor market outcomes for certification exam takers whether or not they become teachers. Appendix B.1 provides details on variable definitions.

[7]We observe race/ethnicity for 78 percent of certification exam takers. For analyses that restrict to employed teachers, we use race/ethnicity measured directly in the TEA data.

2.3 Summary statistics. Table 1 displays demographics, certification exam performance, employment in teaching, and earnings for individuals who took the TExES exams in 2002–2021. For comparison, we also show statistics for first-year teachers, all bachelor's degree (BA) recipients in Texas, and K–12 public school students. Roughly three quarters of certification exam takers and first-year teachers are women. Hispanics make up nearly half of the K–12 student population but only 26.5 percent of first-year teachers. Black teachers are also underrepresented relative to the student population (11.6 percent vs. 13.3 percent). About 45 percent of certification exam takers are employed in Texas public schools one year after their first certification exam attempt. Teacher attrition is high, with less than half of first-year teachers still in Texas public schools 10 years later. First-year teachers earn roughly $48,000 on average (in 2019 dollars).

Historically, white students have outperformed Blacks and Hispanics on standardized tests, and this pattern also arises on teacher certification tests.[8] Among certification exam takers, Table 1 shows that, on average, URMs score nearly 0.4 standard deviations (SDs) lower than whites on high school math and ELA exams. Similarly, URMs score 16 points lower on their first TExES content exam attempt than whites on average (0.7 SDs in the distribution of certification exam takers). More importantly, URMs are 26pp less likely to pass the content exams on their first attempt and 14pp less likely to pass the PPR exams. Appendix Figure A1 shows that whites make up over 80 percent of exam takers with the highest content scores and only 20 percent of those with the lowest scores. Many candidates retake the exams, but there are still substantial racial/ethnic gaps in overall pass rates. 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 reflect discrimination or unequal resources prior to taking a teacher certification exam.

3 Costs of failing certification exams

3.1 Regression model. We begin our analysis with a regression discontinuity (RD) design that estimates the causal impacts of failing a teacher certification exam on prospective teachers' careers. Our sample includes individuals who took their first teacher certification exam in any subject/grade-level between 2003–2021. We use a local linear RD model:

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

[8] Racial/ethnic gaps in test scores appear as early as at the start of formal schooling (Jencks and Phillips, 1998; Fryer Jr and Levitt, 2006). Many reasons have been posited for these gaps, including segregation, school quality, and test content (Card and Rothstein, 2007; Penney, 2017; Bond and Lang, 2018).

Y_{ie} is an outcome for individual i who took certification exam e.[9] 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, and we weight observations using a triangular kernel following the default options in the authors' `rdrobust` package.[10] We estimate equation (1) for all certification exam takers and separately for white and URM candidates. We cluster standard errors at the individual level.

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 a variety of individual traits as dependent variables, including demographics, high school test scores, characteristics of individuals' K–12 schools, and prior year earnings. Most RD estimates are statistically insignificant, and we cannot reject the hypothesis that these coefficients are jointly equal to zero for the full sample ($p = 0.24$), white ($p = 0.74$), and URM candidates ($p = 0.52$).

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.[11] 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-

[9] 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.

[10] Except where noted below, our results are consistent across a wide range of RD bandwidths (see Appendix Figures A6–A7).

[11] 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 Effects of failing certification exams on short- and long-run outcomes. Table 2 presents RD estimates of the impacts of failing a certification exam on candidates' teaching careers, with corresponding RD graphs in Figures 1–2. We show outcomes measured one, five, and ten years after individuals first took the certification exam.[12]

Failing a certification exam both delays and dissuades entry into teaching. On average, candidates who marginally fail are 11.8pp less likely to be employed as a teacher one year later. Roughly half of individuals above the passing threshold work as teachers one year later, which means failing significantly delays entry into the profession for more than one in five individuals who marginally fail. The effect of failing on teacher employment declines over time, with RD coefficients of −4.8pp and −2.0pp measured five and ten years later. Thus while many individuals who marginally fail ultimately become teachers, a small proportion are deterred from teaching altogether. Among those who become teachers, individuals who marginally failed have roughly 0.5 fewer years of teaching experience ten years later.

The causal effect of failing on teacher employment one year later is nearly twice as large for URMs (−15.2pp) as it is for whites (−8.7pp). This gap persists over time, although the difference is smaller and statistically insignificant after ten years (−2.5pp vs. −1.7pp). Failing is more consequential for URMs for two reasons. First, while URMs and whites retake exams at similar rates, it is harder for URMs to pass conditional on retaking (Appendix Table A3). Second, URMs are more likely to work as teachers if they pass. Among individuals who marginally pass, the proportion who go into teaching is 2.8pp higher for URMs measured one year later and 9.6pp higher after ten years. This may reflect a greater interest in teaching among URM candidates. It also reflects differences in female labor supply; among certification exam takers, URMs are more likely than whites to be employed in any job in Texas (Appendix Figure A4), and this is largely driven by women (Appendix Figure A5).

Our main finding is that failing a certification exam causes a significant loss in early-

[12]Appendix Figure A3 plots RD coefficients for each year in 0–10 years since the certification exam using a balanced panel that we observe for all the years. The findings mirror those in Table 2 and Figures 1–2.

career earnings. On average, candidates who marginally fail earn $2,831 less one year after the exam. A teaching job represents a significant pay increase for the population taking their first certification exam, and thus this earnings loss reflects forgone earnings from delayed entry into the profession.[13] We find smaller RD effects on annual earnings five (−$552) and ten (−$1,595) years later; these estimates are indistinguishable from zero for some choices of the RD bandwidth (Appendix Figure A7). Although the annual earnings losses fade out as individuals retake the exams and become teachers, the cumulative loss in earnings is substantial: the present value of total earnings 10 years after the exam is $9,668 lower for candidates who marginally fail than for those who marginally passed.

The effect of failing on earnings is more negative for URM candidates than for white candidates. The RD coefficient for annual earnings one year later is twice as large for URMs (−$3,942) as it is for whites (−$1,919), consistent with the RD coefficients for teacher employment. Similarly, the loss in cumulative earnings is larger for URMs by roughly $7,300 measured over five years and $4,400 measured over ten years, although the latter difference is statistically insignificant due to a smaller sample size. Intuitively, failing causes a larger earnings loss for URMs because it leads to a longer delay in entry into teaching.[14]

3.4 Extrapolating effects to all exam takers. While our RD results show that URMs have more negative impacts of failing than whites, URM candidates are disproportionately impacted by certification exams for a more fundamental reason: they are more likely to fail. To quantify the costs of certification exams that include the likelihood of failing, we use the observed relationship between certification scores and outcomes among individuals who passed with scores between 240 and 260, and then perform a linear extrapolation to project counterfactual outcomes for all candidates with scores below the threshold.[15] Appendix Figure A8 illustrates our extrapolation method. This method relies on a strong functional form assumption for potential outcomes, but we believe it is plausible for our main outcomes

[13]Mean earnings are roughly $20,000 higher for first-year teachers than for certification exam takers in the year of their first exam (Table 1). Our RD coefficient for annual earnings (−$2,831) is approximately equal to this $20,000 pay gap multiplied by our RD coefficient for teaching employment one year later (−11.8pp).

[14]Figure 2 shows that URMs tend to have *higher* earnings than whites with the same certification scores. Appendix Table A1 shows that this pattern is due to three main factors: 1) URMs are more likely to have earnings from teaching jobs; 2) URMs have higher non-teaching earnings; and 3) conditional on teaching, URMs tend to work in school districts with higher average salaries. While labor supply decisions partly explain the URM/white gap in average earnings, they do not explain the heterogeneity in RD coefficients. Failing has only a small effect on the likelihood of employment in any Texas job (−1pp), and this effect is not statistically different between whites and URMs (Appendix Table A4). Further, our RD results are robust to including or excluding zeroes for individuals with no earnings (Appendix Table A4).

[15]We define the causal effect of failing for each exam taker as the difference between their observed and extrapolated outcomes. We use linear extrapolation to estimate causal effects below the threshold rather than the method in Angrist and Rokkanen (2015) because Angrist and Rokkanen's recommended tests of the conditional independence assumption fail given our set of observable covariates.

because many low-scoring individuals would likely have gone into teaching if they had passed and because there is limited variation in teacher salaries.

The extrapolation estimates show that certification exams impose a much larger cost on the typical URM candidate than on the typical white candidate. Figure 3 shows our extrapolation estimates for employment as a teacher and the present value of cumulative earnings (see Appendix Table A6 for all outcomes). Our RD estimates (top section of each panel) are similar to the average causal effect for exam takers with scores 1–10 points below the threshold. The effects are more negative when averaged across all failing exam takers because individuals with very low scores are less likely to retake and pass the exams. The last section of each panel shows the effects of certification exams averaged across *all* exam takers assuming a zero causal effect for those who passed. Certification exams reduce the likelihood of employment as a teacher one year later by 8.2pp for the average URM candidate as compared with 2.1pp for the average white candidate. Similarly, certification exams cost the average URM candidate roughly $12,000 in forgone earnings as compared with roughly $2,000 for the average white candidate.

Although these extrapolated estimates rely on strong assumptions, they illustrate the intuitive point that the costs of certification exams are much greater for URMs due to their lower pass rates. These disparities are notable in light of the fact that URMs are more likely to become teachers and persist in the profession conditional on passing (Figure 1).

4 Disparate impacts of certification exams

Having demonstrated that certification exams impose disproportionate costs on URMs, we now examine the relationship between certification exam performance and teaching quality. Our goal is to measure whether the exams have a disparate impact, defined as a difference in pass rates between URMs and whites of equivalent productivity. 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. To overcome this challenge, we modify the framework in Arnold et al. (2022) to focus on what we call *policy-relevant disparate impact*, which we estimate by exploiting a reform of the TExES certification exam for elementary teachers. Unlike the population-level disparate impact in Arnold et al. (2022), our measure quantifies the disparate impact experienced by the subset of individuals whose exam outcomes were affected by the TExES reform. We first describing the reform and its impacts on exam passing rates, and then we define and 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. Table 3 shows the TExES exams that led to elementary and middle school 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.[16] The EC-6 exam was harder than the EC-4 exam because it covered more material and more advanced content.[17] In contrast, the middle school exams did not change from 2002–2015; individuals could teach grades 4–8 by passing either a Generalist 4-8 exam or by passing subject-specific exams (e.g., Mathematics 4-8).

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) show 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).[18] 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 passed on subsequent attempts 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). Figure 4 presents event study versions of these results, which show that passing rates declined immediately in 2010 and remained lower in subsequent years. The TExES reform increased the total number of exam attempts but not the number of *individuals* attempting (Panel C of Table 4). The harder exam also

[16] Texas also replaced its Bilingual Generalist EC-4 and English as a Second Language (ESL)/Generalist EC-4 exams with EC-6 versions of these exams, which we include as part of the reform change.

[17] Appendix Table A8 provides examples of topics covered on the EC-4 and EC-6 exams based on CliffsNotes test prep books. The EC-4 topics are mostly a subset of those covered on the EC-6 exam.

[18] 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 4 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).

increased the mean ability (measured by high school test scores) of individuals who passed on their first attempt, although the average ability of individuals who ever passed increased only modestly (Panel D).

4.2 Policy-relevant disparate impact. This subsection defines our measure of disparate impact and our key identification assumptions (see Appendix C.1 for details). We begin by considering a population of prospective elementary school teachers characterized by their race/ethnicity, potential exam performance, and potential teaching quality. Let $R_i \in \{U, W\}$ denote the race or ethnicity 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.[19] 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 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

[19] 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.

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.

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:[20]

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

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 they 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.

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

Appendix Figure A9 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

[20] 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.

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. 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 (Appendix Table A8), so we believe it is plausible that the exam was uniformly more difficult.

Assumption 2. *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 2 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 elementary teachers. In our robustness analysis below, we test this assumption by computing placebo disparate impact estimates for middle school exams that were unaffected by the reform. We also relax Assumption 2 by using middle school teachers to control for potential time trends in pass rates or teaching quality (see Assumption 2B in Appendix Section C.2).

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 observed value-added is the same among individuals who took the easy and hard exams. While we cannot test this assumption directly because we do not observe Y_{it}^* for all exam passers, Appendix Table A9 presents balance tests using observable teacher characteristics such as demographics, test scores, and prior year earnings. Specifically, these balance tests ask whether the average gap in characteristics between EC-4 exam passers and EC-6 exam passers differs depending on whether or not we observe their value-added (i.e., whether they went on to teach subjects and grades with student testing). Consistent with Assumption 3, we find limited differences in these gaps between teachers with and without observed value-added, and we cannot reject the hypothesis that the differences in these gaps are jointly

equal to zero across all characteristics for both URM and white candidates.[21]

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), difference these across URMs and whites 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 our overall measure of disparate impact 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 URM teachers.[22] 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

[21] Another consideration is that the sample of individuals who took the EC-6 exam includes some individuals who wished to teach grade 5–6, while these individuals do not appear in the EC-4 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 slightly higher average high school test scores (although these differences are not detectable in the balance tests in Appendix Table A9) 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.

[22] 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.

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 (see Appendix B.3 for details).

4.3 Estimating disparate impact. We focus our analysis of disparate impacts on fourth grade math and ELA teachers, for two reasons. First, the EC-4/EC-6 exam reform had the most direct effects on EC–4 teachers because the exams for these grades became unambiguously more difficult. (The effects on grades 5–6 are more ambiguous.) 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 EC-4 or EC-6 exams (see Appendix Table A7). 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 EC–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 math and ELA teachers for the pre-Covid 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 5 shows that the distributions of math and ELA value-added are similar for fourth grade teachers who passed the EC-4 and EC-6 exams. 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. Similarly, the dashed lines show 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.[23] The value-added distributions overlap significantly between the two groups. In fact, the vertical lines show that the means of the EC-4 distributions are 0.01–0.03 SDs *higher* than that for the EC-6 exam, despite the fact that EC-6 exam passers are more likely to be white and have higher average ability (Appendix Table A10). This suggests that the harder EC-6 exam was not any more effective at screening out low-quality teachers.

Figure 6 shows that certification exam performance is only weakly related to value-added, particularly near the passing thresholds. Using the same samples as in Figure 5, we plot

[23]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.

fourth grade teachers' math and ELA value-added against certification exam scores separately by race/ethnicity and whether individuals took the pre- or post-reform elementary certification exams. We find positive relationships between exam performance and teacher value-added for individuals with scores above 270, consistent with the literature (e.g., Clotfelter et al., 2007). But 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 7 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 raw 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 benchmark estimates, we compute the 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 7 shows the difference 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 relative to the easier EC-4 exam. Panel A of Figure 7 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 for math value-added differences changes this pass rate only slightly to 81.3 percent. Thus our 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. Using ELA value-added as a measure of teaching quality, we similarly find a policy-relevant disparate impact estimate of 11.7pp (Panel B). In both cases, adjusting for differences in teaching quality has only a small effect on our estimates, consistent with the descriptive results that exam scores are only weakly related to value-added (Figures 5 and 6).

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 A11 and A12 provide details on all of these estimates.

We first consider the sensitivity of our results to Assumption 2, which states that prospective teachers are as good as randomly assigned to the EC-4 or EC-6 exam. For this we estimate placebo disparate impact estimates using grade 7–8 teachers. 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 went on to teach grades 7–8. 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 for fourth grade teachers. We also use grade 7–8 teachers as a control group to explicitly relax Assumption 2 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," and the estimator is a ratio-of-ratios rather than a difference-in-differences. See Appendix C.2 for details on this specification and the identification assumptions.

Figure 7 shows our placebo and ratio-of-ratios estimates, which corroborate our benchmark disparate impact results. 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. We find similar estimates using high school certification exams and ninth grade math/ELA teachers as the control group (Figure 8, Panel B, Specification I).

Figure 8 examines robustness to methodological decisions for computing disparate impact. Our primary specification discretizes the continuous value-added into deciles, so we examine robustness by grouping value-added into 20 quantiles (B), by estimating kernel densities and then grouping density values into 100 percentiles (C), and by splitting value-added into two binary groups: above/below median (D) or bottom 10 percent/top 90 percent (E). Since value-added is a *relative* measure of teacher quality, a potential concern is that our value-added estimates may be biased toward zero.[24] To address this, we estimate a leave-

[24]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

out measure of value-added where the scale for test scores is defined only by students with 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 (F). We test whether other forms of teacher quality could be driving differences in pass rates by utilizing non-cognitive value-added (index of grade retention, attendance, and suspension) following Jackson (2018) (G). Lastly, we check how our estimates change if we average values of PRDI(y, e) using the distribution of value-added for *all* teachers who passed the EC-4 exam, rather than just URM teachers (H).

Our disparate impact estimates are highly robust to these methodological choices. Across specifications, our disparate impact estimates 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 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 only weakly related to certification exam performance (Figures 5 and 6), which means our method of controlling for value-added make little substantive difference.

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. A back-of-the-envelope calculation suggests that differential pass rates on the EC-6 exam cost URM prospective teachers approximately $8.5 million in present value earnings in any given year.[25]

5 Overall effects of TExES certification reform

While Section 4 showed that certification exam performance is not strongly related to teacher value-added, changes in the difficulty of certification exams may have additional impacts on teachers and students through supply side channels. For example, schools still have to fill open teaching vacancies even if harder exams reduce the number of certified teachers.

This section develops a new strategy to identify the overall impacts of certification exam difficulty on teachers and students. We first identify a set of open teaching vacancies by

were also employed.

[25]In our data there are approximately 4,700 URMs who attempt a certification exam yearly. Given our estimate of policy-relevant disparate impact for math value-added (14.9pp), this implies that equalizing white/URM pass rates would increase the number of URMs who passed the EC-6 exam by about 700. Combined with our RD estimates on the loss in discounted earnings over 10 years for URMs ($12,200), this suggests a total earnings gain from equalizing pass rate disparities of approximately $8.5 million.

using the departure of experienced teachers as a shock to schools' labor supply. We then combine this with variation from the TExES exam reform in an RD-DiD model. Together, these two sources of variation allow us to examine how schools fill teaching vacancies under harder or easier certification exam regimes and the resulting impacts on student achievement. Importantly, this strategy circumvents the issue in related literature (and our analysis in Section 4) that teacher value-added is censored for many exam takers by directly measuring impacts of harder certification exams on student test scores.

5.1 Teacher departures. To examine the overall impacts of the TExES reform, we bring in a second source of variation: the departure of experienced teachers. 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.[26] This strategy is similar to work that uses worker deaths as a shock to firm labor supply to ask how firms find substitutes for workers (e.g., Jäger and Heining, 2022). Analogously, we use teacher departures to ask how schools fill open vacancies when easier or harder certification exams are in place.[27]

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 model to isolate sharp changes in teacher composition due to departures.[28] Specifically, we

[26]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.4.1 for details on our definition of teacher departures.

[27]Our strategy also follows Chetty et al. (2014a) in using departures to identify impacts of teacher quality.

[28]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.

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.[29] Standard errors are clustered at the school level to allow for correlation in outcomes within the same school. The coefficient of interest, β, estimates the change in average outcomes in the first year after the teacher departure ($\tau_{ty} = 0$). For example, if departing teachers are higher quality than the teachers that replace them, we would find $\beta < 0$ for the outcome of student achievement.

5.2 RD-DiD specification. Our main specification combines our teacher departure RD model with DiD variation in exposure to the TExES reform. Specifically, we estimate the RD regression (7) separately for grades that were more and less affected by the exam reform and for teacher departures that occurred before and after the reform. 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 A7). 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.[30] 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 β_{gp}.

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

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

where Treated_g is a dummy for grades 3–4 and Post_p is a dummy for teacher departures in $y \in$

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

[30] 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.

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 harder certification exams are a net benefit for student achievement, then we would find $\theta > 0$ for the outcome of student test scores.[31]

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 (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 with the TExES reform.[32] Below we test the validity of the parallel trends assumption by using student characteristics as outcome variables, Y_{st}, and using an event-study version of our RD-DiD specification.

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) shows the mean of each outcome in the year prior to the departure of a grade 3–4 teacher ($\tau_{ty} = -1$) in school/grades that experienced a departure in 2011–2016. Columns (B)–(C) show RD coefficients β from equation (7) estimated separately for grade 3–4 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–A11 show RD graphs for our main outcomes.

5.3 Exam difficulty and teacher composition. We first show that our RD-DiD approach identifies open teaching vacancies with differential exposure to the Generalist EC-4

[31] 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):

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

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.

[32] 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.

and EC-6 exams, as intended. 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 (Panel A of Table 5). 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.

Our main finding on teacher composition is that the TExES reform reduced the URM share of newly-hired elementary teachers. The RD coefficients in columns (C)–(E) of Table 5 (Panel B) show that teacher departures typically increased the URM share of teachers in a school/grade, which reflects the growing diversity of Texas' population over time. But grade 3–4 teachers who departed in the post-reform period were just as likely to be URMs as their replacements (column B). Thus the RD-DiD estimate shows that the TExES reform reduced the URM share of *employed* teachers by 1.9pp (column F), which is similar to our estimate of how the reform affected the share of URM exam *passers* in our earlier DiD analysis (Table 4, −2.6pp). This change in teacher composition was driven by a 2.5pp decline in the share of teachers who were recently-certified URMs and a 1.7pp increase in white teachers who were certified 6+ years ago. We also find a 2.4pp increase in the share of teachers who were in their first year of teaching and a decline in average teaching experience of −0.38 years. Taken together, these findings may suggest that the harder certification exams required schools to hire more white teachers who may have had breaks in their tenure (e.g., maternity leave).

The TExES reform led to a small increase in the average ability of employed teachers, consistent with the effects for exam passers (Table 4, Panel D). Specifically, the high school math scores of newly-hired elementary teachers increased by 0.084 SDs, although we find no effect on high school ELA scores (0.001 SDs) and both estimates are statistically insignificant.

5.4 Exam difficulty and student achievement. We find no systematic evidence that the TExES reform affected the number of students in school/grades with departing teachers or their demographic characteristics (Table 6, Panel A). We find small and statistically insignificant RD-DiD coefficients on a variety of student characteristics, including indices that combine all demographic traits based on their relationship with test scores, with one exception: the proportion of students at risk of dropping out ($\theta = -2.0$pp). These null findings support the validity of our key assumption of parallel trends in RD coefficients.

Our main finding from the RD-DiD analysis is that the harder elementary certification exams neither increased nor decreased student test scores (Table 6, Panel B). Our outcome variables include math and ELA scores in SD units as well as test score residuals from value-added-like regressions that include individual, school/grade mean, and school mean student characteristics and lagged test scores as covariates. The RD coefficients in columns (B)–(E)

show that, on average, teacher departures have mostly small and insignificant effects on student achievement. Our RD-DiD estimates in column (F) 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). All our RD-DiD coefficients are statistically insignificant, and they are small relative estimates from the teacher quality literature, which often finds that the standard deviation of teacher value-added is 0.1–0.2 SDs (e.g., Chetty et al., 2014a).

We do not find systematic patterns of heterogeneity in the RD-DiD estimates for student test scores by race and ethnicity (Appendix Table A13). Although research finds that URM students benefit from having URM teachers (e.g., Dee, 2004), our estimate of the effect of the TExES reform on the URM share of teachers (−1.9pp) is not large enough to yield significant heterogeneity in student achievement.

5.5 Robustness. Our finding that the TExES reform did not improve student achievement is robust to a variety of other specifications. A caveat with our RD-DiD approach is that it only estimates effects at the RD threshold, i.e., in the first year after a teacher departure ($\tau_{ty} = 0$). Appendix Table A14 modifies equation (7) so that it 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 some student demographics, which is why we prefer the RD model that estimates effects at $\tau_{ty} = 0$. In our main RD-DiD specification, our point estimates for math and ELA scores are relatively stable and statistically insignificant across a wide range of RD bandwidths (Appendix Figure A12). We find little evidence of pre-trends in our main outcomes of interest in an event-study version of our RD-DiD specification (Appendix Figure A13). Appendix Table A15 presents results from an alternative RD-DiD specification that restricts the sample to the subset of grade 3–4 teachers who taught *only* math or *only* ELA and uses two different control groups: 1) test scores in the subject *not* taught by the departing teaching; and 2) schools where a grade 3–4 teacher departed in the *other* subject. We find no effects on ELA scores in these specifications, and some evidence that the TExES reform *reduced* math scores.

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

6 Conclusion

This paper presented new evidence on racial/ethnic disparities in the economic costs of teacher certification exams. Using administrative data from Texas, we found that individuals who marginally fail certification exams are delayed in starting their teaching careers, costing them roughly $10,000 in forgone earnings. URM teaching candidates are more likely to bear these costs given their lower pass rates, and they also have larger earnings losses if they fail.

We also showed that these disproportionate costs are not justified by differences in teaching quality between white and URM candidates. We developed a new measure of disparate impact for certification exams and estimated it using a reform that increased the difficulty of Texas' exam for elementary teachers. We found that the increase in exam difficulty had a disparate impact in the sense that URMs who would have passed the easier exam were 10–15pp less likely to pass the hard exam than white candidates with comparable teaching quality. A complementarity analysis of the net impacts of the reform confirmed that the harder exam reduced the URM share of teachers but did not raise student achievement.

Our findings suggest that policies that increase certification exam pass rates can reduce costs on URM candidates and help diversify the teaching workforce without harming student achievement. Multiple states have adopted such policies in recent years. For example, Mississippi lowered passing thresholds on some teacher certification exams (Amy, 2018), while Alabama, Delaware, and Missouri now allow candidates who marginally fail the exams to become certified anyway if they meet GPA or other academic requirements (Swisher, 2022). Although such policies are often motivated by teacher shortages, our paper highlights their potential to promote racial and ethnic equity without compromising student outcomes.

Our results on student achievement are based on the elementary certification exams in Texas, and so it is an open question whether exams are more informative for teaching quality in other states or at higher grade levels. Similarly, our results do not necessarily generalize to policies that eliminate certification exams altogether, which may have a greater impact on the composition of individuals who pursue teaching careers. But the tools developed in this paper can help researchers exploit this type of policy variation to study the costs and benefits of teacher certification policies in other contexts.

More broadly, standardized exams are widely used to determine certification and licensing in other professions such as medicine, law, and accounting. These exams also exhibit large disparities in pass rates between white and URM exam takers, but there is little evidence on their validity and the economic consequences for candidates who fail. We hope future research will shed light on the potential disparate impacts of licensing exams in other settings.

References

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.

Amy, J. (2018). Mississippi cuts math teacher test score, citing shortage. *Associated Press*. Accessed in January 2026 at https://apnews.com/general-news-35610a0e283b4590a8a5d535b24def6b.

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.

Angrist, J. D. and Rokkanen, M. (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344.

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

Arrow, K. (1971). The theory of discrimination. Industrial Relations Section, Princeton University, Working Paper No. 30A.

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.

Becker, G. S. (1959). *The Economics of Discrimination*. University of Chicago Press.

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. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6):673–682.

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)*.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, 86(1):195–210.

Farronato, C., Fradkin, A., Larsen, B. J., and Brynjolfsson, E. (2024). Consumer protection in an online world: An analysis of occupational licensing. *American Economic Journal: Applied Economics*, 16:549–579.

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.

Gittleman, M., Klee, M. A., and Kleiner, M. M. (2018). Analyzing the labor market outcomes of occupational licensing. *Industrial Relations: A Journal of Economy and Society*, 57(1):57–100.

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.

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. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. National Bureau of Economic Research.

Kleiner, M. and Xu, M. (2024). Occupational licensing and labor market fluidity. *Journal of Labor Economics*.

Kleiner, M. M. and Krueger, A. B. (2013). Analyzing the extent and influence of occupational licensing on the labor market. *Journal of Labor Economics*, 31.

Kleiner, M. M. and Kudrle, R. T. (2000). Does regulation affect economic outcomes? the case of dentistry. *Journal of Law and Economics*, 43:547–582.

Kleiner, M. M., Marier, A., Park, K. W., and Wing, C. (2016). Relaxing occupational licensing requirements: Analyzing wages and prices for a medical service. *The Journal of Law and Economics*, 59(2):261–291.

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.

Rodman, B. (1987). Arkansas union drops suit over teacher tests. *Education Week*. Accessed in January 2026 at https://www.edweek.org/education/arkansas-union-drops-suit-over-teacher-tests/1987/11.

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.

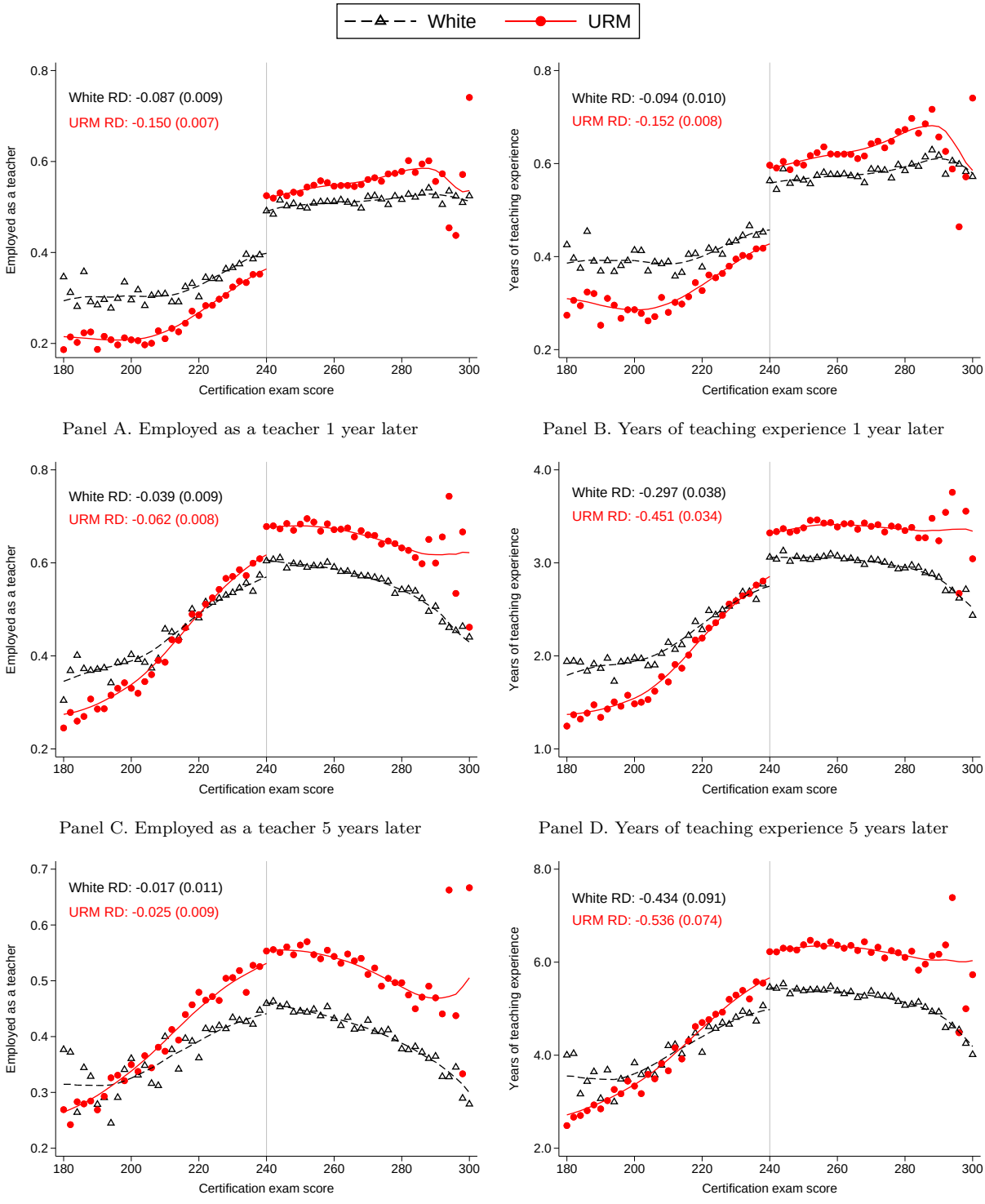
Swisher, A. (2022). Setting sights lower: States back away from elementary teacher licensure tests. Accessed in January 2026 at https://www.nctq.org/research-insights/setting-sights-lower-states-back-away-from-elementary-teacher-licensure-tests.

Thornton, R. J. and Timmons, E. J. (2013). Licensing one of the world's oldest professions: Massage. *The Journal of Law and Economics*, 56(2):371–388.

Tsao, C. (2025). It's not (just) about the money: Pay and the value of working conditions in teaching. Working Paper.

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

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

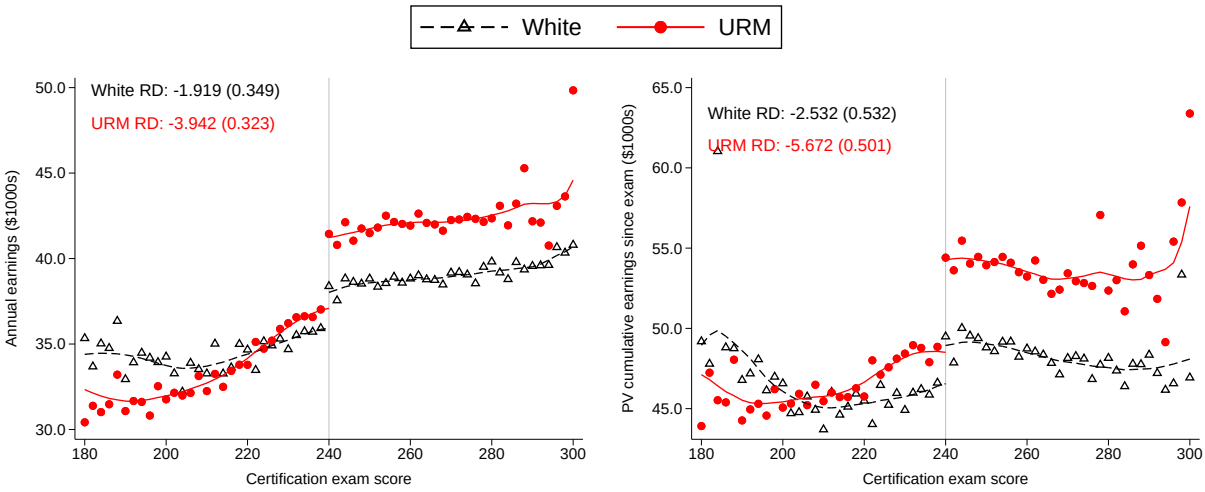
Panel D. Years of teaching experience 5 years later

Panel E. Employed as a teacher 10 years later

Panel F. Years of teaching experience 10 years later

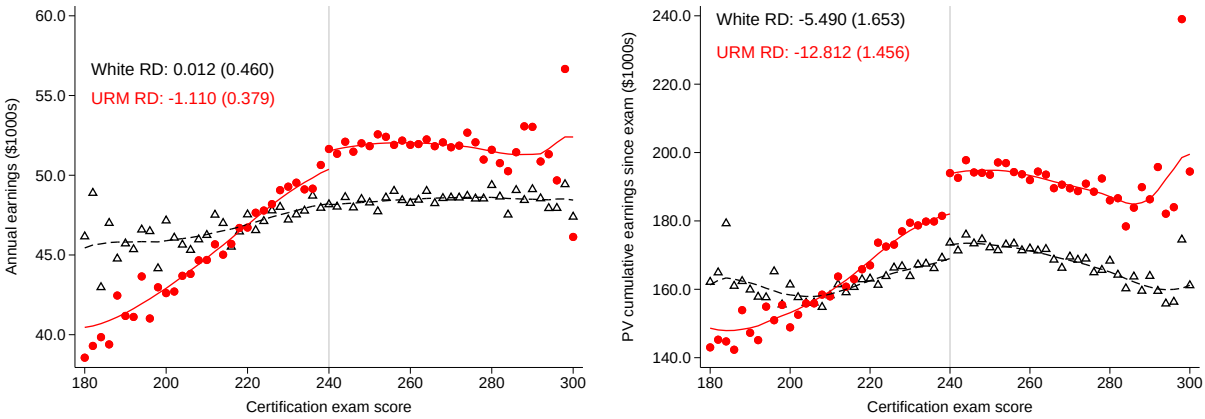
Figure 1: RD plots: Teaching outcomes

Notes: This figure presents RD effects of failing a certification exam on teaching outcomes for white and URM candidates. The x-axis of each panel is the certification exam score with a vertical line at the passing threshold (240). The outcomes are teacher employment and years of teaching experience measured one, five, and ten years after the exam, as indicated by the panel title. The upper-left corner of each panel shows RD coefficients with standard errors clustered at the individual level in parentheses.



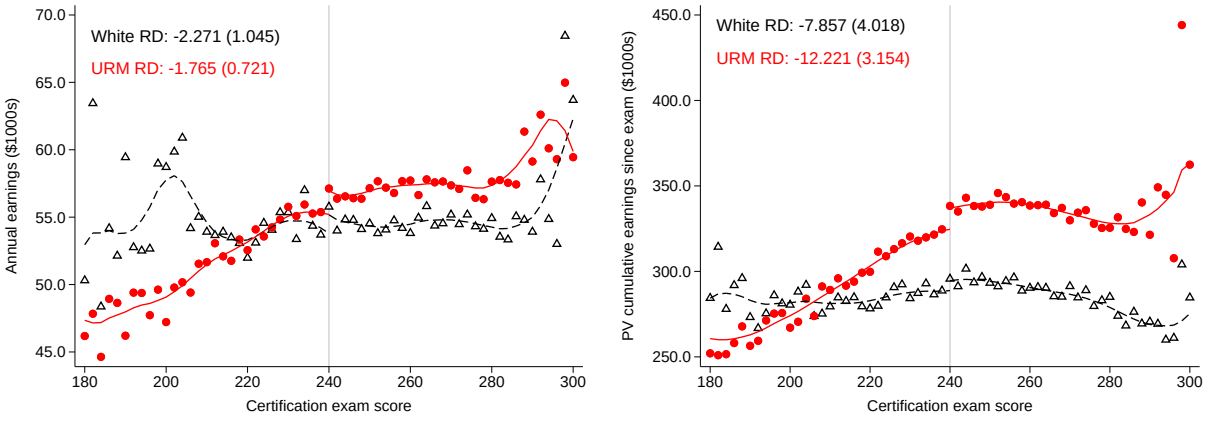
Panel A. Annual earnings 1 year later ($1000s)

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)

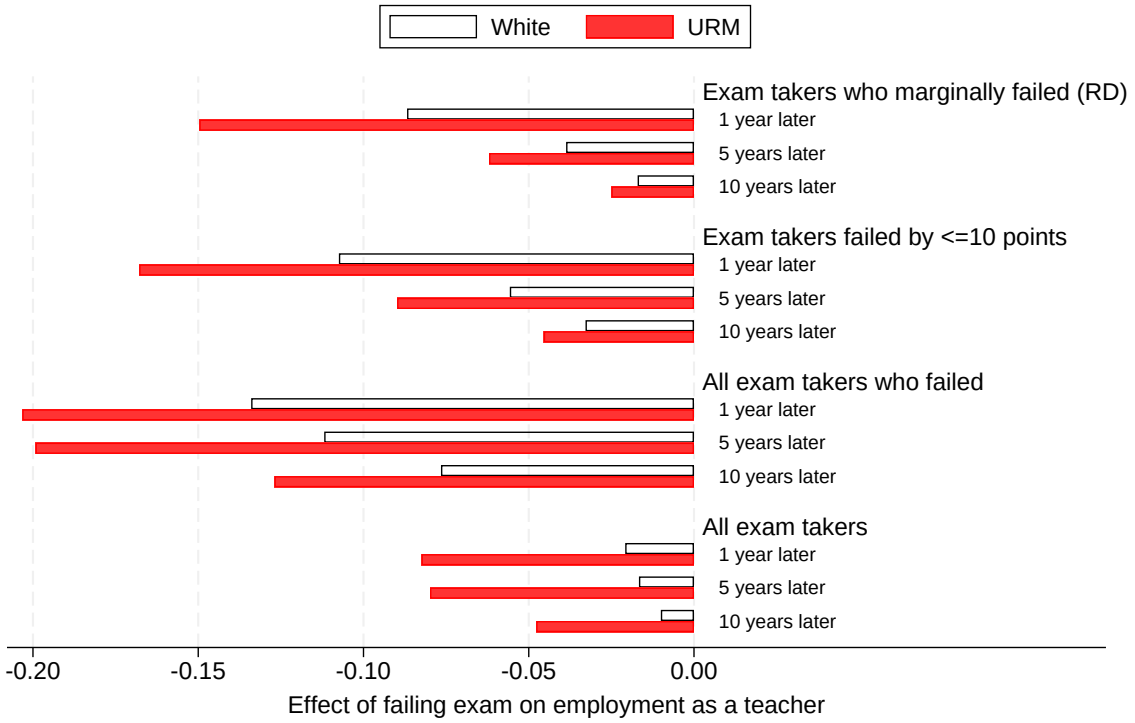


Panel E. Annual earnings 10 year later ($1000s)

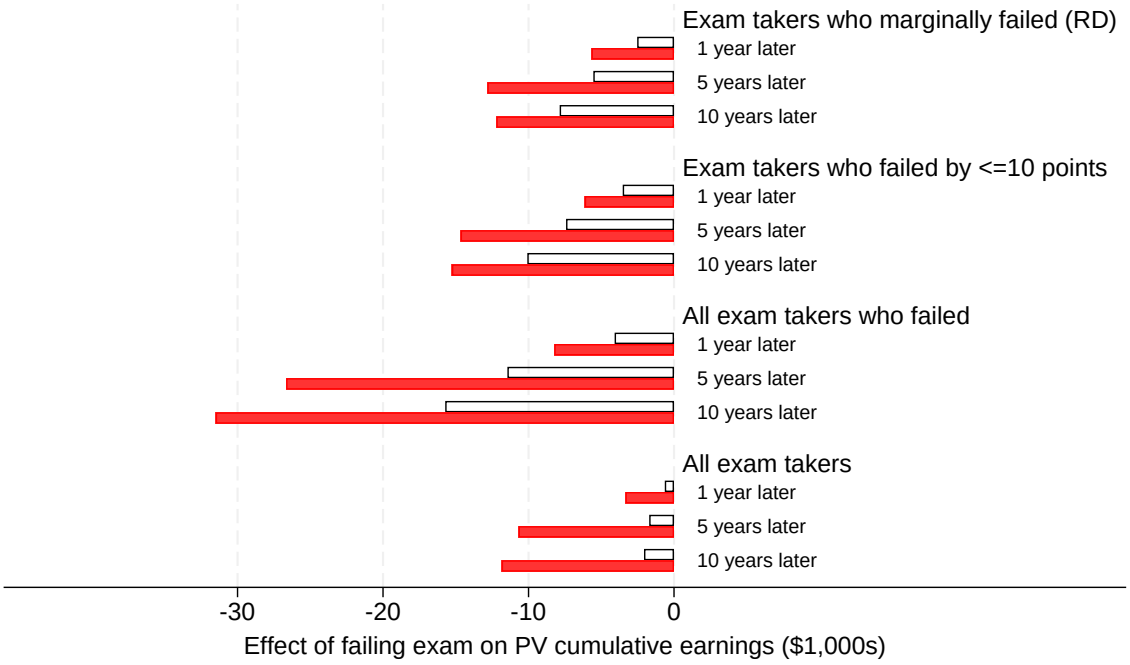
Panel F. PV cum. earnings after 10 years ($1000s)

Figure 2: RD plots: Earnings outcomes

Notes: This figure presents RD effects of failing a certification exam on earnings outcomes for white and URM candidates. The x-axis of each panel is the certification exam score with a vertical line at the passing threshold (240). The outcomes are annual earnings and the present value of cumulative earnings measured one, five, and ten years after the exam, as indicated by the panel title. The upper-left corner of each panel shows RD coefficients with standard errors clustered at the individual level in parentheses.



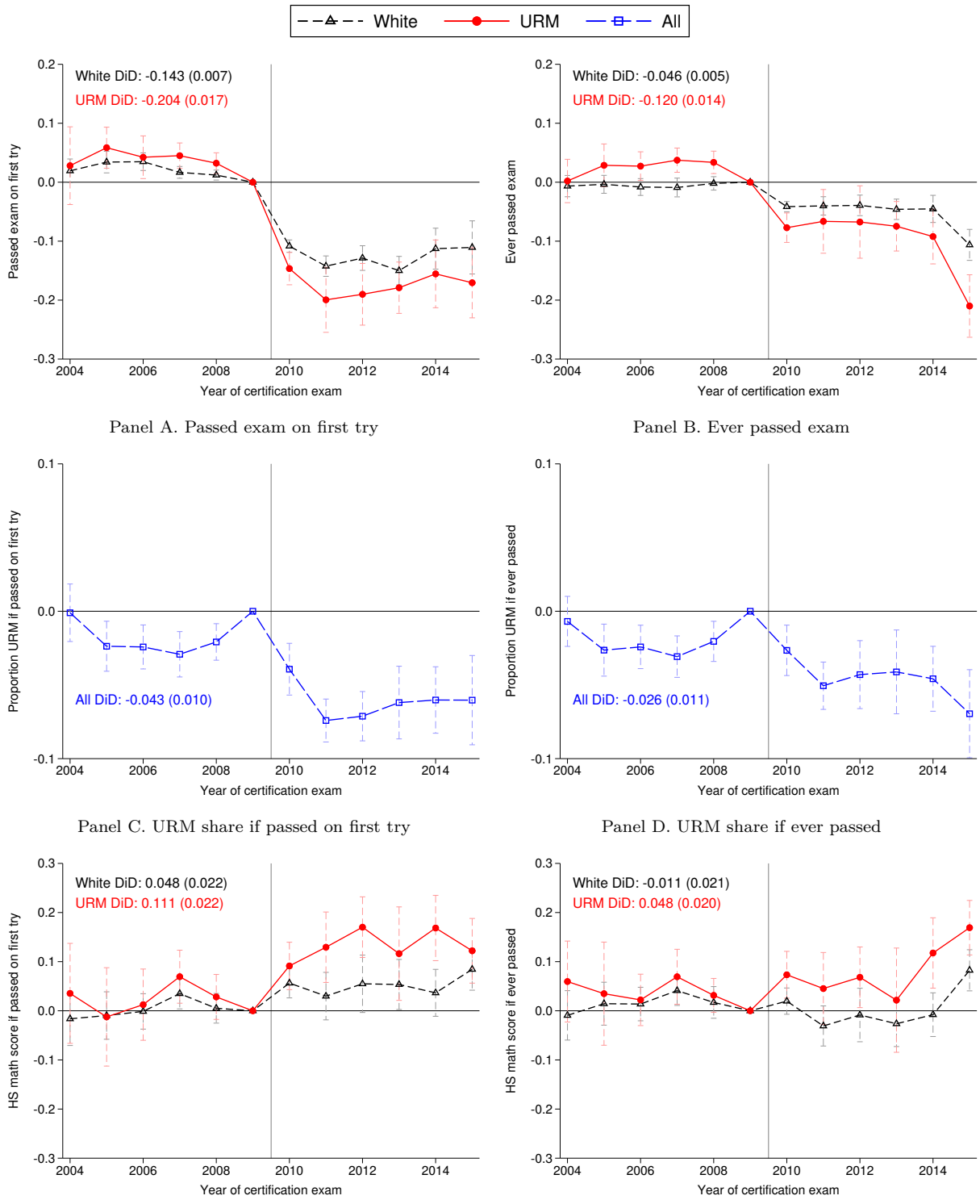
Panel A. Employed as a teacher



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

Figure 3: Extrapolated effects of failing certification exam on outcomes

Notes: This figure summarizes the baseline RD effects and extrapolated estimates in three samples: exam takers with a score within ten points of the passing score, exam takers with a score below the 240 cutoff, and all exam takers. The outcome variables in Panels A and B are teaching employment and the present value of cumulative earnings, respectively. In each case we show separate estimates for white and URM candidates measured one, five, and ten years after the exam. See Section 3.4, Appendix Figure A8, and Appendix Table A6 for details on our extrapolation method.



Panel A. Passed exam on first try

Panel B. Ever passed exam

Panel C. URM share if passed on first try

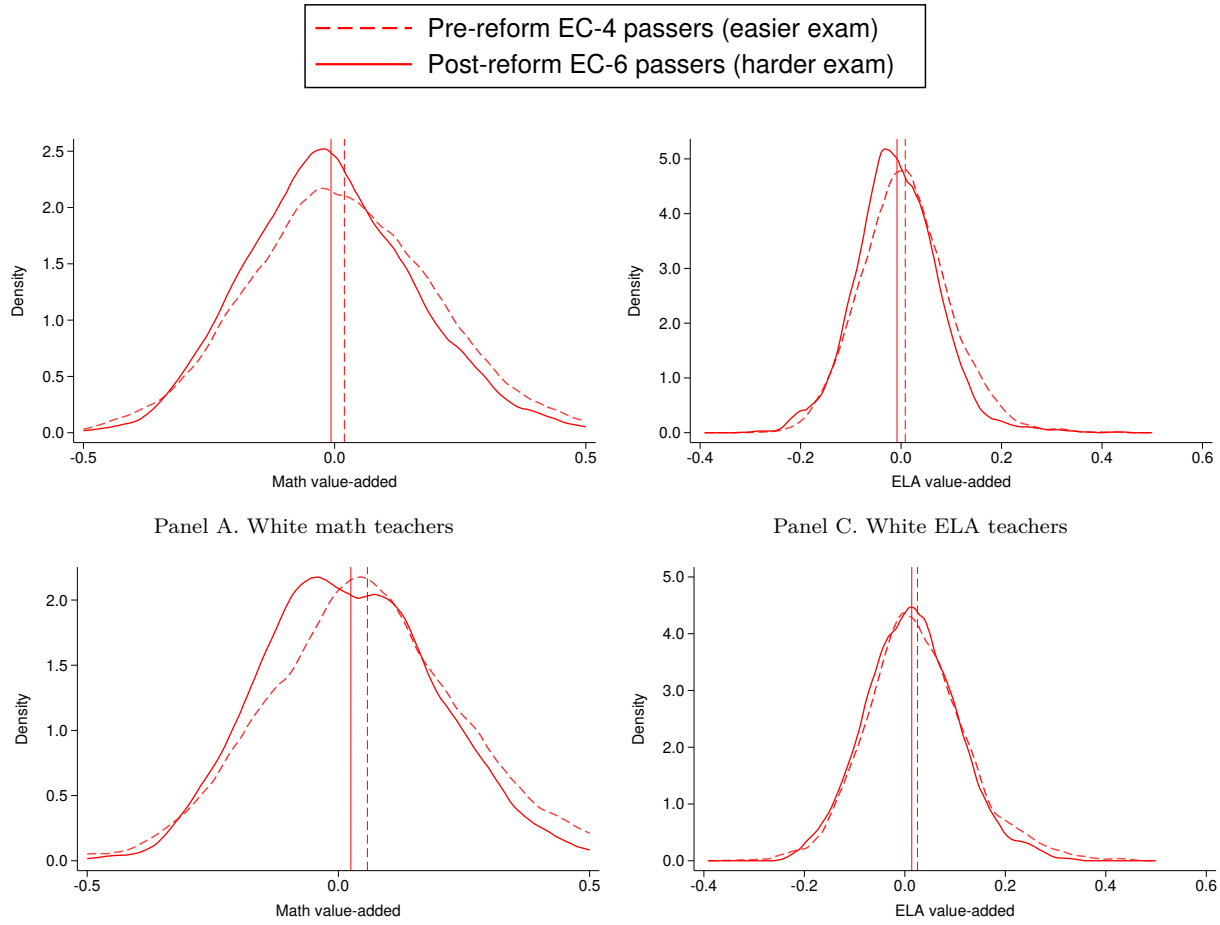
Panel D. URM share if ever passed

Panel E. HS math score if passed on first try

Panel F. HS math score if ever passed

Figure 4: 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

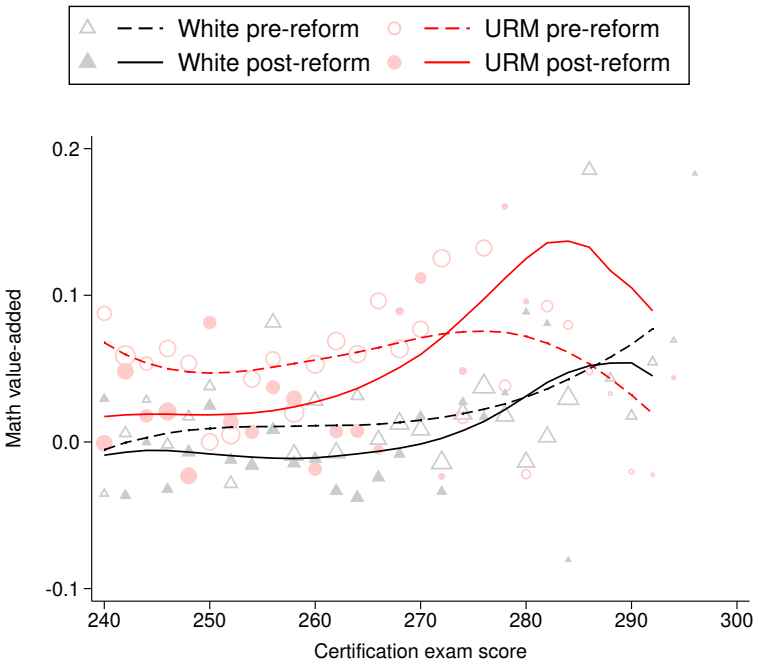
Panel C. White ELA teachers

Panel B. URM math teachers

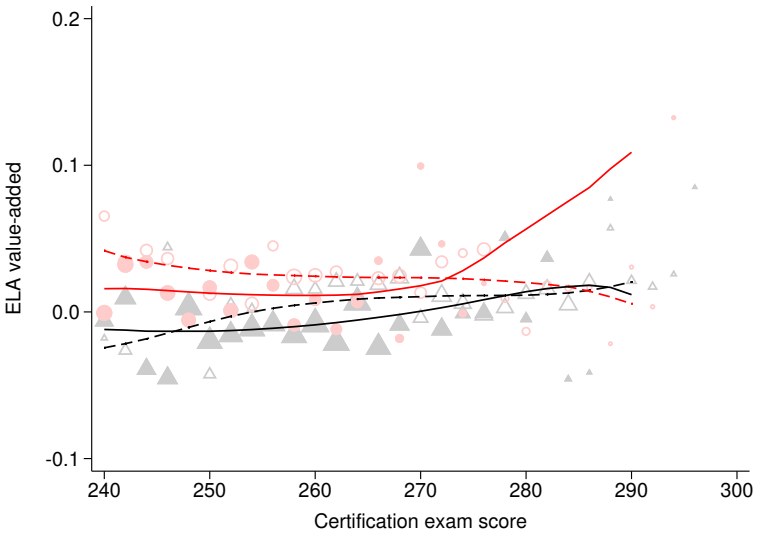
Panel D. URM ELA teachers

Figure 5: 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 teachers 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 teachers 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.



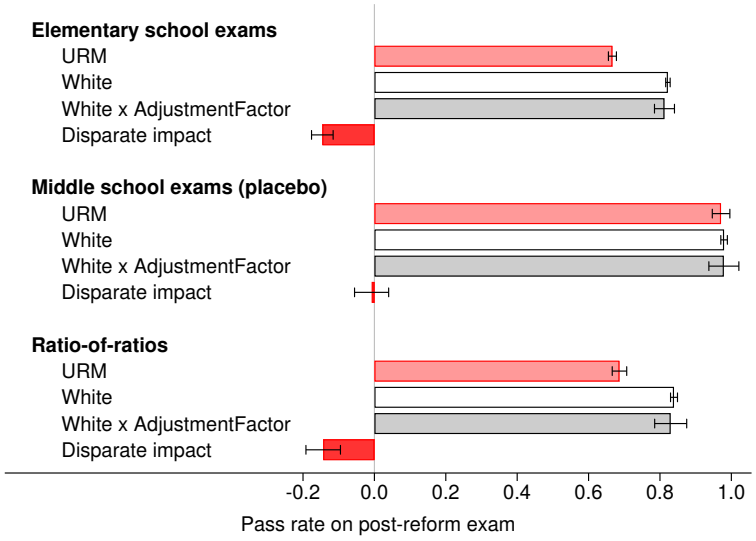
Panel A. Math value-added



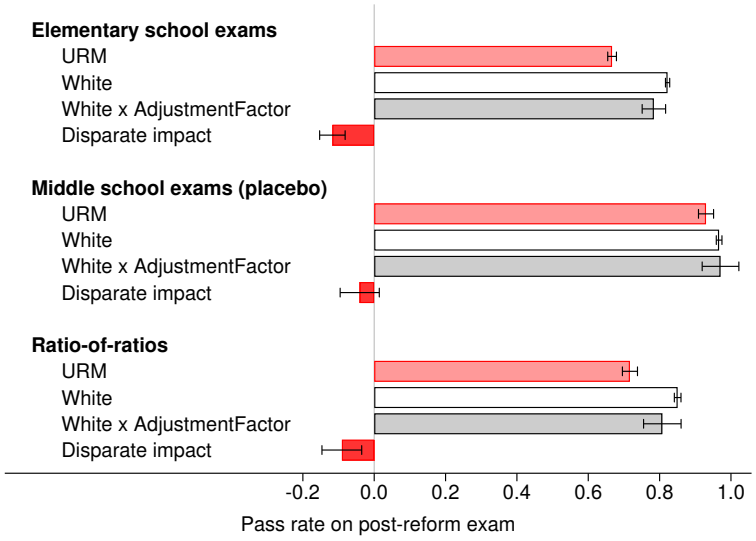
Panel B. ELA value-added

Figure 6: 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 teachers 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.



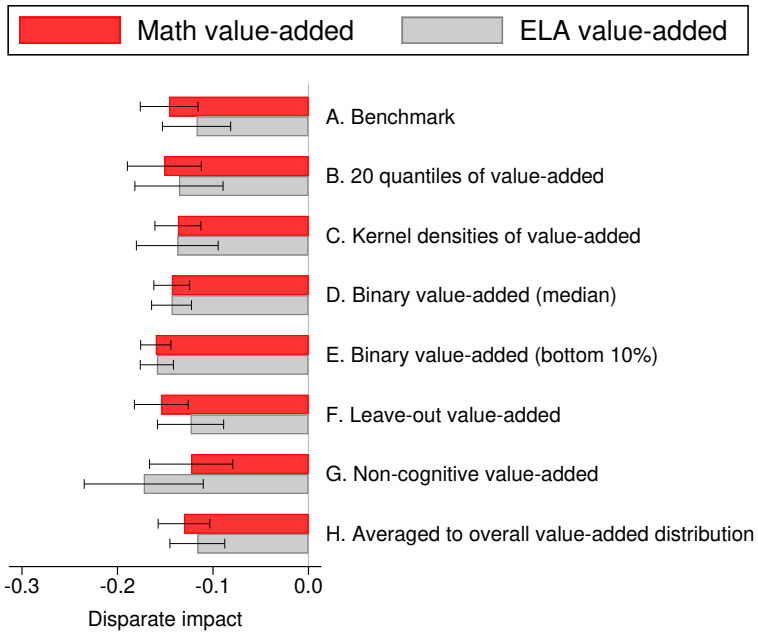
Panel A. Math value-added



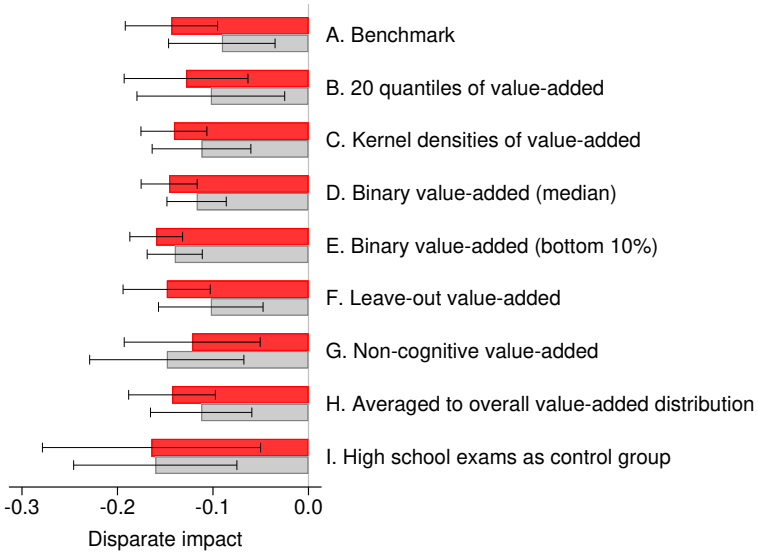
Panel B. ELA value-added

Figure 7: 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 A11 and A12 for details on these computations.



Panel A. Elementary school exams only



Panel B. Ratio-of-ratios

Figure 8: 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 suspension. 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 and Appendix Tables A11–A12 for details on these specifications and computations.

Tables

Table 1: Summary statistics for TExES exam takers and comparison populations in 2002–2021

	(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	0.776	0.743	0.762	0.747	0.581	0.487
White	1.000	0.000	0.582	0.582	0.506	0.320
Black	0.000	0.303	0.119	0.116	0.095	0.133
Hispanic	0.000	0.697	0.274	0.265	0.278	0.493
HS math score	0.619	0.243	0.473	0.530	0.668	0.043
HS ELA score	0.697	0.307	0.542	0.591	0.633	0.039
Panel B. TExES content exam performance						
Content exam score on first try	257.7	241.5	252.1	253.6		
Passed content exam on first try	0.829	0.573	0.736	0.773		
Ever passed first content exam	0.925	0.771	0.863	0.922		
Ever passed any content exam	0.958	0.850	0.911	0.976		
Panel C. TExES PPR exam performance						
PPR exam score on first try	266.4	254.6	262.4	263.0		
Passed PPR exam on first try	0.944	0.807	0.893	0.905		
Ever passed first PPR exam	0.991	0.953	0.975	0.986		
Ever passed any PPR exam	0.993	0.963	0.981	0.991		
Panel D. Teaching in public schools						
Taught in year of exam/first teaching/graduation	0.163	0.128	0.189	0.975	0.001	
Taught 1 year later	0.432	0.375	0.447	0.862	0.082	
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	26,445	26,805	29,044	47,886	24,667	
Earnings 1 year later	37,246	37,502	39,316	48,562	38,529	
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)	314,972	212,711	716,002	388,437	1,702,097	14,331,974

Notes: This table presents the summary statistics for first-time TExES exam takers by race/ethnicity (columns A–C), first-year teachers (column D), bachelor degree recipients (column E), and K–12 students (column F) in 2002–2021. Panel A displays demographic variables, including gender, race, and high school math and ELA test scores. Panel B presents TExES content exam scores and passing rates. Panel C reports the TExES PPR exam scores and passing rates. Panel D reports the fraction of individuals teaching in Texas public schools in the exam/first teaching/graduation year as well as one, five, and ten years later. Panel E shows annual earnings in 2019 US dollars in the exam/first teaching/graduation year as well as one, five, and ten years later.

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: This table presents RD estimates of the impacts of failing a certification exam on teaching and earning outcomes. Panels A–C show outcomes measured one, five, and ten years after the exam, respectively. The dependent variables are an indicator for being employed as a teacher, years of teaching experience, annual earnings (in thousands of 2019 dollars), and the present value of cumulative earnings since the year of the exam (in thousands of 2019 dollars). Columns (A) and (B) display dependent variable means for candidates with exam scores 0–10 points above the threshold. Columns (C)–(E) show RD coefficients β from equation (1) estimated separately for all, white, and URM candidates. Column (F) shows the p-value from a test of equality of the RD coefficients for white and URM candidates. "N (Employed as a teacher)" indicates the sample size for the outcome of employment as a teacher. 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						
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)
Years of teaching experience	11.424	-3.093***	-3.426***	-3.690***	-4.406***	-0.383*
		(0.081)	(0.098)	(0.101)	(0.127)	(0.207)
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)
N (# *sty* observations)	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) and demographics (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 * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

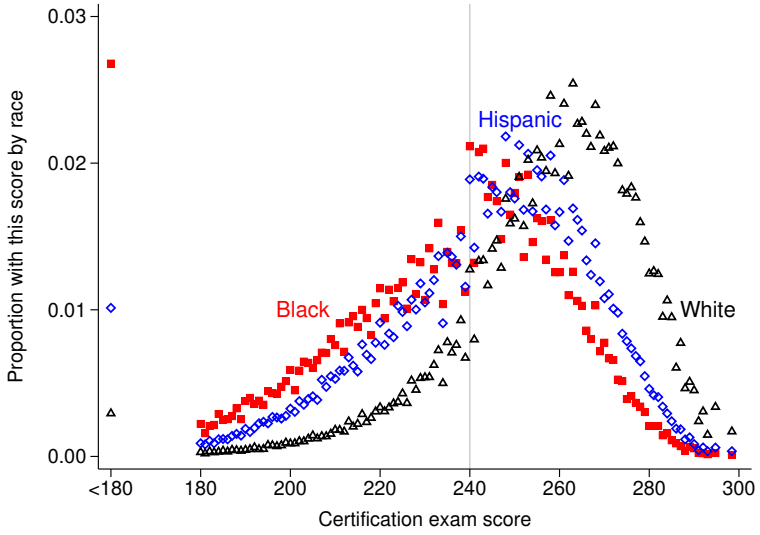
Table 6: RD-DiD effects of harder certification exams on student achievement

	(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 (# *sty* observations)	4,887	105,829	86,102	123,706	107,812	423,449

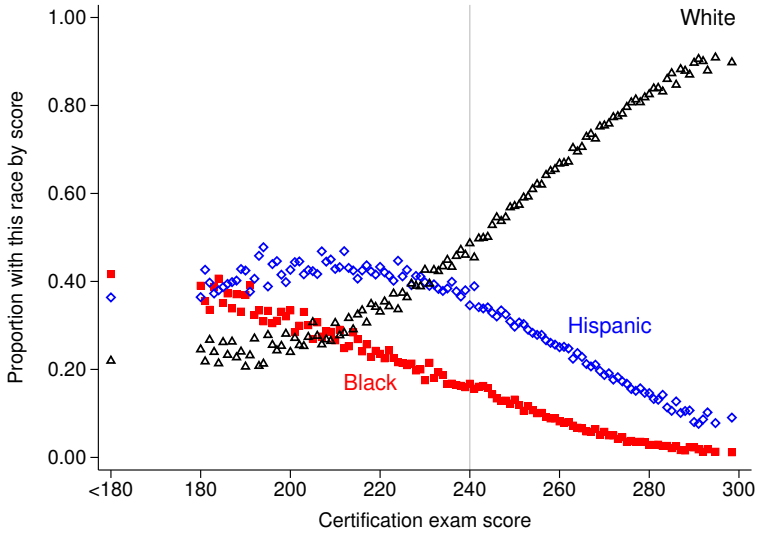
Notes: This table displays RD and RD-DiD estimates of the effects of the TExES reform on student composition (Panel A) and student achievement (Panel B). Column (A) shows the mean of each outcome in the year prior to the teacher departure ($\tau_{ty} = -1$) in school/grades that experienced a departure 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 * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix — For Online Publication Only

A Appendix Figures and Tables



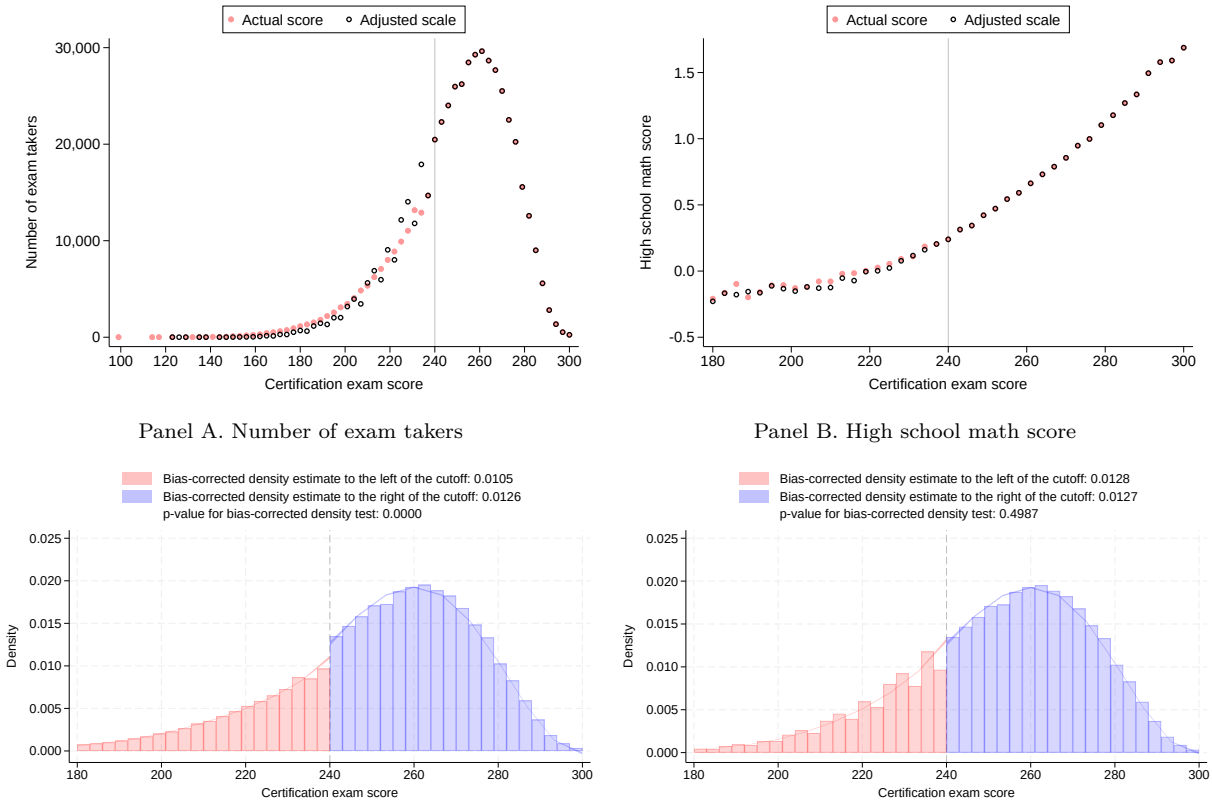
Panel A. Score distribution for each racial group



Panel B. Racial distribution for each score value

Figure A1: First-attempt certification exam scores

Notes: This figure shows the distribution of first-attempt certification exam scores by race for individuals taking their first TExES teacher certification exam in 2002–2021. 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

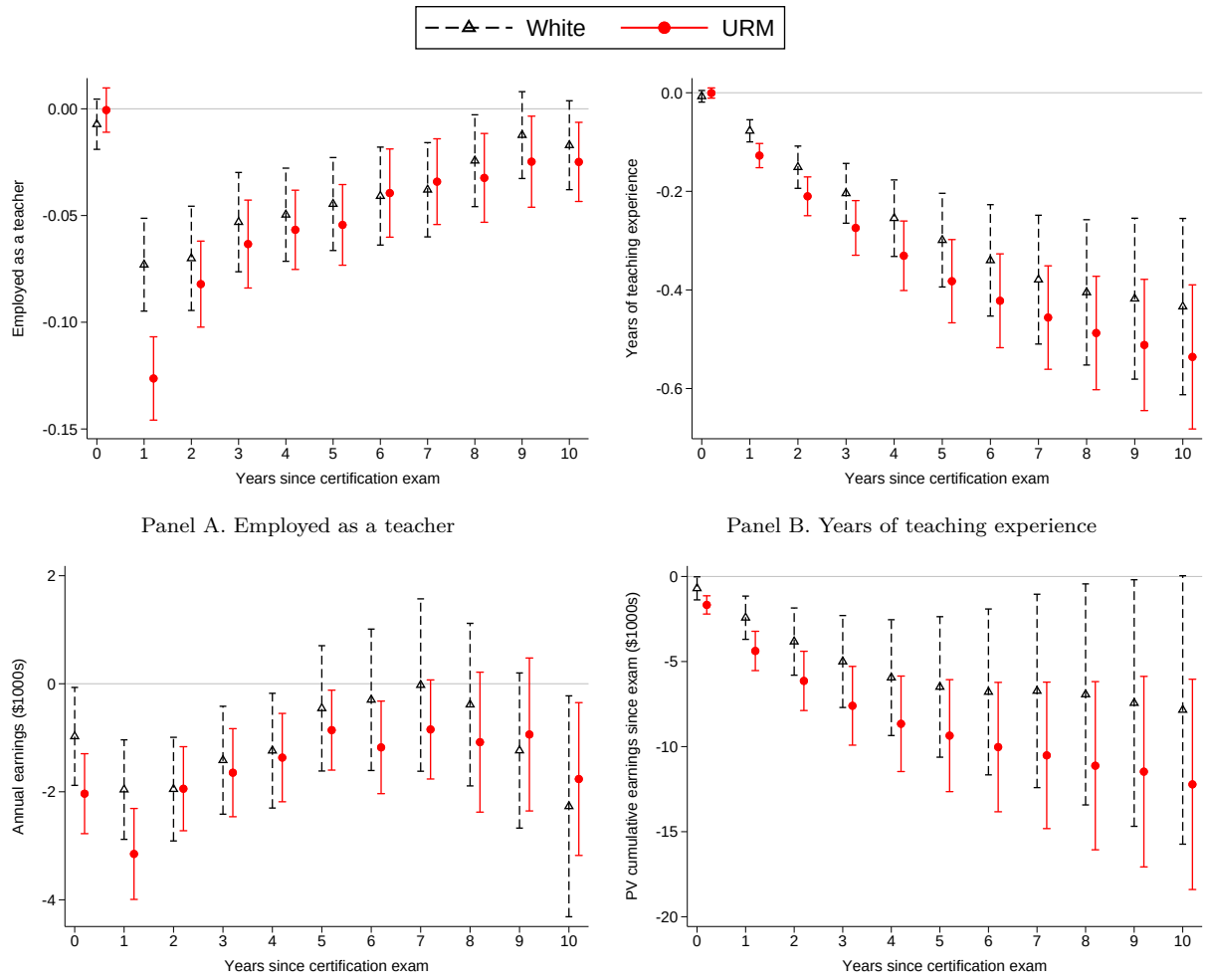
Panel D. rddensity test: Adjusted scale

Figure A2: Density tests

Notes: This 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, and Panel B illustrates the relationship between high school math scores and certification exam scores. Both panels depict actual scores (solid circles) and adjusted scores (hollow circles). Adjusted scores are rescaled below the passing threshold (240) so that they have the same linear relationship with high school math scores as passing scores in the vicinity of the threshold. Specifically, we run linear regressions of high school math scores on certification scores using certification scores both slightly above the threshold (240–244) and slightly below the threshold (235–239). This gives an above threshold OLS coefficient β_a and a below threshold OLS coefficient β_b. We then compute adjusted scores for each exam taker i using the formula:

$$\text{AdjustedScore}_i = \text{ActualScore}_i \text{ if } \text{ActualScore}_i \geq 240$$
$$\text{AdjustedScore}_i = 240 + (\text{ActualScore}_i - 240) \times \beta_b/\beta_a \text{ if } \text{ActualScore}_i < 240.$$

Panels C and D display regression discontinuity density tests using the `rddensity` package of Cattaneo et al. (2020), based on the actual scores (Panel C) and adjusted scores (Panel D).



Panel A. Employed as a teacher

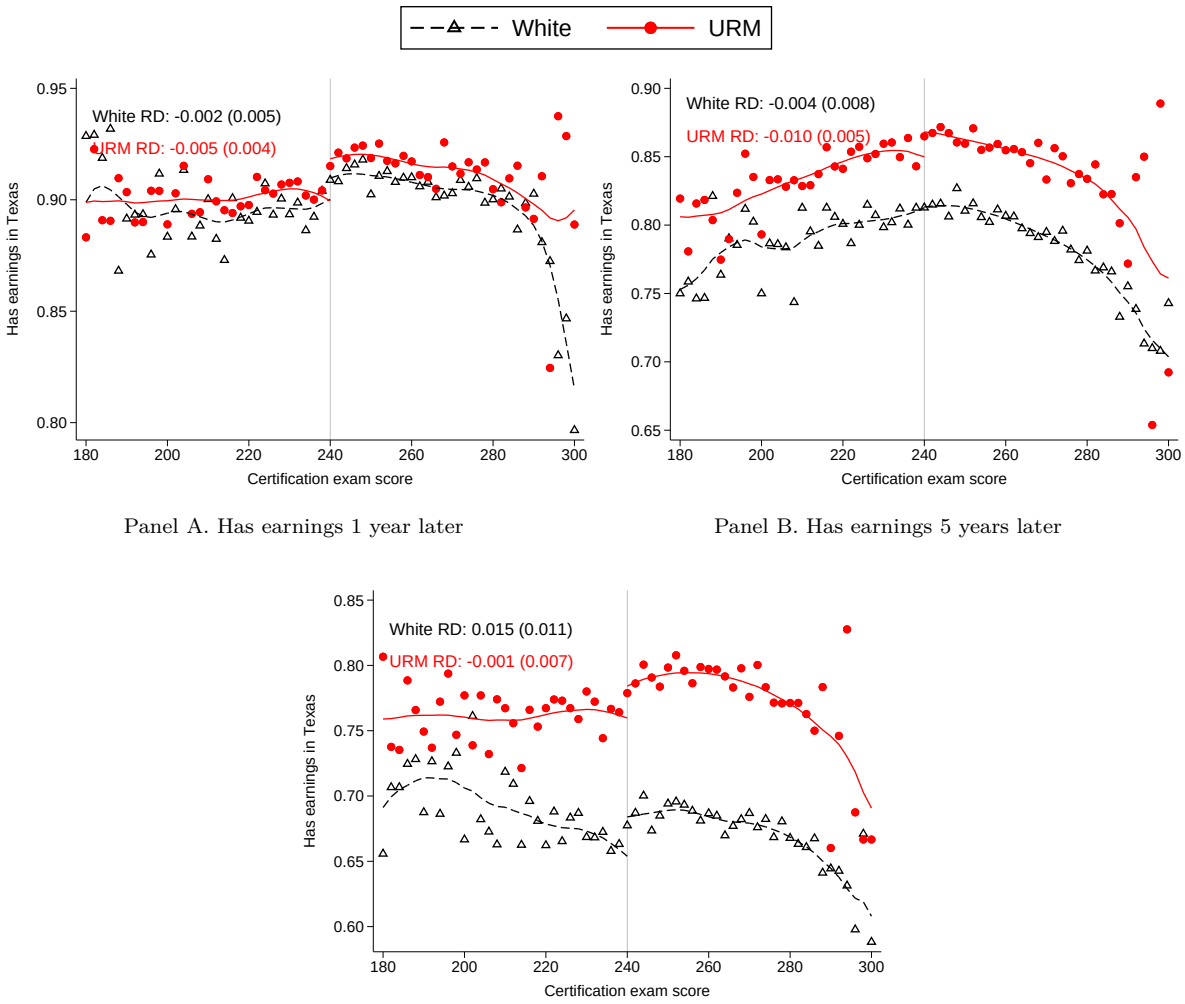
Panel B. Years of teaching experience

Panel C. Annual earnings ($1000s)

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

Figure A3: RD effects of failing a certification exam over time — Balanced panel

Notes: This figure displays time variation in the RD effects of failing a certification exam on teaching and earnings outcomes. Each panel plots RD coefficients β from equation (1) estimated separately for white and URM candidates and for each year from 0–10 years since taking the certification exam. We use a balanced panel of exam takers that we can observe in every year given the timing of our data, i.e., individuals who took their first TExES certification exam in 2003–2012. The dependent variables are an indicator for being employed as a teacher, years of teaching experience, annual earnings (in thousands of 2019 dollars), and the present value of cumulative earnings since the year of the exam (in thousands of 2019 dollars), as indicated in the panel titles.



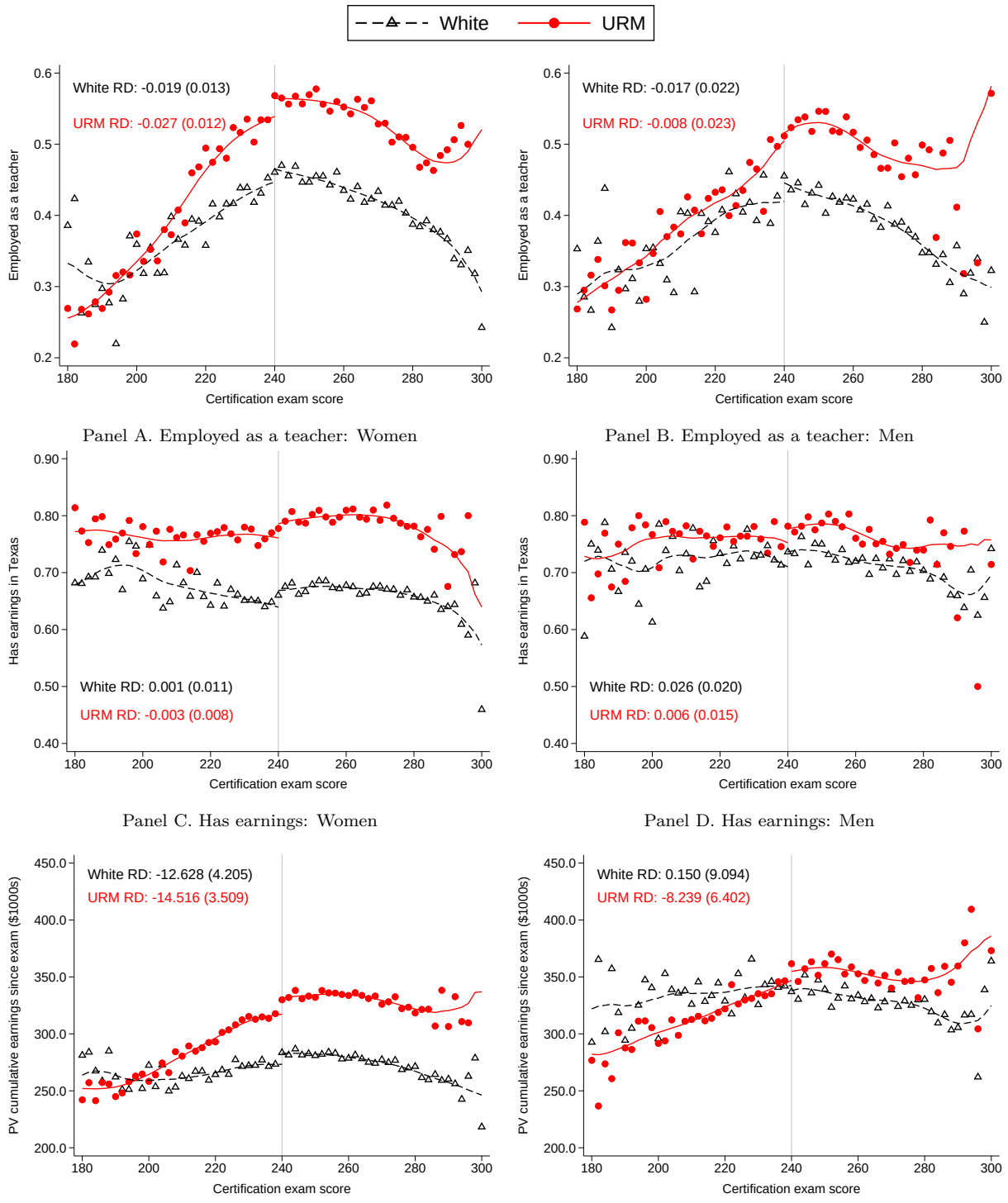
Panel A. Has earnings 1 year later

Panel B. Has earnings 5 years later

Panel C. Has earnings 10 years later

Figure A4: RD plots: Has earnings in Texas

Notes: This figure presents RD effects of failing a certification exam on the likelihood of appearing in the TWC earnings data for white and URM candidates. The x-axis of each panel is the certification exam score with a vertical line at the passing threshold (240). The outcome is an indicator for appearing the TWC data measured one, five, and ten years after the exam, as indicated by the panel title. The upper-left corner of each panel shows RD coefficients with standard errors clustered at the individual level in parentheses.



Panel A. Employed as a teacher: Women

Panel B. Employed as a teacher: Men

Panel C. Has earnings: Women

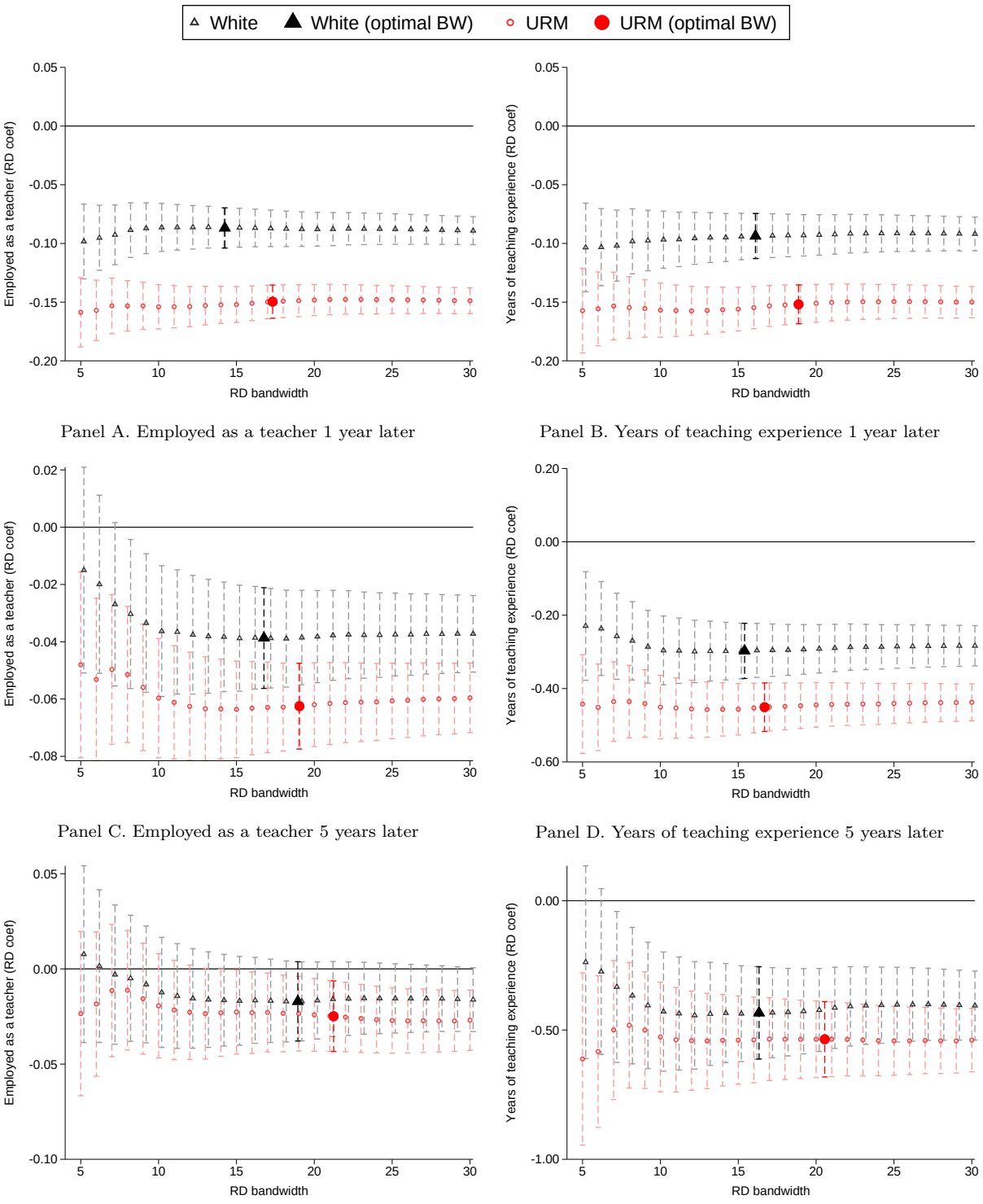
Panel D. Has earnings: Men

Panel E. PV cumulative earnings: Women

Panel F. PV cumulative earnings: Men

Figure A5: RD plots: Teaching and labor market outcomes by gender

Notes: This figure presents RD effects of failing a certification exam on teaching and labor market outcomes separately by gender and for white and URM candidates. The x-axis of each panel is the certification exam score with a vertical line at the passing threshold (240). The outcomes are teacher employment, an indicator for appearing the TWC data, and the present value of cumulative earnings measured one, five, and ten years after the exam, as indicated by the panel title. Panels A, C, and E show results for female candidates, while Panels B, D, and F show results for male candidates. The upper-left corner of each panel shows RD coefficients with standard errors clustered at the individual level in 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

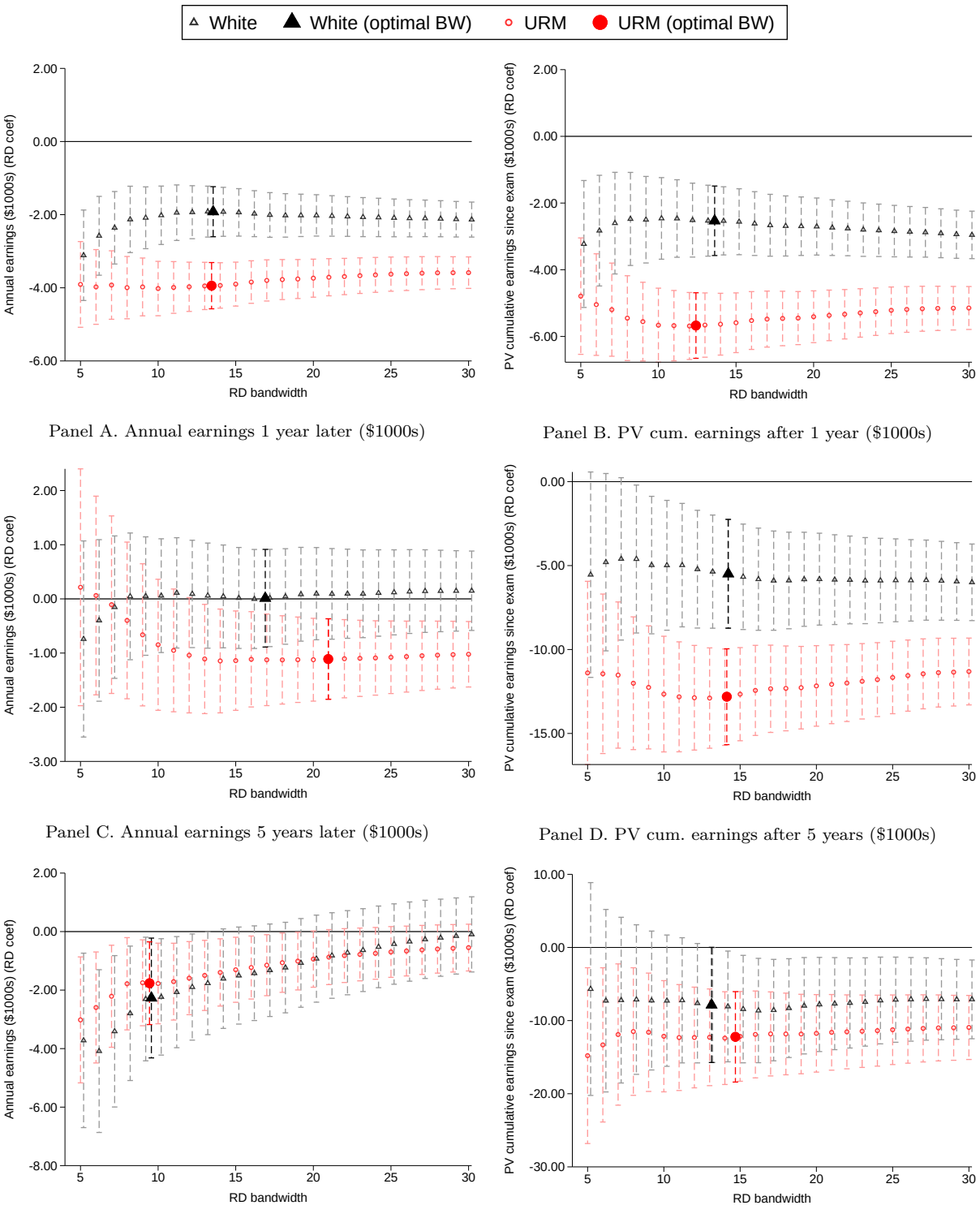
Panel D. Years of teaching experience 5 years later

Panel E. Employed as a teacher 10 years later

Panel F. Years of teaching experience 10 years later

Figure A6: Robustness to RD bandwidth: Teaching outcomes by race

Notes: This figure presents RD effects of failing a certification exam on teaching outcomes by race/ethnicity using different RD bandwidths (x-axis). The outcomes are teacher employment and years of teaching experience measured one, five, and ten years after the exam, as indicated by the panel title. Hollow markers depict RD coefficients computed using integer bandwidths from 5–30 certification score points. The large solid markers show our benchmark estimates from Table 2 using the Calonico et al. (2019) bandwidth. 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)

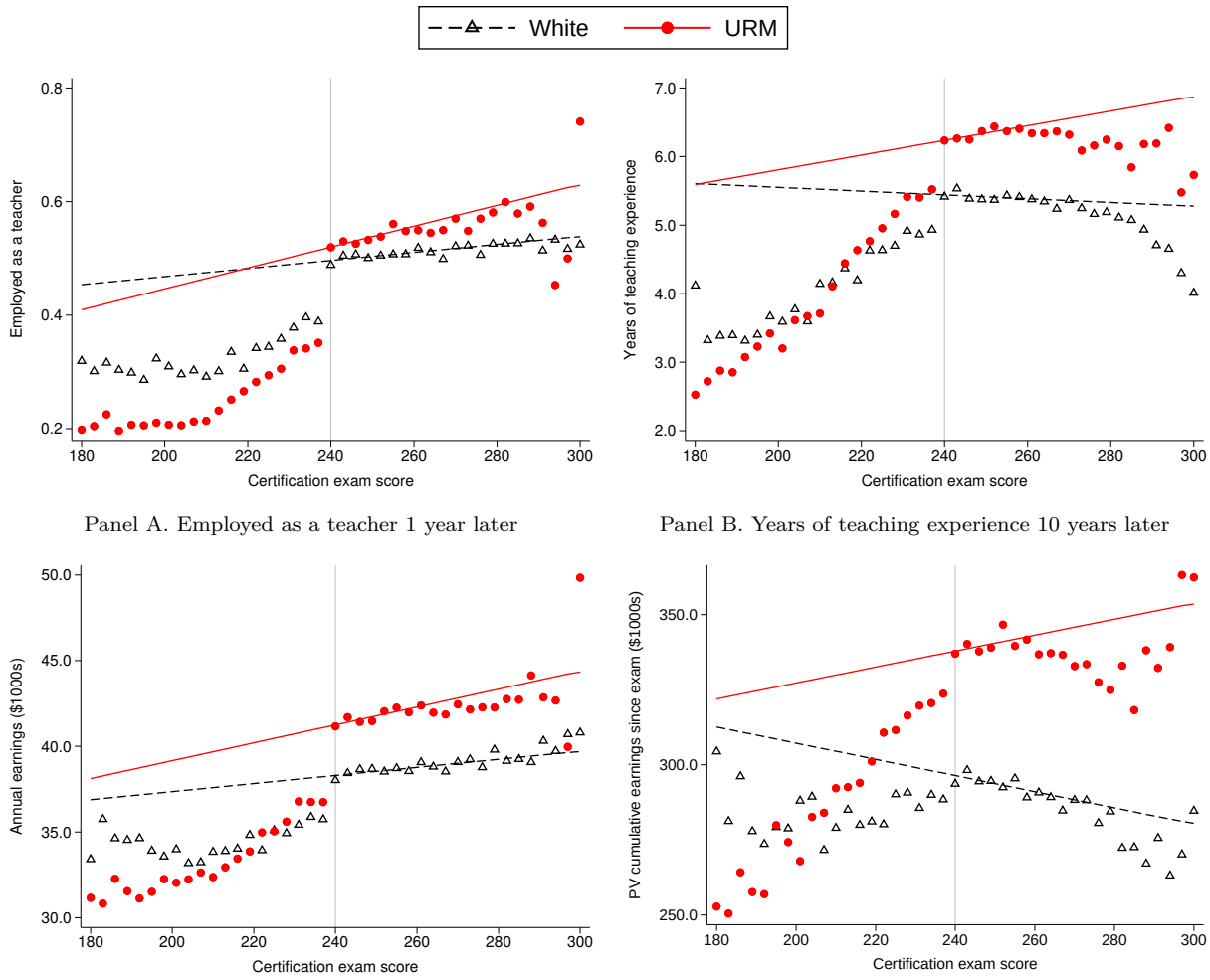
Panel D. PV cum. earnings after 5 years ($1000s)

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

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

Figure A7: Robustness to RD bandwidth: Earnings outcomes by race

Notes: This figure presents RD effects of failing a certification exam on earnings outcomes by race/ethnicity using different RD bandwidths (x-axis). The outcomes are annual earnings and the present value of cumulative earnings measured one, five, and ten years after the exam, as indicated by the panel title. Hollow markers depict RD coefficients computed using integer bandwidths from 5–30 certification score points. The large solid markers show our benchmark estimates from Table 2 using the Calonico et al. (2019) bandwidth. Dashed lines depict 95 percent confidence intervals using standard errors clustered at the individual level.



Panel A. Employed as a teacher 1 year later

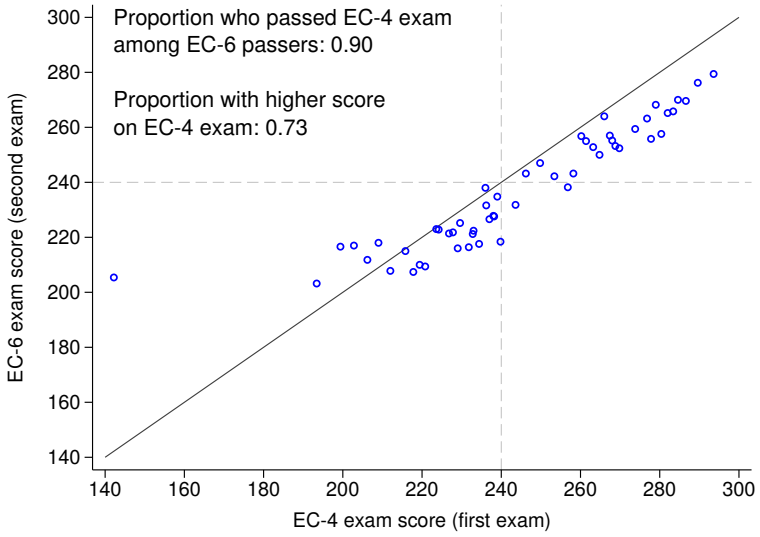
Panel B. Years of teaching experience 10 years later

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

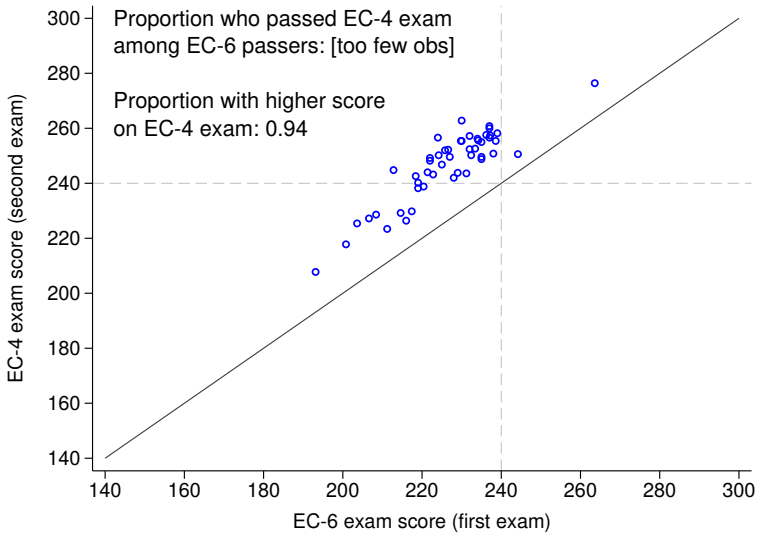
Panel D. PV cumulative earnings after 10 years ($1000s)

Figure A8: Extrapolated effects of failing certification exam on outcomes

Notes: This figure illustrates our extrapolation method for estimating the effects of failing a certification exam below the passing threshold. Our extrapolation computes the linear relationship between outcomes and certification exam scores in the 240–260 score range, and then uses a linear projection to predict counterfactual outcomes below the threshold (240). We perform this extrapolation separately for white (dashed black lines) and URM (solid red lines) exam takers. The dependent variables are an indicator for being employed as a teacher, years of teaching experience, annual earnings (in thousands of 2019 dollars), and the present value of cumulative earnings since the year of the exam (in thousands of 2019 dollars), as indicated in the panel titles.



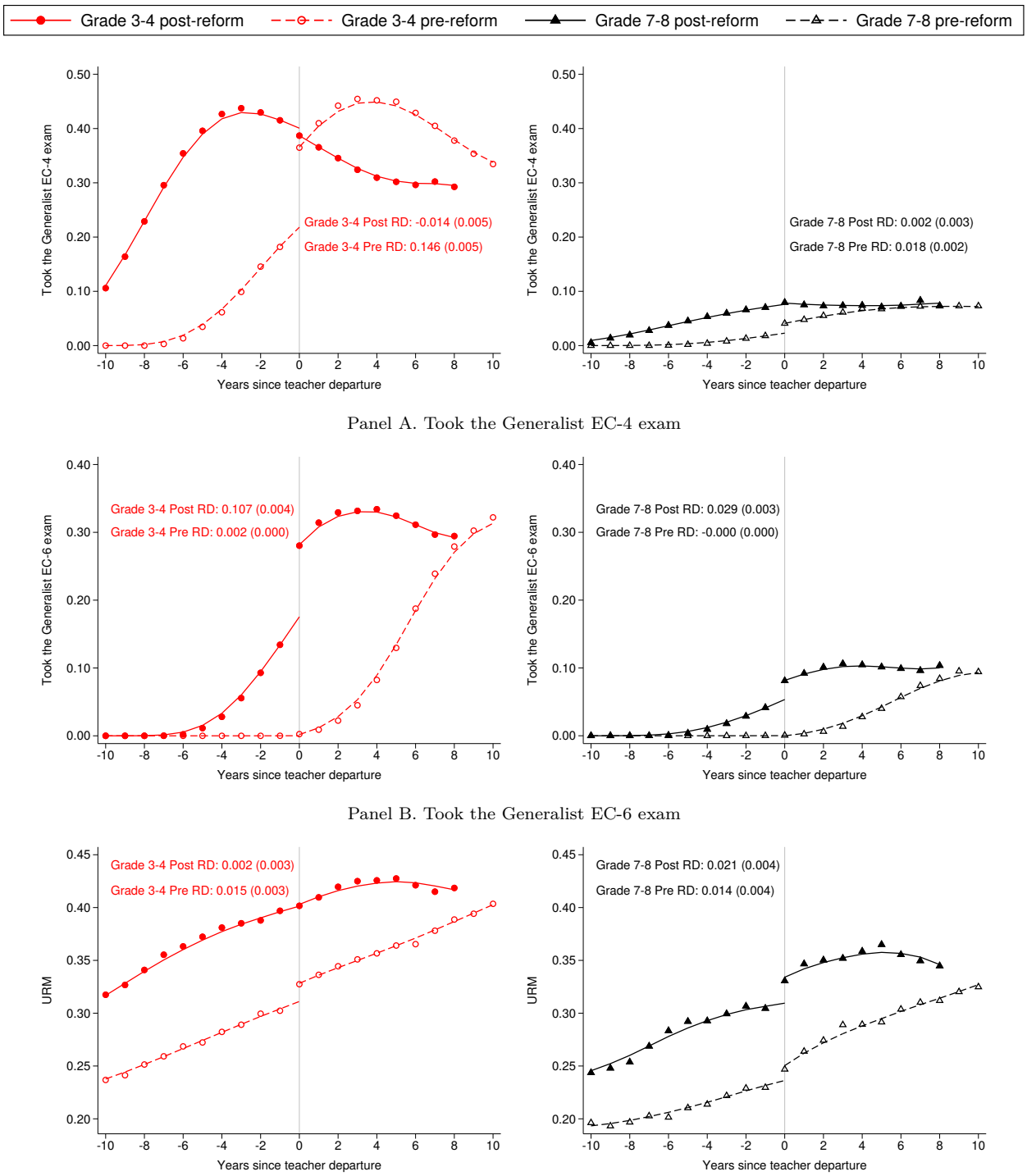
Panel A. Exam takers who took the EC-4 exam first



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

Figure A9: 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. 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.

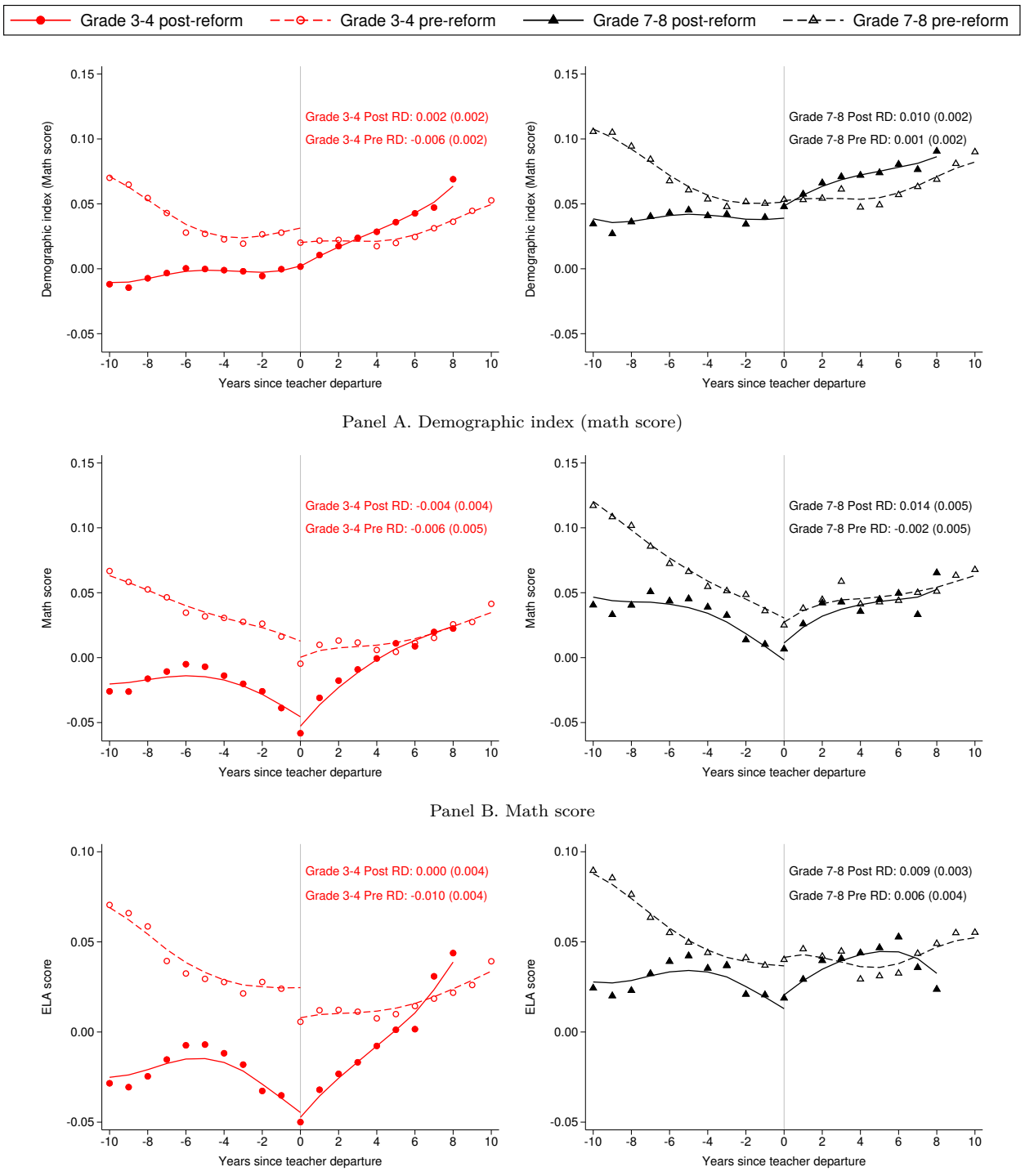
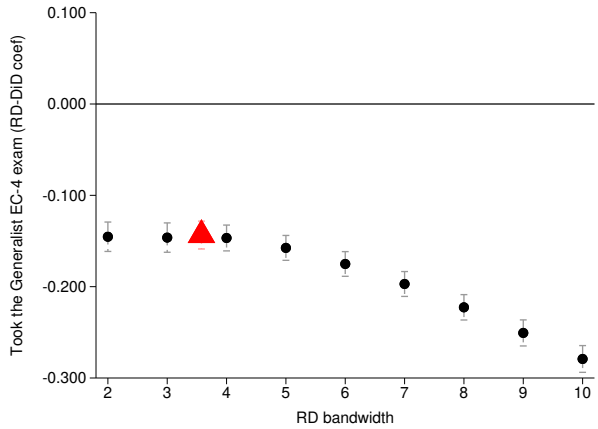
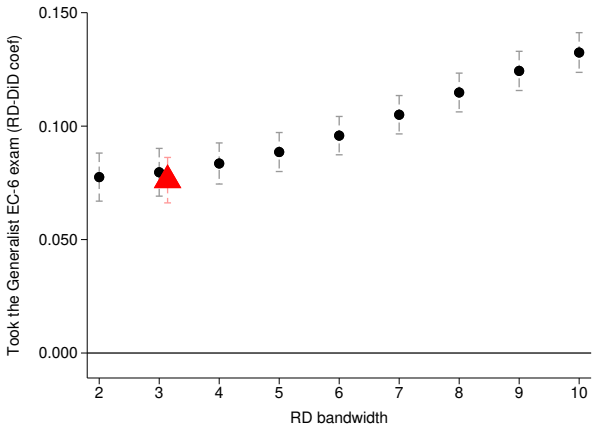


Figure A11: Effects of teacher departures on student achievement

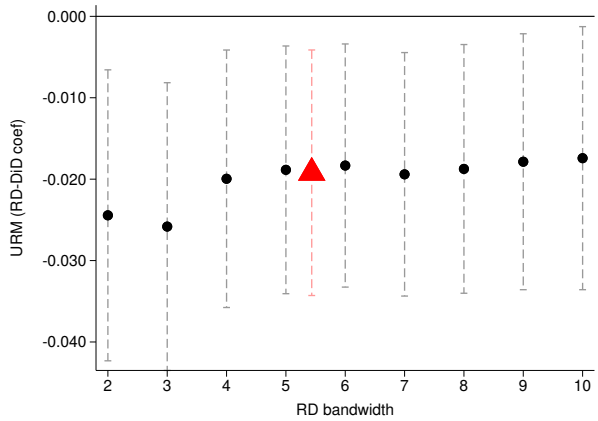
Notes: This figure contains RD graphs that show how teacher departures affect the math/ELA achievement of students in a school/grade. Red circles represent grade 3–4 teachers. Black triangles represent grade 7–8 teachers. Hollow symbols depict teacher departures that occurred in $y \in$ 2005–2010 (pre-reform). 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.



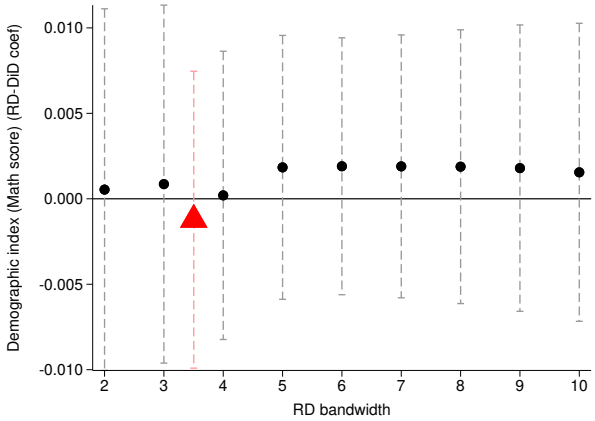
Panel A. Took the Generalist EC-4 exam



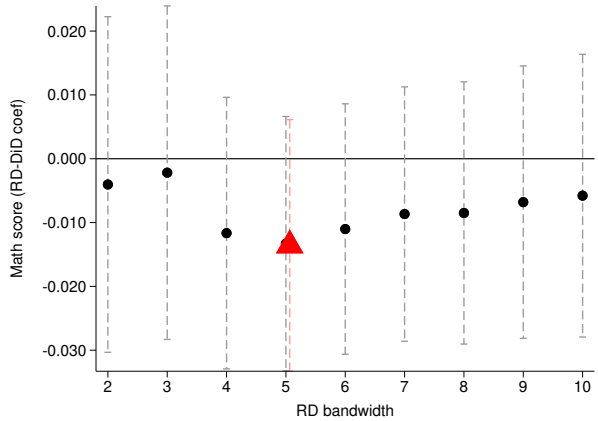
Panel B. Took the Generalist EC-6 exam



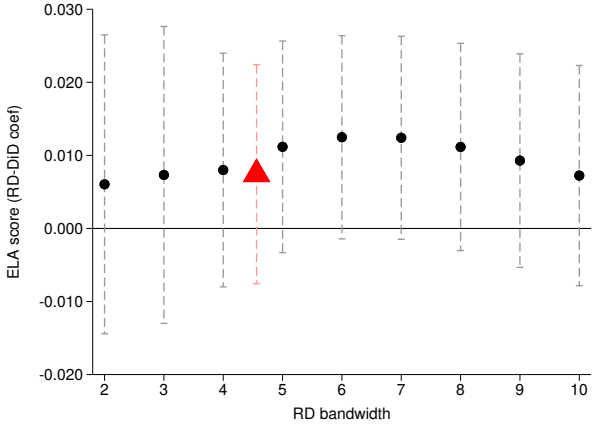
Panel C. URM



Panel D. Demographic index (Math score)



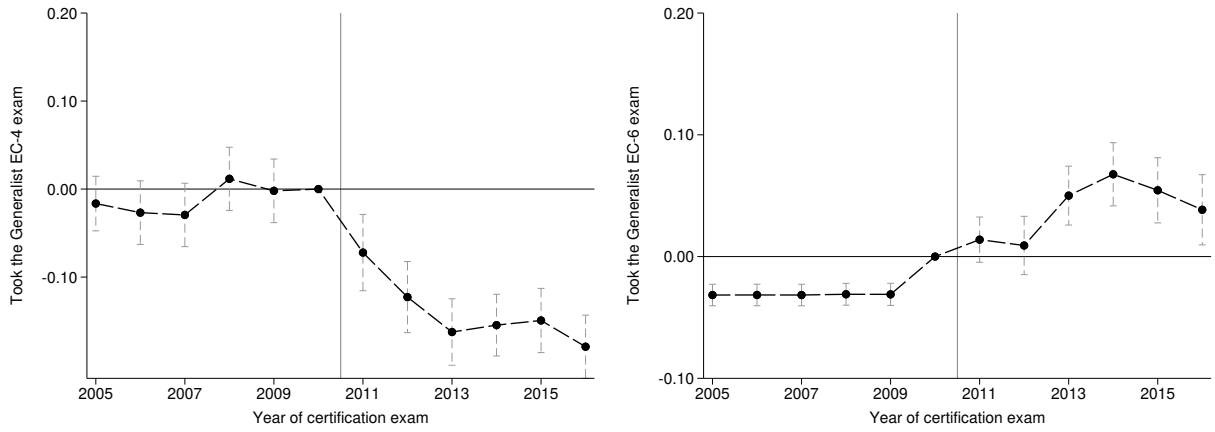
Panel E. Math score



Panel F. ELA score

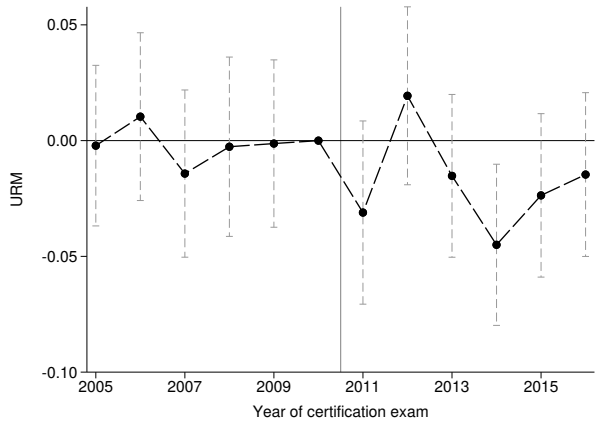
Figure A12: RD-DiD effects of harder certification exams — Robustness to RD bandwidth

Notes: This figure displays the robustness of our RD-DiD estimates 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). Circular markers depict RD coefficients computed using integer bandwidths from $h^Y \in$ 2–10 years relative to the year of the teacher departure. The large triangular markers show our benchmark estimates from Tables 5 and 6 using the Calonico et al. (2019) bandwidth. Note that the Calonico et al. (2019) RD bandwidths (triangular markers) vary across treated/control groups and pre/post periods since we compute them separately for each RD regression, while our robustness bandwidths (circular markers) are constant across these groups. Dashed lines represent 95 percent confidence intervals using standard errors clustered at the school level.

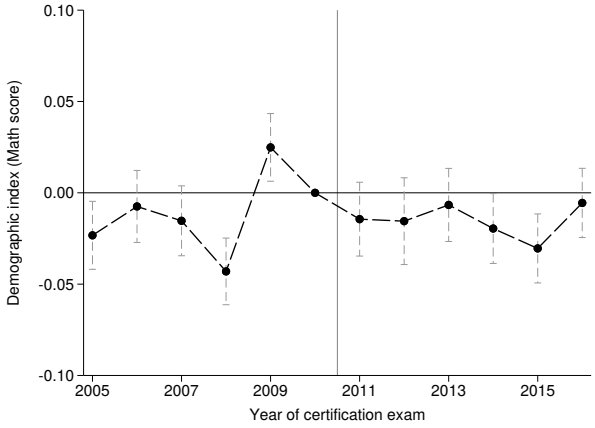


Panel A. Took the Generalist EC-4 exam

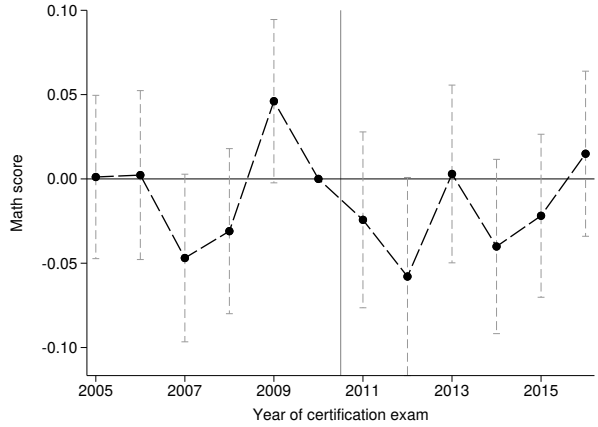
Panel B. Took the Generalist EC-6 exam



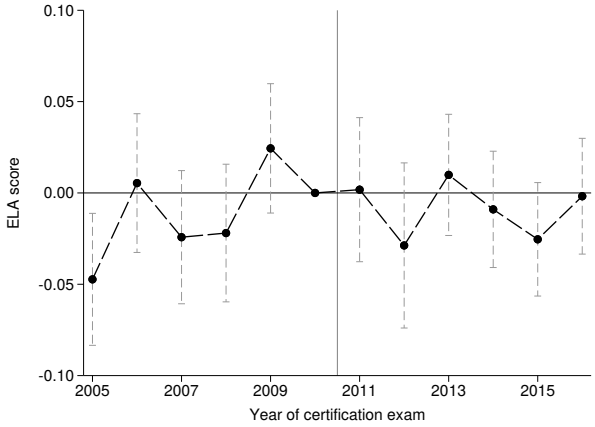
Panel C. URM



Panel D. Demographic index (Math score)



Panel E. Math score



Panel F. ELA score

Figure A13: RD-DiD effects of harder certification exams — Event studies

Notes: This figure displays event study versions of our RD-DiD estimates. For this specification, we replace the Post_p indicators in equation (8) with dummies for each teacher departure year $y \in \{2005, \ldots, 2016\}$, omitting the 2010 dummy. This yields θ_y coefficients on the interactions between Treated_g and the departure year dummies. These coefficients represent the differential change in the effects of teacher departures in grades 3–4 and grades 7–8 between 2010 and year y. Circular markers represent the θ_y coefficients from this event study specification. 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: This table presents RD balance tests using pre-determined exam taker characteristics as outcome variables. The dependent variables include demographic characteristics measured during K–12 schooling, high school math/ELA test scores, and pre-exam employment and earnings in Texas. Columns (A) and (B) display dependent variable means for candidates with exam scores 0–10 points above the threshold. Columns (C)–(E) show RD coefficients β from equation (1) estimated separately for all, white, and URM candidates. Column (F) shows the p-value from a test of equality of the RD coefficients for white and URM candidates. "N (Male)" indicates the sample size for the outcome of male. 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						
Retook this exam by now	0.001	0.001	0.779***	0.806***	0.791***	0.052
			(0.004)	(0.006)	(0.006)	
Passed this certification exam by now	1.000	1.000	-0.294***	-0.258***	-0.295***	0.000
			(0.005)	(0.008)	(0.007)	
# 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)	
Passed any certification exam by now	1.000	1.000	-0.191***	-0.152***	-0.198***	0.000
			(0.005)	(0.007)	(0.006)	
Total # exams taken	1.634	1.701	0.938***	0.991***	0.966***	0.442
			(0.014)	(0.024)	(0.021)	
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						
Retook this exam by now	0.001	0.001	0.805***	0.831***	0.823***	0.338
			(0.004)	(0.006)	(0.005)	
Passed this certification exam by now	1.000	1.000	-0.260***	-0.232***	-0.251***	0.093
			(0.005)	(0.009)	(0.007)	
# 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)	
Passed any certification exam by now	1.000	1.000	-0.142***	-0.124***	-0.131***	0.421
			(0.004)	(0.007)	(0.006)	
Total # exams taken	2.208	2.278	0.995***	1.009***	1.071***	0.231
			(0.021)	(0.035)	(0.037)	
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						
Retook this exam by now	0.001	0.001	0.813***	0.840***	0.839***	0.923
			(0.005)	(0.008)	(0.007)	
Passed this certification exam by now	1.000	1.000	-0.241***	-0.214***	-0.221***	0.582
			(0.006)	(0.010)	(0.008)	
# 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)	
Passed any certification exam by now	1.000	1.000	-0.143***	-0.123***	-0.122***	0.930
			(0.006)	(0.009)	(0.007)	
Total # exams taken	2.499	2.512	1.008***	0.974***	1.113***	0.076
			(0.034)	(0.052)	(0.056)	
N (Passed this certification exam by now)	9,848	9,734	310,404	145,790	88,718	234,508

Notes: This table presents RD estimates of the impacts of failing a certification exam on exam outcomes. Panels A–C show outcomes measured one, five, and ten years after the exam, respectively. The dependent variables include indicators for retaking and passing the first certification that the individual attempted, the number of days to pass the initial certification exam (conditional on passing), an indicator for passing any certification exam, and the total number of certification exams taken. Columns (A) and (B) display dependent variable means for candidates with exam scores 0–10 points above the threshold. Columns (C)–(E) show RD coefficients β from equation (1) estimated separately for all, white, and URM candidates. Column (F) shows the p-value from a test of equality of the RD coefficients for white and URM candidates. "N (Passed this certification exam by now)" indicates the sample size for the outcome of passed this certification exam by now. 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: This table presents RD estimates of the impacts of failing a certification exam on earning outcomes. Panels A–C show outcomes measured one, five, and ten years after the exam, respectively. The dependent variables are an indicator for appearing the TWC earnings data, annual earnings (excluding zeroes, as in Table 2), annual earnings including zeros, earnings from teaching (excluding zeros), and earnings from teaching including zeros. All earnings are measured in thousands of 2019 dollars. Columns (A) and (B) display dependent variable means for candidates with exam scores 0–10 points above the threshold. Columns (C)–(E) show RD coefficients β from equation (1) estimated separately for all, white, and URM candidates. Column (F) shows the p-value from a test of equality of the RD coefficients for white and URM candidates. "N (Annual earnings)" indicates the sample size for the outcome of annual earnings (excluding zeroes). 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: This table presents RD estimates of the impacts of failing a certification exam on teaching and earning outcomes. This table is identical to Table 2, except we use the running variable with the adjusted scale as described in Section 3.2 and Appendix Figure A2. Panels A–C show outcomes measured one, five, and ten years after the exam, respectively. The dependent variables are an indicator for being employed as a teacher, years of teaching experience, annual earnings (in thousands of 2019 dollars), and the present value of cumulative earnings since the year of the exam (in thousands of 2019 dollars). Columns (A) and (B) display dependent variable means for candidates with exam scores 0–10 points above the threshold. Columns (C)–(E) show RD coefficients β from equation (1) estimated separately for all, white, and URM candidates. Column (F) shows the p-value from a test of equality of the RD coefficients for white and URM candidates. "N (Employed as a teacher)" indicates the sample size for the outcome of employment as a teacher. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: 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: This table presents results from our extrapolation method for estimating the effects of failing a certification exam below the passing threshold. Our extrapolation computes the linear relationship between outcomes and certification exam scores in the 240–260 score range, and then uses a linear projection to predict counterfactual outcomes below the threshold (240). We perform this extrapolation separately for white and URM exam takers. We compute the difference between the extrapolated counterfactual outcome and the actual outcome for each individual (defined to be equal to zero for individuals above the passing threshold) and then compute the average difference in each sample defined by the column headers. Columns (A)–(B) show results for white and URM exam takers with a score within ten points of the passing score (230–239). Columns (C)–(D) show results for all white and URM exam takers who scored below the threshold. Columns (E)–(F) show results for all white and URM exam takers (both above and below the threshold). Panels A–C show results for outcomes measured one, five, and ten years after the exam. The dependent variables are an indicator for being employed as a teacher, years of teaching experience, annual earnings (in thousands of 2019 dollars), and the present value of cumulative earnings since the year of the exam (in thousands of 2019 dollars), as indicated in the panel titles. "N (Employed as a teacher)" indicates the sample size for the outcome of employment as a teacher. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: 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 A8: Example of topics covered on the EC-4 and EC-6 exams

(A) EC-4 exam topics	(B) EC-6 exam topics
Panel A. Mathematics: "Patterns and Algebra" topics	
Illustrate numerical relationships	Illustrate numerical relationships
	Model functions
Understand algebraic thought	Understand algebraic thought
	Formulate rules
Know how to use patterns	Know how to use patterns
Use relationships in real-world applications	Use relationships in real-world applications
Translate real-life applications to algebraic expressions	Translate real-life applications to algebraic expressions
Model problem-solving	Model problem-solving
	Determine linear function model
	Understand algebraic "phrases" and use them in problem solving
Panel B. English Language Arts and Reading: "Reading Comprehension" topics	
	Understand reading comprehension as construction of meaning
	Help students increase their reading vocabulary
Understand influences on reading comprehension	Understand factors that affect reading comprehension
Understand and teach different levels of comprehension	Understand and teach levels of comprehension
Facilitate the transition to "reading to learn"	Facilitate the transition to "reading to learn"
Use instructional approaches that build comprehension	Use instructional approaches that build comprehension
Help children monitor and improve their comprehension	Teach strategies that aid comprehension before, during, and after reading
	Understand metacognition and its role in comprehension
	Provide explicit instruction in comprehension strategies
Use multicultural instructional approaches	Use multicultural instructional approaches
Use assessments to meet individual needs	Understand grade-level expectations for comprehension
Teach elements of literary analysis	Teach elements of literary analysis
	Distinguish between and support guided and independent reading
Foster collaboration with families and other professionals	Foster collaboration with families and other professionals

Notes: This table provides examples of topic covered on the EC-4 (column A) and EC-6 (column B) exams. The sources for this table are the CliffsNotes *TExES Generalist EC-4* and *EC-6* test prep books. Topics are section headers under the Mathematics "Patterns and Algebra" competency (Panel A) and the English Language Arts and Reading "Reading Comprehension" competency (Panel B). See the Table of Contents in both books.

Table A9: Balance tests for Assumption 3

	(A)	(B)	(C)	(D)	(E)
	Mean diff. between EC-6 and EC-4 exam passers	Coefficient on Passed EC-6 × Has math VA		Coefficient on Passed EC-6 × Has ELA VA	
	All	URM	White	URM	White
Certification exam score	-8.261	0.168	0.202	-0.727	-0.307
		(0.614)	(0.514)	(0.619)	(0.510)
Male	0.005	-0.009	-0.012	0.020	-0.023***
		(0.017)	(0.010)	(0.015)	(0.008)
Hispanic	-0.044	-0.018		0.012	
		(0.021)		(0.021)	
Black	-0.008	0.018		-0.012	
		(0.021)		(0.021)	
Mean proportion white at school during K-12	0.038	0.041**	0.002	0.029	0.003
		(0.018)	(0.010)	(0.019)	(0.009)
High school graduation cohort	4.563	0.157	0.348*	-0.120	0.428**
		(0.224)	(0.179)	(0.240)	(0.166)
Ever identified as an immigrant during K-12	-0.003	0.023	-0.001	0.010	-0.003
		(0.014)	(0.004)	(0.013)	(0.004)
Economically disadvantaged during K-12	0.009	-0.014	-0.003	-0.024	-0.003
		(0.028)	(0.010)	(0.029)	(0.010)
Identified as at risk of dropping out during K-12	-0.044	0.030	-0.006	0.028	-0.004
		(0.020)	(0.011)	(0.021)	(0.010)
Gifted education during K-12	-0.010	-0.035*	0.000	-0.026	-0.001
		(0.020)	(0.017)	(0.021)	(0.016)
In special education during K-12	0.005	-0.001	0.006	-0.001	0.003
		(0.005)	(0.005)	(0.004)	(0.005)
Ever in bilingual education during K-12	0.028	0.034	-0.002	0.035	-0.001
		(0.026)	(0.001)	(0.026)	(0.002)
Ever in English as a Second Language during K-12	0.005	0.029	0.004	0.003	0.003
		(0.020)	(0.003)	(0.020)	(0.002)
Ever had limited English proficiency during K-12	0.022	0.035	0.003	0.036	0.002
		(0.030)	(0.004)	(0.030)	(0.003)
High school math score	0.043	-0.087*	0.015	-0.096*	0.023
		(0.051)	(0.039)	(0.052)	(0.038)
High school ELA score	0.104	0.018	0.004	-0.015	0.029
		(0.048)	(0.039)	(0.050)	(0.040)
Had earnings in Texas in prior year	-0.055	0.001	0.007	0.014	-0.017
		(0.024)	(0.019)	(0.024)	(0.019)
Prior year annual earnings ($1000s)	-2.437	0.154	0.861	-0.294	0.428
		(0.859)	(0.754)	(0.827)	(0.641)
N (Certification exam score)	591,495	101,360	204,704	101,358	204,658
P-value: All coefs zero		0.334	0.541	0.364	0.126

Notes: This table presents balance tests related to Assumption 3 in our disparate impact analysis (see Section 4.2). The sample is exam takers who took and passed an EC-4 or EC-6 exam as their first certification exam. The data is at the individual (i) × year (t) level with an observation for each year in 2012–2019 that is 3–8 years after the individual took the exam. The dependent variables include the certification exam score, demographic characteristics measured during K–12 schooling, high school math/ELA test scores, and pre-exam employment and earnings in Texas. For each variable, column (A) shows the mean for EC-6 exam passers minus the mean for EC-4 exam passers. Columns (B)–(E) show δ coefficients from the regression:

$$Y_i = \alpha_{e(it)} + \beta \text{EC6}_i + \gamma \text{HasVA}_{it} + \delta \text{EC6}_i \text{HasVA}_{it} + \epsilon_{it}$$

where Y_i is the individual characteristic, $\alpha_{e(it)}$ are dummies for years $e(it)$ since the certification exam, EC6_i is an indicator for taking the EC-6 exam, and HasVA_{it} is indicator equal to one if we observe value-added for individual i in year t. We estimate this regression separately for URM and white exam takers and for HasVA_{it} defined by math and ELA value-added. "N (Certification exam score)" indicates the sample size for the outcome of certification exam score. Standard errors in parentheses are clustered at the individual level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: 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 A11: 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 below.

Table A12: 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 below.

Notes for Tables A11 and A12

Tables A11 and A12 present our estimates of the policy-relevant disparate impact of the EC-6 exam relative to the EC-4 exam. Table A11 presents results for math teachers using math value-added as a measure of teaching quality. Table A12 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 $\times$ 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 difference 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 * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. See Appendix B.3 for details on our bootstrap procedure.

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.
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 value-added using an index of grade retention, attendance, and suspension (see Appendix B.2.2).
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 A13: 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 (# *sty* observations)	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 * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A14: 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 B4 in Appendix B.4.2) . Standard errors in parentheses are clustered at the school level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A15: 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 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.4.3 for details on the sample and these RD-DiD specifications. Standard errors in parentheses are clustered at the school level with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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. Individuals who do not appear in the TWC have missing values of this variable (except when noted in Appendix Table A4).

- **At risk of dropping out.** The proportion of students who are reported as being at risk of dropping out in the TEA data.

- **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 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 in the TEA data.

- **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.
- **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 zero years of experience.
- **Has earnings in Texas.** An indicator for appearing the TWC data.
- **High school math/ELA score.** The math/ELA score we have for a teacher's high school math/ELA exams, standardized to mean zero and SD one within the population of individuals who took the same exam in the same year. 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 testing regimes.
- **In gifted education.** The proportion of students who are reported as being in gifted and talented programs in the TEA data.
- **Math/ELA score (SD units).** Scale scores on standardized grade 3–8 and high school 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. See the definition of high school math/ELA score for details on the high school exams.
- **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. Controls for grade 3 students include only those based on demographic variables, not lagged test scores.

- **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 attempt.** 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. We use the individual's score on their first attempt at their first PPR exam. Scores range from 100–300 with 240 representing a passing score.
- **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. We set earnings equal to zero for years in which the individual has no earnings in the TWC data; thus this variable is defined for all individuals in our sample.
- **Recently-certified teacher.** The proportion of teachers who received their initial teaching certification within five years relative to the year in which they are teaching.
- **Teacher certified 6+ years ago.** The proportion of teachers received their initial teaching certification more than five years ago relative to the year in which they are teaching.
- **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. These racial/ethnic 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 Teacher Value-Added.

B.2.1 Value-added for test scores. This section describes how we compute teacher value-added for student math/ELA scores. Our methods closely follow those in Chetty et al. (2014a).

We select a sample of students and math/ELA teachers for whom we can cleanly compute value added. We exclude students with missing values in current or lagged test scores in the subject for which we are estimating value-added. We restrict the sample to 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), and we also 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 students with current and lagged scores in the relevant subject.

To compute value-added in this sample, we begin by estimating the following regression:

$$A_{it}^{*} = \beta' \mathbf{X}_{it} + \alpha_{j(i)} + \nu_{it}, \tag{B1}$$

where A_{it}^{*} is the standardized math or ELA score of student i in year t. This regression includes fixed effects for teachers, $\alpha_{j(i)}$, and a large vector of controls, $\mathbf{X}_{it}$. Our control vector $\mathbf{X}_{it}$ follows the specification in Chetty et al. (2014a). At the individual level, we include lagged standardized math and ELA scores and their squares and cubes as well as student demographic characteristics (economic disadvantage, ethnicity/race, sex, whether they are in special education, whether they are at risk, and whether they are gifted). We also include average lagged standardized test scores in math and ELA and their squares and cubes and average individual demographic characteristics at both the classroom level and the school level.[33] We interact all individual, classroom, and school-level controls with dummies for grade levels. Lastly, we include grade $\times$ year dummies and dummies for grade $\times$ student population type (regular, honors, gifted and talented, bilingual, and special education).

Our measure of value-added for teacher j in year t is a prediction based on average test score residuals from the same teacher in other years. Specifically, we estimate equation (B1) separately for math and ELA scores and compute residuals excluding teacher fixed effects:

$$A_{it} = A_{it}^{*} - \hat{\beta}' \mathbf{X}_{it}. \tag{B2}$$

[33] Our sample restrictions require that students have non-missing lagged test scores in the subject defined by the outcome variable, A_{it}^{*}, but we implement a similar procedure as Chetty et al. (2014a) to deal with missing values in other covariates. Specifically: (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.

We then compute the average of these residuals at the teacher (j) by year (t) level, i.e., $\bar{A}_{jt} = \frac{1}{n_{jt}} \sum_{i \in \{i:j(i,t)=j\}} A_{it}$. Letting $\mathbf{A_j^{-t}} = \{\bar{A}_{js}\}_{s:s \neq t}$ denote the vector of mean residuals for teacher j excluding year t, our teacher $\times$ year value-added estimate is given by:

$$\hat{\mu}_{jt} = \psi' \mathbf{A_j^{-t}}, \tag{B3}$$

where $\psi \equiv \Sigma_A^{-1}\gamma$ is a vector of coefficients chosen to minimize the mean-squared error of test score forecasts. $\gamma = \{\text{cov}(\bar{A}_{jt}, \bar{A}_{js})\}_{s:s \neq t}$ is a vector of auto-covariances of mean test scores taught by a given teacher, and Σ_A^{-1} is the inverse of the variance-covariance matrix of $\mathbf{A_j^{-t}}$.[34] Following Chetty et al. (2014a), we assume that the auto-covariances are constant across teachers and constant for all pairs of years that share a common time lag s, i.e., $\text{cov}(\bar{A}_{jt}, \bar{A}_{j,t-s}) = \sigma_{As}$ for all j and all t. Our estimate of value-added for teacher j in year t is given by $\hat{\mu}_{jt}$ in equation (B3).

B.2.2 Value-added for non-cognitive outcomes. In robustness analyses, we also compute teacher value-added for non-cognitive outcomes (Jackson, 2018). We use the same sample as that for test score value-added; in other words, we examine the same set of math-/ELA teachers, but we compute these teachers' value-added for other outcomes. Following Jackson (2018), our main outcome is an index that measures student i's non-cognitive outcomes in year t. The index is constructed using students' behavioral outcomes: attendance (percentage of days that student was present), grade retention (indicator), and suspension (indicator).[35] To compute non-cognitive value added, we use the non-cognitive index as the outcome variable in equation (B1), and we add the following controls to the vector $\mathbf{X}_{it}$. At the individual level, we include lagged values of the non-cognitive index, attendance, grade retention, and suspension and, for non-binary variables, their squares and cubes. We also include average lagged values of the non-cognitive index, attendance, grade retention, and suspension and their squares and cubes at both the classroom level and the school level.[36] We interact all individual, classroom, and school-level controls with dummies for grade levels. Aside from these modifications, our method for estimating non-cognitive value added is the same as that described in Appendix B.2.1.

B.3 Disparate impact bootstrap. For our disparate impact analysis in Section 4, 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. This appendix subsection describes our bootstrap procedure.

For each bootstrap iteration b, we first compute teacher value-added by drawing a random sample of students with replacement holding class sizes fixed. For example, if teacher j has n_{jkt} students in their classroom k in year t, we randomly draw n_{jkt} observations with

[34] The diagonal elements of Σ_A are the variance of mean test score residuals and the off-diagonal elements are the auto-covariances of mean test score residuals for the same teacher across years, i.e., $\text{cov}(\bar{A}_{jt}, \bar{A}_{j,t-s})$.

[35] 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 compute 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).

[36] Our sample restrictions require that students have non-missing lagged values of each of the non-cognitive variables (attendance, grade retention, and suspension).

replacement from their actual set of students. Using this bootstrap sample, we compute teacher value-added estimates $\hat{\mu}^b_{jt}$ following the methodology described in Appendix B.2.

We then compute disparate impact by drawing a random sample of certification exam takers with replacement holding fixed the distribution of exams and race/ethnicity. For example, if n_{re} individuals with race/ethnicty r (white or URM) took exam e, we randomly draw n_{re} observations with replacement from the actual set of certification exam takers. Using this bootstrap sample, we compute disparate impact using equation (6) and the methodology described in Section 4.3. Note that for each bootstrap iteration b, we use the teacher value-added estimates $\hat{\mu}^b_{jt}$ from iteration b in computing disparate impact.

We run $b = 1 \ldots 500$ bootstrap iterations and use the standard deviation of our disparate impact estimates to compute standard errors and confidence intervals.

B.4 RD-DiD specification. This section provides details on our regression discontinuity difference-in-differences (RD-DiD) specification that we use in Section 5.

B.4.1 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 A15, 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. 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 cases in which the departure causes a significant change in teacher composition at the school/grade/subject level, which is the level at which we can connect teachers to students 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.[37]

B.4.2 Benchmark RD-DiD specifications. This subsection provides details on our RD-DiD regression specifications.

In Tables 5–6 (and Appendix Table A13), 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 A14, 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:

$$Y_{st} = \Big(\phi \text{Treated}_g + \delta \text{Post}_p + \theta \text{Treated}_g \text{Post}_p\Big)\mathbf{1}\{\tau_{ty} \geq 0\} + \gamma_{sy} + \varepsilon_{sty} \quad \text{if } |\tau_{ty}| \leq 3. \tag{B4}$$

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.4.3 RD-DiD specifications for Table A15. In Table A15, 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.4.1) 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 A15, 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 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:

$$\begin{aligned} Y_{st} = &\Big(\phi \text{SameSubject}_e + \delta \text{Post}_p + \theta \text{SameSubject}_g \text{Post}_p\Big)\mathbf{1}\{\tau_{ty} \geq 0\} + \\ &\alpha_{gp}\tau_{ty} + \psi_{gp}\mathbf{1}\{\tau_{ty} \geq 0\}\tau_{ty} + \gamma_{sy} + \varepsilon_{sty} \quad \text{if } |\tau_{ty}| \leq h^Y_{gp}. \end{aligned} \tag{B5}$$

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

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

departs). All other variables in equation (B5) 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 A15, 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 (B5) 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:[38]

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

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

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

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

[38] 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:[39]

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

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:

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

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

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

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 2 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 elementary 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 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 2, 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 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 VARatioRoR(r) under Assumption 2B; and 2) the y- and e-specific RoR estimator VARatioRoR(r, y, e) under Assumption 2C. However, in cases where y takes many different values, the 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:

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