Will Job Testing Harm Minority Workers?*

David H. Autor
MIT and NBER

David Scarborough
Unicru, Inc.

July 27, 2004
Revised from December 2003

Abstract

Because of the near-universal finding that minorities fare poorly on standardized tests, the use of such tests for employment screening is thought to pose an equity-efficiency trade-off: improved selection comes at a cost of screening out more minority applicants. This paper investigates the consequences of standardized testing for minority employment and productivity. The data come from a large, geographically dispersed retail firm whose 1,363 stores switched from paper to electronic job applications during 1999 and 2000. Both hiring methods use face to face interviews, while the test-based regime also places substantial weight on a computer administered personality assessment. We find strong evidence that the move to standardized testing raised productivity at treated stores: increasing mean and median employee tenure by 10 percent, and slightly lowering the frequency at which workers were fired for cause. Analysis of electronic applications reveals that minority applicants performed significantly worse on the screen and were less likely to be hired conditional on their scores. Despite this, the use of standardized testing had no adverse consequences for minority hiring, and productivity gains were equally large among minority and non-minority workers. We provide a model that explains these facts as a consequence of statistical discrimination used prior to the introduction of standardized testing.

*We thank Daron Acemoglu, Joshua Angrist, David Card, Edward Lazear, Michael Greenstone, Sendhil Mullainathan, Roberto Fernandez, and numerous seminar participants for insightful suggestions. We acknowledge Tal Gross and Jared Gross for superb research assistance, and Alan Baumbusch of Unicru, Inc. for generous assistance with all data matters. Autor gratefully acknowledges financial support from the Alfred P. Sloan foundation and National Science Foundation grant SES-0239538.

1 Introduction

In the early 20th century, the majority of unskilled, industrial employees in the United States were hired with no systematic efforts at selection (Wilk and Cappelli, 2003). Sanford Jacoby's well-known industrial relations text describes an early 20th century Philadelphia factory at which foremen tossed apples into crowds of job-seekers, and hired the men who caught them (Jacoby, 1985, p. 17). More recently, Murnane and Levy (1996, p. 19) quote a company manager describing Ford Motor Company's hiring process in 1967: "If we had a vacancy, we would look outside in the plant waiting room to see if there were any warm bodies standing there." These hiring practices are no longer commonplace. During the 1980s, as much as one-third of large employers adopted systematic skills testing for job applicants (Bureau of National Affairs, 1980 and 1988). But skills testing has remained rare in hiring for hourly wage jobs, where training investments are typically modest and employment spells brief (Aberdeen, 2001). Due to advances in information technology, these practices are now poised for change. With increasing prevalence, employers use computerized job applications and assessments to administer and score personality tests, perform online background checks and guide hiring decisions. Over time, these tools are likely to become increasingly sophisticated, as for example has occurred in the consumer credit industry.

Widespread use of job testing has the potential to raise aggregate productivity by improving the quality of matches between workers and firms. But there is a pervasive concern, reflected in public policy, that job testing may have adverse distributional consequences, commonly called 'disparate impacts.' Because of the near universal finding that minorities, less-educated and low-socioeconomic-status (SES) individuals fare relatively poorly on standardized tests (Neal and Johnson, 1996; Jencks and Phillips, 1998), job testing is thought to pose a trade off between efficiency and equity; better candidate selection comes at a cost of reduced opportunity for groups with lower average test scores (Hartigan and Wigdor, 1989; Hunter and Schmidt, 1982).[1] This concern is forcefully articulated by Hartigan and Widgor in the introduction to their influential National Academy of Sciences Report, *Fairness in Employment Testing* (p. vii):

> "What is the appropriate balance between anticipated productivity gains from better employee selection and the well-being of individual job seekers? Can equal employment opportunity be said to exist if screening methods systematically filter out very large proportions of minority candidates?"

This presumed trade-off has garnered substantial academic, legal and regulatory attention, including a

[1] Jencks and Phillips (1998) report that in 1986, the mean black-white test score gap on the Armed Forces Qualification Test (an IQ test) was 0.7 to 0.9 standard deviations.

landmark Supreme Court decision limiting use of employment tests that are not directly job-relevant (*Griggs v. Duke Power Co.,* 3 FEP Cases 175, 1971), a series of Equal Employment Opportunity Commission guidelines regulating employee selection procedures (U.S. Department of Labor, 1978), and two National Academy of Sciences studies evaluating the efficacy and fairness of job testing (Hartigan and Wigdor, 1989; Wigdor and Green, 1991).

Yet, despite a substantial body of research and policy, the evidence for an equity-efficiency trade-off in job testing is not well established. As our illustrative model below demonstrates, there are two assumptions underlying the presumed trade-off, and these assumptions do not appear equally palatable. The first assumption is that employment tests provide a valid predictor of worker productivity; if so, testing has the potential to improve applicant selection.[2] The second assumption is that, absent job testing, firms hire in a manner that is blind to, or weakly correlated with, the tested attribute; if so, testing will reduce hiring rates from demographic groups with below average test scores (a disparate impact).

Because competitive employers face a strong incentive to select and remunerate workers according to productivity, a setting where hiring is blind to an important productive characteristic appears artificial.[3] Consider instead a case where firms screen informally for a tested attribute and testing improves the accuracy of screening. Will the resulting gain in screening precision reduce hiring from low scoring groups? As we show below, the answer is ambiguous without further assumptions; hiring rates from groups with low scores could rise or fall slightly. Moreover, the gains from testing in these cases will primarily accrue from better selection *within* applicant groups (i.e., minorities, non-minorities) rather than from differential shifts in cross-group shifts in hiring. The reason is that if firms already screen *imperfectly* for a tested attribute, improved precision has no *intrinsic* implications for relative hiring of different worker group, yet it does unambiguously raise productivity.

The preceding discussion suggests that the trade-off between efficiency and equity in hiring is an empirical possibility rather than a theoretical certainty. To evaluate this trade-off requires a comparison of the hiring and productivity of comparable workers hired with and without employment testing at comparable employers. To our knowledge, there is no prior research that performs this

[2] In an exhaustive assessment, Wigdor and Green (1991) find that military recruits' scores on the Armed Forces Qualification Test (AFQT) accurately predict their performance on objective measures of job proficiency. Similarly, based on an analysis of 800 studies, Hartigan and Wigdor (1989) conclude that the General Aptitude Test Battery (GATB), used by the U.S. Employment Service to refer job searchers to private sector employers, is a valid predictor of job performance across a broad set of occupations. The personnel psychology literature also finds that commonly administered personality tests based on the "five factor model" are significant predictors of employee job proficiency across almost all occupational categories (Barrick and Mount, 1991; Tett, Jackson and Rothstein,1991; Goodstein and Lanyon, 1999).

[3] Not all researchers fail to recognize that this assumption is problematic. Hartigan and Wigdor (1989, chapter 12) critque Hunter and Schmidt's (1982) widely cited analysis of the potential economic gains from job testing, noting that Hunter and Schmidt's results depend upon the unrealistic assumption that absent testing, worker assignment is random.

comparison.[4] In this paper, we empirically evaluate the consequences of private sector applicant testing for minority employment and productivity. We study the experience of a large, geographically dispersed retail firm whose 1,363 establishments switched from informal, paper-based hiring methods to a computer-supported screening process during 1999 and 2000. Both hiring methods use face to face interviews, while the electronic assessment tool also places substantial weight on a computer-administered personality test. We use the rollout of this technology over a twelve month period to contrast contemporaneous changes in productivity and minority hiring at establishments differing only in whether or not they adopted employment testing in a given time interval.

We find strong evidence that testing yielded more productive hires – increasing median employee tenure by 10 percent, and slightly lowering the frequency at which workers were fired for cause. Consistent with a large body of work, analysis of applicant data reveals that minorities and low SES applicants performed significantly worse on the employment test. Had managers initially been hiring unsystematically (i.e., in a manner uncorrelated with the test), simple calculations suggest that testing would have lowered minority hiring by approximately 10 to 25 percent. This did not occur. We find no evidence that employment testing changed the racial composition of hiring at this firm's 1,363 sites. Moreover, productivity gains were uniformly large among both minority and non-minority hires. The combination of uniform productivity gains and a lack of adverse hiring impacts suggests that, prior to the introduction of employment testing, employers were already implicitly screening (albeit imperfectly) for the skills measured by the test.

Our paper is related to a broad theoretical and empirical literature on the economics of worker screening. Key theoretical contributions include Spence (1973), Stiglitz (1975) and Salop and Salop (1976), who analyze models of screening, signaling, and self-selection, and Phelps (1972) and Aigner and Cain (1977), who provide the classic theoretical treatments of statistical discrimination. A number of recent empirical studies assess the role of race in employers' hiring decisions. Altonji and Pierret (2001) develop a dynamic learning model to test for employer statistical discrimination in a longitudinal panel of worker earnings, and find little evidence of race-based statistical discrimination.[5] Holzer, Raphael, and Stoll (2002) analyze the effect of employer-initiated criminal background checks on the likelihood that employers hire black workers and conclude that, in the absence of criminal background checks, employers statistically discriminate against black applicants. Bertrand and Mullainathan (forthcoming) conduct an audit study of employer callback rates for job applications. They find

[4] Although a large literature evaluates the likely impacts of testing on private sector hiring, all studies that we are aware of compare anticipated or actual hiring outcomes using an employment test to a *hypothetical* 'unsystematic hiring' case in which no alternative formal or informal applicant screen is used. As explained above, we view this hypothetical case as unlikely.

[5] See also the closely related learning model by Farber and Gibbons (1996).

that applicants with 'black-sounding' names receive significantly fewer callbacks than applicants with 'white-sounding' names, a result that is potentially consistent with either taste-based or statistical discrimination.[6]

Our analysis is most closely related to studies of ability testing used for military selection. Eitelberg et. al. (1984) provide a comprehensive history of ability testing in the U.S. military and discuss its implications for racial composition.[7] Wigdor and Green (1991) provide the definitive validation study of the Armed Forces Qualification Test (AFQT) as a predictor of soldiers' in-field performance. Closest in spirit to our paper, Angrist (1993) analyzes the impacts of successive increases in the military's AFQT qualification standard on military recruiting, and finds that increases in screening stringency differentially reduce minority enlistment.[8]

Our study differs from the existing literature in several respects. First, distinct from the large literature on the use of testing for military selection and public sector job placement, we study testing at competitive, private sector employers. Second, whereas almost all prior work evaluates the effect of race on hiring in a static employment setting – that is, one where screening policies are fixed – the rollout of testing at the $1,363$ stores in our sample provides a unique opportunity to analyze how the use of testing changes hiring in a previously informal hiring environment. A final unusual feature of our study is that we are able to extend the analysis beyond the hiring phase to evaluate how job testing affects the productivity of hires, as measured by turnover and firing for cause. As we show below, these two outcomes – hiring and productivity – are closely linked theoretically and hence provide complementary evidence on the consequences of job testing for employee selection.

The next section describes our data and details the hiring procedures at the firm under study before and after the introduction of testing. Section (3) offers a model to illustrate how the potential disparate impacts of employment testing on minority hiring and productivity depend on pre-testing hiring practices. Sections (4) and (5) provide our empirical analysis of the consequences of testing for

[6] Giuliano (2003) finds that nonblack managers of establishments of a large service sector firm are disproportionately likely to hire nonblack workers. Using data from the same firm, Levine, Leonard and Giuliano (2003) find that dismissals and quits are also higher if managers and subordinates are not of the same race. In a related vein, Montgomery (1991) provides a theoretical model of the use of job referrals for worker selection, and Fernandez and Fernandez-Mateo (2004) analyze the role of employee referral networks in connecting applicants to desirable jobs.

[7] The United States Military's Alpha literacy exam, initiated during World War I, probably represents the first systematic effort to screen U.S. workers for 'employment.' But it wasn't until World War II that rigorous employment screening first confronted the issue of equality. In 1940, when the Army began screening draftees for the "ability to read and write English at the fourth grade level," Southern Congressmen pressured the military to relax standards. Because Southern blacks failed the literacy test in large numbers, a disproportionate share of Southern whites was inducted (Eitelberg et al., 1984). Prior to 1940, the standard had been "ability to comprehend simple orders in the English language."

[8] A key contrast between Angrist's study and our own lies in how testing changes the hiring environment. In Angrist (1993), the experimental variation comes from changes in screening *stringency*. In our study, the variation comes from changes in screening *precision* with stringency roughly held constant. This allows us to analyze how improvements in the employer's information set affect minority and non-minority hiring.

productivity and hiring. Section (6) concludes.

2 Informal and test-based applicant screening at a service sector firm

We analyze the application, hiring, and employment outcome data of a large, geographically dispersed service sector firm with outlets in 47 continental U.S. states. Our data includes all 1,363 outlets of this firm operating during our sample period. All sites are company-owned, each employing approximately 10 to 20 workers in line positions, and offering near-identical products and services. Line positions account for approximately 75 percent of total (non-headquarters) employment, and a much larger share of hiring. Line job responsibilities include checkout, inventory, stocking, and general customer assistance. These tasks are comparable at each store, and most line workers perform all of them. Line workers are primarily young, ages 18 - 30, and many hold their jobs for short durations. As is shown in the first panel of Table 1, 70 percent of line workers are white, 18 percent are black, and 12 percent are Hispanic. Median tenure of line workers is 99 days, and mean tenure is 174 days (panel B).[9]

Worker screening

Prior to June 1999, hiring procedures at this firm were informal, as is typical for this industry and job type. Workers applied for jobs by completing brief, paper job application forms, available from store employees. If the store had an opening or a potential hiring need, the lead store manager would typically phone the applicant for a job interview and make a hiring decision shortly thereafter. On some occasions, applicants were interviewed and hired at the time of application.

Commencing in June 1999, the firm began rolling out electronic application kiosks provided by Unicru, Incorporated in all of its stores. By June of 2000, all 1,363 stores in our sample were equipped with the technology. This technology supplanted the paper application process. At the kiosk, applicants complete a questionnaire administered by a screen-phone or computer terminal, or in a minority of cases, by a web application. Like the paper application form, the electronic questionnaire gathers basic demographic information such as age, gender, race, education, and prior experience. In addition, applicants sign a release authorizing a criminal background check and a search of records in commercial retail offender databases.

A major component of the electronic application process is a computer-administered personality test, which has 100 items and takes approximately 20 minutes to complete. This test measures five personality attributes that collectively constitute the 'Five Factor' model: conscientiousness, agreeableness, extroversion, openness and neuroticism. These factors are widely viewed by psychologists as core personality traits (Digman, 1990; Wiggins, 1996). The particular test instrument used by this

[9] Means exclude incomplete employment spells. Over 98 percent of the spells in our data are complete.

firm focuses on three of the five traits – conscientiousness, agreeableness and extroversion – which have been found by a large industrial psychology literature to be effective predictors of worker productivity, training proficiency, and tenure (Barrick and Mount, 1991; Tett, Jackson, and Rothstein, 1991; Goodstein and Lanyon, 1999).

Once the electronic application is completed, the data are sent to the vendor of the electronic application system, Unicru Incorporated, for automated processing. Unicru's computers transmit the results of processing (typically within a few minutes) to the store's manager by web-posting, email or fax. Two types of output are provided. One is a document summarizing the applicant's contact information, demographics, employment history and work availability. This is roughly a facsimile of the conventional paper application form. Second is a 'Hiring Report' that recommends specific interview questions and highlights potential problem areas with the application, such as criminal background or self-reported prior drug test failure. Of greatest interest, the report provides the applicant's computed customer service test score percentile, along with a color code denoting the following score ranges: lowest quartile ('red'), second-to-lowest quartile ('yellow'), and two highest quartiles ('green').[10]

Following the employment testing, hiring proceeds largely as before. Store managers choose whether to offer an interview (sometimes before the applicant has left the store) and, ultimately, whether to offer a job. Managers are strongly discouraged from hiring 'red' applicants, and, is shown in Table 2, fewer than 1 percent of all 'red' applicants are hired. Beyond this near-prohibition, managers retain considerable discretion. There are many more applicants than jobs, and only 8.9 percent of applicants are hired: approximately 1 in 11. Even for those who score well above the 'red' threshold, the customer service test score has substantial predictive power for hiring. As shown in panel C of Table 2, hiring rates are strongly monotonically increasing in the test score. Only 1 in 18 of those scoring in the fourth decile (in the 'yellow' range) is hired, relative to 1 in 5 applicants scoring in the highest decile.

Hiring and termination data

Our analysis draws on company personnel records that contain worker demographics (gender, race), hire date, and (if relevant) termination date and termination reason for each worker hired during the sample frame. These data allow us to calculate length of service for employment spells in our sample, 98 percent of which are completed by the close of the sample. We code worker terminations into two groups: neutral terminations and terminations for cause. Neutral terminations include return to school, geographic relocation, or any separation that is initiated by the worker except for job abandonment. Firings for cause include incidents of theft, insubordination, unreliability, unacceptable

[10] An identical paper and pencil personality test could readily have been used in the pre-electronic application hiring regime. Administering and scoring this test manually would have been time-consuming, however.

performance or job abandonment. In addition, we utilize data on applicant's self-reported gender, race (white, black, Hispanic, other), and the zip code of the store to which they applied for employment. We merge these zip codes to data from the 2000 U.S. Census of Populations Summary Files 1 and 3 (U.S. Census Bureau, 2001 and 2003) to obtain information on the racial composition and median household income in each store's location.

An important feature of our analysis is that personnel (but not application) records are available for workers hired prior to implementation of the Unicru system at each store. Hence, we build a sample that includes all line workers hired from January 1999, five months prior to the first Unicru rollout, through May 2000, when all stores had gone online. After dropping observations in which applicants had incompletely reported gender or race, we were left with 34,247 workers hired into line positions, 25,820 of whom were hired without use of testing and 8,427 of whom were hired after receiving the test.[11]

Notably absent from our data are standard human capital variables such as age, education and earnings. Because most line workers at this firm are relatively young and many have not yet completed schooling, we are not particularly concerned about the absence of demographic variables. The omission of wage data is potentially a greater concern. Our understanding, however, is that wages for line jobs are largely set centrally, and the majority of these positions pay the minimum wage. We therefore suspect that controlling for year and month of hire, as is done in all models, should purge much of the wage variation in the data.

Applicant test scores

To analyze test score differences in our sample, we draw on a database containing all applications (214,688 total) submitted to the 1,363 stores in our sample during the one year following the rollout of job testing (June 2000 through May 2001). Although we would ideally analyze applications submitted during the rollout, these records were not retained. In Appendix 2, we demonstrate that applicant test scores from this database are highly correlated with the productivity of workers hired at each store *before and after* the introduction of employment testing (see also Appendix Table 2). This suggests that the applicant sample provides a reasonable characterization of workers applying for work during the rollout period.

As shown in Table 2, there are marked differences in the distribution of test scores among white, black and Hispanic applicants. Mean black and Hispanic test scores are, respectively, 5.4 points and 3.5 points below the mean score of whites. Kernel density comparisons of standardized raw test scores, shown in Figure 1, also underscore the pervasiveness of these differences. Relative to the white test

[11] We closed the sample at the point when all hires were made through the Unicru system. Because the rollout accelerated very rapidly in the final three of twelve months, the majority of hires during the rollout period were non-tested hires. Twenty-five percent of the hires in our sample were made prior to the first rollout.

score distribution, the black and Hispanic test score densities are visibly left-shifted. These racial gaps, equal to 0.19 and 0.12 of standard deviations, accord closely with the representative test data reported by Goldberg et al. (1998).[12] As we show below, these test score gaps are also economically significant.[13]

Before beginning our empirical analysis of these outcomes, we provide a brief conceptual model to explore the conditions under which disparate impacts are likely to occur.

3 When does job testing have disparate impacts?[Preliminary]

How does the introduction of job testing affect the employment opportunities of minority job seekers in a competitive labor market? As discussed in the Introduction, the presumed answer to this question is that testing reduces the labor market opportunities of members of low scoring groups. Here, we present a brief, illustrative model to explore when this presumption is likely to hold. Our conceptual framework is closely related to well known models of statistical discrimination by Phelps (1972), Aigner and Cain (1977), Lundberg and Startz (1984), Coate and Loury (1983) and Altonji and Pierret (2001). The contribution of our model is to analyze how an improvement in the employer's information set – that is, a rise in screening precision – affects the employment opportunities and productivity (conditional on hire) of minority and non-minority workers.[14]

Consider a large set of firms facing job applications from two identifiable demographic groups $x \in \{a, b\}$ that differ only in mean productivity. For simplicity, we assume that a and b applicants each comprise half of the population. Applicants have productivity η_i, which is distributed $\eta \sim N\left(\bar{\eta}_x, \sigma_\eta^2\right)$ with $\sigma_\eta^2 > 0$, identical for a and b, and $\bar{\eta}_a > \bar{\eta}_b$. We can write $\eta = \bar{\eta}_x + \varepsilon_\eta$. Firms in our model have linear, constant returns to scale production technology, a positive discount rate, and are risk neutral. Works produce output $f(\eta_i) = \eta_i$, in flow terms, which is priced at unity. Job spell durations are independent of η and wages are fixed at $\omega < \bar{\eta}_a, \bar{\eta}_b$ (also in flow terms).[15]

[12] Goldberg et al. (1998), using a representative sample of the U.S. workforce, find that conditional on age, education and gender, blacks and Hispanics score, respectively, -0.22 and -0.18 standard deviations below whites on the Conscientious trait. Blacks also score lower on Extroversion and Hispanics lower on Agreeableness (in both cases significant), but these discrepancies are smaller in magnitude.

[13] We explored the robustness of these unconditional comparisons by regressing applicant test scores (in percentiles) on dummy variables for race and gender, month × year of application, and store fixed effects. Conditional on gender and month-year of application, black applicants score 5.5 percentiles below white applicants ($t = 24$). For Hispanics, this gap is 3.6 percentiles ($t = 14$). When store fixed effects are added, the race coefficients decline in magnitude by about 30 percent and remain highly significant, indicating that minority applicants are overrepresented at stores where white applicants have below average scores. We also find that, conditional on race and store-effects, applicants from high minority and low-income zip codes have significantly lower test scores than others.

[14] Masters (2004) provides a theoretical analysis of the impact of culturally-biased testing on the welfare of minority workers.

[15] As above, the majority of line workers at the establishments we study are paid the minimum wage.

Firms in our model do not observe the productivity of individual applicants, η_i. Instead, they observe group membership, $x_i \in \{a, b\}$, and a noisy productivity signal, η_{0i}, with $\eta_{0i} = \eta_i + \varepsilon_0$ where $\varepsilon_0 \sim N(0, \sigma_0^2)$ with $\sigma_0^2 > 0$. We think of η_0 as representing observable applicant attributes, such as attitude, dress and speech, that will not be measured by our data. Job testing in our model provides firms with a second productivity signal, η_1, which is unbiased and is independent of η_0 conditional on η. In particular, $\eta_{1i} = \eta_i + \varepsilon_1$ where $\varepsilon_1 \sim N(0, \sigma_1^2)$ with $\sigma_1^2 > 0$ and $E(\varepsilon_0 \varepsilon_1) = 0$.

Firms in our model employ one worker at a time and search for a replacement when a vacancy opens. While holding a vacancy, firms receive applications drawn at random from the pooled distribution of a and b workers. Firms can choose either to hire the current applicant or to wait a non-zero interval for a new applicant. In this case, the prior applicant becomes unavailable. Since wages are fixed, firms strictly prefer to employ workers with higher η. Because holding a vacancy forfeits potential profits, firms will apply a hiring rule that trades off the costs and benefits of waiting for a superior applicant. As is well understood, this trade-off leads to a hiring threshold: firms hire applicants whose expected productivity exceeds an optimally chosen value, and a constant fraction of worker-firm matches lead to hire. We analyze a reduced form of this setup. In our model, firms select applicants using a hiring threshold, and this produces a constant hire rate of $K > 0$.[16] In a complete model, this hiring threshold would depend on technology and labor market conditions. In our reduced form model, the unconditional hiring probability is held constant at $\Pr(H) = K$. This simplification focuses our analysis on the first-order impacts of job testing on the *distribution* of hiring across applicant types $\{a, b\}$, leaving total employment fixed.[17]

The question asked by our model is: will job testing have a disparate impact on the hiring rates and productivity (conditional on hire) of a versus b workers? As we demonstrate, the answer to this question depends on how firms screen applicants in the absence of testing. To demonstrate the importance of screening practices, we consider three boundary cases that differ only in how firms use available applicant information. The first is unsystematic selection. Here, firms do not act upon – or, equivalently, do not observe – applicant productivity information (that is, η_0 and x). The second practice is what we term 'naive' selection. Here, firms select workers using the error-ridden productivity signal, η_0, but do not adjust for (or do not observe) the additional information conveyed by the applicant's demographic group (x). In the third case, firms statistical discriminate. That is, they combine information from both η_0 and x to form 'rational expectations' for worker productivity.[18]

[16] To reduce the number of cases considered, we also assume that $K < 1/2$. As above, fewer than 1 in 10 applicants at the stores in our sample are hired.

[17] Endogeneizing K in our model would require many additional assumptions that we believe detract from the simple points we wish to underscore.

[18] Note that U.S. employment law does not permit use of protected group membership (i.e., race, sex, age over 40, disability, or union status) as an indicator of productivity. Statistical discrimination is probably difficult to detect,

To provide a metric for disparate impact, let $\psi = \Pr(H|b) - \Pr(H|a)$ equal the expected difference in the hiring rate of a and b applicants, and let $\pi = E(\eta|H,b) - E(\eta|H,a)$ equal the expected productivity difference between a and b hires. We say that job testing has a disparate impact if it systematically alters ψ or π, that is if $E(\Delta\psi) \neq 0$ or $E(\Delta\pi) \neq 0$.

3.1 Unsystematic selection

We begin with unsystematic selection. Since all productivity information is ignored, firms hire a representative subset of all applicants, each with probability K. Hence, hiring is effectively random. (In this case, the notion of a screening 'threshold' does not apply.) Under unsystematic selection, a and b applicants face equal probability of hire: $\psi_u = 0$. The expected productivity gap between a and b hires is equal to the difference in population means: $\pi_u = \bar{\eta}_b - \bar{\eta}_a < 0$. Though unsystematic selection unrealistic, it provides a useful baseline case because it corresponds to the setting primarily considered by the literature on the impact of testing on minority employment (e.g., Hartigan and Wigdor, 1989, and cites therein).[19]

We now consider the introduction of job testing in the unsystematic selection environment. Job testing provides firms with an informative productivity signal, η_1, for each applicant.[20] Per our earlier assumption, firms will apply a selection threshold to the test score, and workers with a value of η_1 exceeding the threshold will be hired. Let κ_u be the selection threshold that solves:

$$\begin{aligned} K &= \frac{1}{2}\left[\Pr(H|x=a) + \Pr(H|x=b)\right] \\ &= \frac{1}{2}\left[\Pr(\eta_1 > \kappa_u|x=a) + \Pr(\eta_1 > \kappa_u|x=b)\right] \\ &= \frac{1}{2}\left[1 - \Phi\left(\frac{\gamma_1(\kappa_u - \bar{\eta}_a)}{\sigma_\eta}\right) + 1 - \Phi\left(\frac{\gamma_1(\kappa_u - \bar{\eta}_b)}{\sigma_\eta}\right)\right], \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, $\sigma_{v1} = \left(\sigma_\eta^2 + \sigma_1^2\right)^{1/2}$ is the standard deviation of the test score, η_1, and $\gamma_1 = \sigma_\eta/\sigma_{\nu 1}$ indexes the precision of the test, expressed on the unit interval.[21]

Since the screening threshold, κ_u, is identical for both applicant groups and average applicant productivity is higher for a than b applicants, the above equation immediately implies that $\Pr(H|x=a)$ falls relative to $\Pr(H|x=b)$ and so $E(\Delta\psi_u) < 0$: testing has a negative disparate impact on b hiring.

however, and so may be commonplace in practice.

[19] Note that we do not need to assume that firms hire unsystematically along *all* dimensions; only that any systematic selection is uncorrelated with η_0 (and, by implication, with a and b).

[20] We continue to assume that other productivity information (η_0, x) is ignored.

[21] Since $\Phi(\cdot)$ is continuous, bounded between 0 and 1, and everywhere decreasing in κ_u, this equation will have a unique solution for κ_u.

Logically, relative to the unsystematic hiring baseline, testing reduces hiring from the less qualified group.

Although aggregate hiring is held constant, testing raises aggregate productivity. We can write the expected productivity of a hired worker from group x as

$$E\left(\eta|x,\eta_1>\kappa_u\right)=\bar{\eta}_x+E\left(\varepsilon_\eta|x,\eta_1>\kappa_u\right)=\bar{\eta}_x+\gamma\sigma_\eta\lambda\left(\frac{\gamma_1\left(\kappa_u-\bar{\eta}_x\right)}{\sigma_\eta}\right), \tag{1}$$

where $\lambda(.)$ is the Inverse Mills Ratio, equal to $\phi(\cdot)/(1-\Phi(\cdot))\geq 0$. This expression decomposes the productivity of hires from each group into two components. The first, $\bar{\eta}_x$, is the expected productivity of a randomly hired applicant from group x. The second term $\gamma_1\sigma_\eta\lambda(\cdot)$ reflects the improvement in selection due to testing. By truncating the lower tail of test-takers (those with $\eta_1<\kappa_u$), testing increases the expected productivity of hires relative to applicants. This improvement is rising in the precision of the test, γ_1, and in the stringency of the threshold (κ_u).[22]

While testing raises the productivity of both a and b hires, the selection effect is not neutral for a versus b productivity. Differentiation of equation (1) demonstrates that testing differentially raises the productivity of b relative to a hires: $\partial E\left(\eta|x,\eta_1>\kappa_u\right)/\partial\bar{\eta}_x=-\gamma_1^2\lambda'(\cdot)<0$. The reason is that testing truncates a relatively larger share of the b distribution and so differentially raises selectivity for this group. Consequently, $E(\Delta\pi_u)>0$.

In brief, introduction of job testing in an unsystematic hiring environment raises the productivity of hires from both groups, reduces the hiring of b relative to a applicants, and raises the productivity of b relative to a hires. Because testing 'systematizes' an unsystematic hiring environment, these effects are first order. Thus, consistent with the large literature on testing and race, improved candidate selection comes at a cost of reduced opportunity for groups with lower average test scores.

3.2 Naive selection

A more plausible setting may be one in which firms hire apply a uniform selection criterion that is blind to demographic characteristics. In this case, firms 'discriminate' on the basis of the productivity information contained in η_0, but they do not use demographics, x, to condition their expectations. Under this assumption, firms assess expected applicant productivity as $E(\eta|\eta_0)=\eta_0$.[23] We refer to this selection rule as 'naive' because a and b applicants with identical signals (η_0) are treated identically although they do not have identical expected productivity.

[22] See Prendergast (1999) for a detailed development of the normal selection equations used here. Also note that $\lambda(\cdot)\geq 0$, $\lambda'(\cdot)\geq 0$.

[23] Hence, 'naive' firms in our model take $\tilde{\eta}$ at face value. A 'quasi-naive' alternative assumption would be that firms calculate $E(\eta|\eta_0)=\Pr(x=a|\eta_0)\cdot E(\eta|\eta_0,x=a)+\Pr(x=b|\eta_0)\cdot E(\eta|\eta_0,x=b)$. That is, they attempt to infer demographic group membership, x, without using the demographic indicator. This alternative complicates the analysis but does not change the fundamental results.

Let $\kappa_n(\gamma_0)$ be the naive selection threshold that solves:

$$K = \frac{1}{2}\left[1 - \Phi\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_a)}{\sigma_\eta}\right) + 1 - \Phi\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_b)}{\sigma_\eta}\right)\right],$$

where $\sigma_{\nu 0} = \left(\sigma_\eta^2 + \sigma_0^2\right)^{1/2}$ and $\gamma_0 = \sigma_\eta/\sigma_{\nu 0}$. We denote $\kappa_n(\gamma_0)$ as explicitly depending upon γ_0 because a change in screening precision, holding K constant, implies a change in κ_n, as we show below.

Under naive selection (and prior to introduction of testing) the hiring rate from each demographic group, x, is:

$$\Pr(H|x) = 1 - \Phi\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_x)}{\sigma_\eta}\right). \tag{2}$$

The expected productivity of group hired workers from each group is:

$$E(\eta|\eta_0 > \kappa_n, x) = \bar{\eta}_x + \gamma_0\sigma_\eta\lambda\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_x)}{\sigma_\eta}\right).$$

Substituting these equations into our measures of relative hiring and productivity gives:

$$\psi_n = \Phi\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_a)}{\sigma_\eta}\right) - \Phi\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_b)}{\sigma_\eta}\right),$$

and

$$\pi_n = (\bar{\eta}_b - \bar{\eta}_a) + \gamma_0\sigma_\eta\left[\lambda\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_b)}{\sigma_\eta}\right) - \lambda\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_a)}{\sigma_\eta}\right)\right].$$

Relative to the unsystematic hiring case, naive hiring yields a lower rate of b relative to a hiring and a smaller (less negative) gap between the productivity of b versus a hires.[24]

We now analyze how job testing changes ψ_n and π_n in the naive hiring environment.

Job testing provides firms with a second applicant productivity signal, η_1.Since both productivity signals, η_0 are η_1, are informative, firms will optimally combine them to assess applicant productivity. The addition of a second signal is identically equal to a rise in screening precision from $\gamma_0 = \left[\sigma_\eta^2/\left(\sigma_\eta^2 + \sigma_0^2\right)\right]^{1/2}$ to:

$$\gamma_2 = \left(\frac{\sigma_\eta^2\left(\sigma_0^2 + \sigma_1^2\right)}{\sigma_\eta^2\sigma_0^2 + \sigma_\eta^2\sigma_1^2 + \sigma_0^2\sigma_1^2}\right)^{1/2}. \tag{3}$$

Here, γ_2 is equal to the population R statistic (i.e., $\sqrt{R^2}$) from a regression (for either demographic group) of η on η_0, η_1 and a constant. We continue to assume that naive firms do not use applicant demographics (x) to form expectations.

We can now assess the disparate impacts of testing by asking if a rise in screening precision systematically alters ψ_n or π_n. Our answer is summarized in the following four propositions:

[24] These equations are identical the case of testing in the unsystematic hiring environment, except that here firms screen on η_0 instead of η_1.

Proposition 1 *A rise in screening precision under naive selection causes a downward adjustment in the screening threshold:* $\partial \kappa_n(\gamma)/\partial\gamma < 0$.

Observe from equation (2) that the hiring odds for each applicant group is declining in screening precision: $\partial \Pr(H|X)/\partial\gamma = -\phi'(\cdot)\left((\kappa_n - \bar{\eta}_x)/\sigma_\eta\right) < 0$. Intuitively, firms engaging in naive selection take observed applicant productivity information, η_0, at face value without adjusting for measurement error. Testing reduces measurement error, and so lowers the fraction of workers whose assessed productivity exceeds a given threshold.[25] Hence, to maintain overall hiring at K, κ_n must decline as screening precision rises.

Proposition 2 *A rise in screening precision under naive selection raises hiring of a relative to b applicants.*

A constant hiring rate implies that

$$\Phi\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_a)}{\sigma_\eta}\right) + \Phi\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_b)}{\sigma_\eta}\right) = \Phi\left(\frac{\gamma_2(\kappa_n(\gamma_2) - \bar{\eta}_a)}{\sigma_\eta}\right) + \Phi\left(\frac{\gamma_2(\kappa_n(\gamma_2) - \bar{\eta}_b)}{\sigma_\eta}\right).$$

Noting that $\gamma_2 > \gamma_0$ and $\kappa_n(\gamma_2) < \kappa_n(\gamma_0)$ (as per the first proposition), this equation implies that $\gamma_2(\kappa_n(\gamma_2) - \bar{\eta}_b) > \gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_b)$ and $\gamma_2(\kappa_n(\gamma_2) - \bar{\eta}_a) < \gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_a)$. Applying these inequalities to the definition of k_n gives:

$$\Phi\left(\frac{\gamma_2(\kappa_n(\gamma_2) - \bar{\eta}_a)}{\sigma_\eta}\right) - \Phi\left(\frac{\gamma_2(\kappa_n(\gamma_2) - \bar{\eta}_b)}{\sigma_\eta}\right) < \Phi\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_a)}{\sigma_\eta}\right) - \Phi\left(\frac{\gamma_0(\kappa_n(\gamma_0) - \bar{\eta}_b)}{\sigma_\eta}\right),$$

which implies that $E(\Delta\psi_n) < 0$.

Hence, as in the unsystematic selection case above, an increase in screening precision generates a disparate negative impact on b hiring. Intuitively, a rise in screening precision reduces hiring of both a and b applicants, ceteris paribus. This differentially reduces hiring of $b's$ since, given their lower mean $\bar{\eta}$, they benefit disproportionately from measurement in the error. (Note that in an extreme case where measurement error in η_1 is unbounded ($\sigma_1^2 \to \infty$) , hiring rates of $a's$ and $b's$ would be identical.) To maintain K constant given higher screening precision, firms lower the hiring threshold from $\kappa_n(\gamma_0)$ to $\kappa_n(\gamma_2)$. This more than fully offsets the reduction in hiring for $a's$ but only partly offsets the loss for $b's$, thereby raising hiring of $a's$ at the expense of $b's$. (Due to the non-linearity of $\Phi(\cdot)$, these offsetting effects cannot 'wash out' for both groups.)

Proposition 3 *A rise in screening precision under naive selection raises the productivity of b relative to a hires.*

[25] Recall that we have assumed that $K > \bar{\eta}_a, \bar{\eta}_b$, and hence κ_n is above the mean of the applicant distribution.

This follows immediately from the fact that hiring of $b's$ falls while hiring of $a's$ rises. Because selectivity of $b's$ has increased while selectivity of $a's$ has decreased, the productivity gap between them must decline: $E(\Delta\pi_n) > 0$.

Proposition 4 *Testing unambiguously raises the productivity of both a and b hires.*

Testing raises the odds that a qualified applicant is hired and that an unqualified applicant is rejected. Consequently, holding total hiring constant, the expected productivity of hires must rise. For a applicants, this gain in productivity is *partly* offset by a rise their aggregate hiring rate. But, as shown in the Appendix 1, the net gain for both $a's$ and $b's$ is positive.

In summary, introduction of testing in a naive selection environment generates disparate impacts comparable in sign than in the unsystematic selection case. However, these effects are necessarily smaller in magnitude than in the unsystematic selection. This reason is that the change in screening induced by a rise in precision is small compared to the change caused by a movement from unsystematic to systematic selection.

3.3 Statistical discrimination

In the prior two cases, firms ignored demographic group membership. Because η_0 is a an error-ridden measure of applicant productivity, however, firms can improve screening precision by also conditioning on demographic group membership – that is, by statistically discriminating.[26] Statistically discriminating firms assess expected applicant productivity as:

$$E(\eta|x,\eta_0) = \bar{\eta}_x + \gamma_0^2(\eta_0 - \bar{\eta}_x),$$

which is equal to a convex combination of the group specific mean, $\bar{\eta}_x$, and the observed applicant signal, η_0, where the weight given to the individual signal is increasing in signal precision, γ_0.

Under statistical discrimination, the hiring gap between a and b applicants will be:

$$\psi_s = \Phi\left(\frac{\kappa_s(\gamma_0) - \bar{\eta}_a}{\gamma_0\sigma_\eta}\right) - \Phi\left(\frac{\kappa_s(\gamma_0) - \bar{\eta}_b}{\gamma_o\sigma_\eta}\right),$$

with productivity gap:

$$\pi_s = (\bar{\eta}_b - \bar{\eta}_a) + \gamma\sigma_\eta\left[\lambda\left(\frac{\kappa_s(\gamma_0) - \bar{\eta}_b}{\gamma_0\sigma_\eta}\right) - \lambda\left(\frac{\kappa_s(\gamma_0) - \bar{\eta}_a}{\gamma_0\sigma_\eta}\right)\right].$$

These terms (ψ_s and π_s) differ from the selection terms for the naive case (ψ_n and π_n) by only one parameter: γ_0, the selectivity term. In the statistical discrimination case, the selectivity term

[26] U.S. employment law does not permit use of protected group membership (i.e., race, sex, age over 40, disability, or union status) as an indicator of productivity. Statistical discrimination is therefore illegal. In pratice, it is probably difficult to detect, however, and so may potentially be commonplace.

appears in the denominator of the selection equations; in the naive screening case, it appears in the numerator. This differences reflects a contrast in how firms use available screening information. Naive firms make no adjustment for measurement error in observed applicant signal (η_1). Consequently, lower precision (more measurement error) reduces selectivity, seen in a reduction in the numerator of the selection equations. By contrast, firms using statistical discrimination discount high and low values of η_0 towards the group specific mean in proportion to the measurement error in the signal. Hence lower precision raises selectivity, seen in a reduction in the denominator of the selection equations.

We now analyze how the introduction of job testing changes ψ_s and π_s in the statistical discrimination environment. The addition of the job test (η_2) is equivalent to a rise in screening precision from γ_0 to γ_2 (see equation (3)). The impacts of rising precision are summarized in the following four propositions:

Proposition 5 *A rise in screening precision under statistical discrimination causes an upward adjustment in the screening threshold:* $\partial\kappa_s(\gamma)/\partial\gamma > 0$.

Opposite to the naive selection case, hiring odds for all applicants are rising in screening precision. Because measurement error causes statistically discriminating firms to discount observed applicants signals towards the group mean, a reduction in measurement error will raise the fraction of applicants whose assessed productivity exceeds κ_s. Therefore, κ_s must rise to maintain overall hiring at K.

Proposition 6 *A rise in screening precision under statistical discrimination raises hiring of b relative to a applicants.*

A constant hiring rate implies that

$$1-\Phi\left(\frac{(\kappa_s(\gamma_0)-\bar{\eta}_a)}{\gamma_0\sigma_\eta}\right)+1-\Phi\left(\frac{(\kappa_s(\gamma_0)-\bar{\eta}_b)}{\gamma_0\sigma_\eta}\right)=1-\Phi\left(\frac{(\kappa_s(\gamma_2)-\bar{\eta}_a)}{\gamma_2\sigma_\eta}\right)+1-\Phi\left(\frac{(\kappa_s(\gamma_2)-\bar{\eta}_b)}{\gamma_2\sigma_\eta}\right).$$

Noting that $\gamma_2 > \gamma_0$ and $\kappa_s(\gamma_2) > \kappa_s(\gamma_0)$, this equation implies that $(\kappa_s(\gamma_2)-\bar{\eta}_b)/\gamma_2\sigma_\eta < (\kappa_s(\gamma_0)-\bar{\eta}_b)/\gamma_0\sigma_\eta$ and $(\kappa_s(\gamma_2)-\bar{\eta}_a)/\gamma_2\sigma_\eta > (\kappa_s(\gamma_0)-\bar{\eta}_a)/\gamma_0\sigma_\eta$. Applying these inequalities to the definition of k_s gives:

$$\Phi\left(\frac{(\kappa_s(\gamma_2)-\bar{\eta}_a)}{\gamma_2\sigma_\eta}\right)-\Phi\left(\frac{(\kappa_s(\gamma_2)-\bar{\eta}_b)}{\gamma_2\sigma_\eta}\right)>\Phi\left(\frac{(\kappa_s(\gamma_0)-\bar{\eta}_a)}{\gamma_0\sigma_\eta}\right)-\Phi\left(\frac{(\kappa_s(\gamma_0)-\bar{\eta}_b)}{\gamma_0\sigma_\eta}\right)\Rightarrow E(\Delta\psi_s)>0.$$

Hence, opposite to the two cases above, an increase in screening precision at statistically discriminating firms differentially benefits b applicants. The reason is that 'discounting' applicant signals towards their group-specific means is particularly harmful to hiring of qualified b group applicants.[27]

[27] Concretely, consider two applicants with identical ability $\eta > \kappa_s$ but different group memberships. So long as there is measurement error in η_0, the b group applicant will be less likely to be hired.

As testing reduces measurement error, it differentially benefits b group applicants. This gain is partly offset by a compensatory rise in the screening threshold (to maintain K constant). But the net effect on b relative to a hiring is unambiguously positive.

Proposition 7 *A rise in screening precision under statistical discrimination lower the productivity of b relative to a hires.*

Analogous to Proposition 3 above, this prediction follows directly from the fact that testing raises hiring of b and lowers hiring of a applicants. Because selectivity of $a's$ has risen while selectivity of $b's$ has fallen, the productivity gap between them must rise: $E(\Delta\pi_s) < 0$.

As with prior cases:

Proposition 8 *Testing unambiguously raises the productivity of both a and b hires.*

See Appendix 1.

Hence, statistical discrimination reverses the predictions of the other two cases (unsystematic and naive selection): if firms have rational expectations, testing does not harm – and may in fact improve – the job prospects of members of low scoring groups. As with the naive case above, these distributional effects are small relative to the first order impact of introducing testing in an unsystematic hiring environment.

3.4 Implications

Only one unambiguous conclusion emerges from the above analysis: testing raises productivity. By contrast, the widely held presumption that testing reduces hiring of applicants from low scoring groups is supported only if firms do not already use available screening information optimally. If firms statistically discriminate initially, a gain in screening precision has the potential to benefit applicants from low scoring groups. If they do not, a gain in precision, may slightly reduce hiring from minority groups. The only case in which large disparate impacts are a certainty is one in which hiring in the pre-test environment is entirely uncorrelated with the test measure (e.g., unsystematic selection). Outside of this case, the productivity gains from testing accrue primarily from better selection within applicant groups rather than substantial shifts in cross-group hiring practices.

Though our model makes many specific assumptions, we view the ambiguity of the results to be quite general; if firms already screen imperfectly for a tested attribute, improved precision has no *intrinsic* implications for relative well-being of different worker groups. To be clear, one can readily construct cases where disparate impacts occur (in either direction). But these cases depend sensitively on assumptions about the shape of applicant distributions or the relative precision of testing across

groups. Based on our analysis, we conclude that there is no a priori presumption that testing will have a disparate impact on employment or productivity of applicants from low-scoring groups.

4 Estimating the productivity consequences of job testing

We begin our empirical analysis by studying the productivity consequences of job testing. As an initial productivity measure, we analyze the length of completed job spell durations of workers hired with and without use of job testing. We think of job spell duration as a proxy for reliability; unreliable workers are likely to quit unexpectedly or be fired for poor performance.[28] In section (4.2), we also consider a second productivity measure: firing for cause.

We initially estimate the following difference-in-difference model for job spell duration:

$$D_{ijt} = \alpha + X_{ijt}\beta_1 + \beta_2 T_{ijt} + \psi_t + \varphi_j + e_{ijt}. \tag{4}$$

In this equation, the dependent variable is the job spell duration (in days) of worker i hired at site j in year and month t. The vector X contains worker race and gender, and T is an indicator variable equal to 1 if the worker was screened via job testing, and 0 otherwise. The vector ψ contains a complete set of month-by-year of hire effects to control for seasonal and macroeconomic factors affecting turnover. Most specifications also include a complete set of store site effects, φ, which absorb fixed factors affecting job duration at each store. Since outcomes may be correlated among workers at a given site, we use Huber-White robust standard errors clustered on store and application method ($T = \{0, 1\}$).[29]

Estimates are found in Table 3. The first estimate excludes both site effects and the T indicator variable. Consistent with the bivariate comparisons in Table 1, black and Hispanic workers have substantially lower conditional mean tenure than white employees. When 1,363 site fixed effects are added in column 2, these race differences fall by approximately 40 percent (though they remain highly significant), indicating that minority workers are overrepresented at establishments where both minorities and non-minorities have high turnover.

Columns 3 and 4 present initial estimates of the impact of testing on job spell duration. In column 3, which excludes site effects and race dummies, we find that the employment spells of tested hires are 8.8 days longer than those of non-tested hires ($t = 2.0$). When site fixed effects are added in column 4, the point estimate rises to 18.8 days ($t = 4.6$).[30] Adding controls for worker race and gender has

[28] Stores of this firm are typically staffed leanly, with 2 to 4 line workers per shift. Unreliable workers and those who quit unexpectedly inconvenience customers by reducing staff availability and impose costs on managers and coworkers who must cover their shifts.

[29] Ninety-eight percent of employment spells that commenced during the sample window of January 1999 to May 2000 were completed by the last observation date in our personnel data (August 2003). We exclude incomplete spells from these OLS models.

[30] The flow of hires in our sample intrinsically overrepresents workers hired at high-turnover stores (relative to the stock

little impact on the magnitude or significance of the job-test effect. When we include state × time interactions in column 6 to account for differential employment trends by state, the job-test point estimate rises slightly to 22.1 days.

In net, these models suggest that testing increased mean job duration by approximately 20 days, or 12 percent.[31] This pattern is also clearly visible in Figure 3, which plots the distribution of completed job spells of tested and non-tested hires. The distribution of spells for tested hires lies noticeably to the right of that for non-tested hires, and generally has greater mass at higher job durations and lower mass at shorter durations.

Instrumental variables estimates

Our estimates could be biased if job-test status is endogenous. This endogeneity might take two forms. A first concern is that we observe in our data that in the 1 to 2 months following the rollout of testing at a site, 10 to 25 percent of new hires are not tested. There are three reasons why this may occur. First, individuals who apply prior to the advent of testing are often not on the payroll for several weeks; they will appear as non-tested, post-testing hires in our data. Second, operational and training issues in the weeks following the Unicru installation may cause the online application system to be unavailable or unused. Third, managers might deliberately circumvent testing to hire preferred candidates.[32]

To purge the possible endogeneity of tested status among hires at a store using the test, we re-estimate equation (4) using a dummy variable indicating store-test-adoption as an instrumental variable for the tested status of all applicants at the store. Since we do not know the exact installation date of the electronic application kiosk at a store, we use the date of the first observed tested hire to proxy for the rollout date. First stage estimates of this equation are found in Appendix Table 1. The coefficient on the store-adoption dummy in the first stage equation of 0.89 ($t = 111$) indicates that once a store has adopted testing, the vast majority of subsequent hires are tested.

Instrumental variables estimate of the effect of testing on job spell durations in panel B of Table 3 are approximately 80 percent as large the OLS estimates and are nearly as precisely estimated. In fact, we cannot reject the hypothesis that IV and OLS estimates are identical. This suggests that the potential endogeneity of tested status within stores is not a substantial source of bias.[33]

A second source of concern is that a store's use of testing may be correlated with potential out-

of hires). Hence, when testing is introduced, a disproptionate share of tested hires are at high turnover establishments. Adding site effects to the model controls for this source of composition bias, which substantially raises the point estimate on the job testing variable (compare columns 3 and 4).

[31] Models that include a full set of state × month-year-of-hire interactions (17×47 dummies) yield nearly identical (and quite precise) point estimates.

[32] Changes to the Unicru system implemented after the close of our sample window effectively barred such overrides.

[33] The fact that IV point estimates are smaller than OLS estimates implies that non-tested hires at stores using testing had below average job duration relative to other *non-tested* hires. This is consistent with some managerial subversion.

comes. Although all stores in our sample adopt testing during our sample, the timing of adoption is not necessarily entirely random. To the best of our understanding, the rollout order of stores was determined by geography, technical infrastructure, and internal personnel decisions. It is this last factor that is of concern. If, for example, stores adopted testing when they experienced a rise in turnover, mean reversion in the length of employment spells could cause us to overestimate the causal effect of testing on workers' job spell durations.

As a check on this possibility, we augmented equation (4) for job spell duration with leads and lags of test adoption. These models, found in Appendix Table 2, estimate the trend in job spell durations for workers hired at each store in the 9 months surrounding introduction of testing: 5 months prior to 4 months post adoption. If job spell durations rose or fell significantly prior to test adoption, the lead and lag models would make this evident.

As shown in the appendix table, the lead estimates are in no case significant and, moreover, do not have consistent signs. By contrast, the lag (post-rollout) dummies show striking evidence of a discontinuous rise in job duration for workers hired immediately after testing was adopted. Workers hired in the first month of testing have 14 days above average duration; workers hired in subsequent months have 19 to 28 days above average duration (in all cases significant). These results indicate that our main estimates above are not confounded by pre-existing trends in job spell duration.[34]

Quantile regressions

Since employment duration data are typically right-skewed, our results could also be driven in part by outliers. As a check on this possibility, Panel A of Table 4 presents quantile (least absolute deviation) regression models for job duration. In these models, we retain the 2 percent of observations in which the job spell had yet to be completed by the end of the sample (August 2003). Since it is not feasible to estimate a large number of store fixed effects in quantile regression models, we instead include 46 state dummies.

The regression estimates for median job spell duration confirm that testing increased the length of job spells. In the models in panel A, we find that testing increased median tenure by 8 to 9 days, which is roughly a 10 percent increase (see Table 1), comparable in effect size to the OLS models. Panel B provides estimates for job spell length at percentiles 10, 25, 50, 75, and 90.[35] The impact of testing on completed tenure is statistically significant and monotonically increasing in magnitude from the 10^{th} to the 75^{th} percentiles. We find no effect at the 90^{th} percentile. In net, these results

[34] As an additional robustness test, we estimated a version of equation (4) augmented with separate test-adoption dummies for each cohort of adopting stores, where a cohort is defined by the month and year of adoption. These estimates find a positive effect of testing on job spell duration for 9 of 12 adopter cohorts, 6 of which are significant at $p < 0.05$. By contrast, none of the 3 negative point estimates is close to significant.A table of estimates is available from the authors.

[35] We exclude incomplete spells since some are at very high percentiles.

provide robust evidence that job testing raised worker tenure.

4.1 Did testing have a disparate impact on productivity?

Tables 1 reveals that, prior to the use of job testing, Hispanic and especially black workers had substantially shorter mean job durations than whites. Job testing could potentially affect this gap. As our model indicates, unless firms were statistically discriminating in the pre-testing regime, an increase in screening precision is predicted to differentially raise the productivity of minority relative to non-minority hires (a disparate impact). We analyze here whether this occurred. Before doing so, we calculate an upper bound on the plausible magnitude of this impact.

Consider a hypothetical case where, prior to testing, screening was uncorrelated with the test score. We refer to this as the 'unsystematic selection benchmark.'[36] Panel A of Table 2 shows that among tested applicants, the black-white test score gap was 5.4 points (47.7 versus 53.1 points). Under the 'unsystematic selection' benchmark, we assume that this gap would have carried over into the hired sample in its entirety. By contrast, Panel B of Table 2 shows that among tested *hires*, the black-white test score gap was only 1.5 points. Hence, relative to the benchmark, hiring using the test reduced the black-white test score gap among hires by 3.9 points. The analogous figure for Hispanic hires is 2.9 points. These gains (3.9 and 2.9 points) place an upper bound on the degree to which testing could plausibly have compressed the minority/non-minority test score gap among hires.

To translate this point difference into a productivity difference, we use the job applicant database summarized in Table 2 to estimate the relationship between applicant test scores and job spell durations. As noted, test scores are not available for applications submitted to the stores in our sample *prior to* the use of testing. In their place, we use job applications submitted in the year after the rollout of employment testing (June 2000 through May 2001). Assuming that applicant characteristics did not change systematically after testing was initiated, these data provide a rough measure of the average characteristics of stores' applicants in the period prior to testing. (Supporting evidence for this assumption is given in Appendix 2.)

Column 1 of Table 5 provides an estimate of the following regression model for job spell durations:

$$D_{ijt} = \alpha + X_{ijt}\beta_3 + \beta_4 \bar{S}_j + \psi_t + e_{ijt}. \tag{5}$$

In this equation, the dependent variable is the completed job spell duration of workers hired at store j *prior to* the use of testing, and $\bar{S}_j$ is the average test score of store $j's$ applicants. Control variables for gender, race, year-month of hire and state are also included. Our expectation is that $\hat{\beta}_4 > 0$: stores

[36] While this case is unlikely, the evidence above that testing significantly raised productivity indicates that the initial screen could not have been perfectly correlated with the test.

that had higher quality applicants (as measured by the test score) should have had longer mean job spell durations prior to the use of testing.[37]

This expectation is confirmed in Table 5. The coefficient of 2.73 ($t = 5.0$) on the mean test score variable indicates that, conditional on race, gender, time and state effects, stores facing applicant pools with below average mean test scores had significantly shorter job spells: a one point lower mean test score is associated with approximately 3 fewer days mean job duration for workers hired prior to the use of testing. The economic magnitude of this relationship is large. A one-standard deviation (3.7 point) difference in average store-level test scores predicts a 10 day difference in mean job duration.[38]

We can calculate an upper bound on expected disparate productivity impacts of testing by using this regression estimate. Under the unsystematic selection benchmark, we calculated that testing could potentially have closed the black-white test gap by 3.9 points. Scaling by $\hat{\beta}_4$, this implies a potential 11 days narrowing of the job duration gap between black and white hires. This is a sizable effect, equal to one third of the initial gap of 33 days (Table 5, column 1). An analogous calculation for Hispanic hires yields a potential disparate impact of 8 days on a baseline of 7 days, i.e., full convergence. Hence, under the null of unsystematic selection, job testing had the potential to substantially raise the tenure of minority relative to non-minority hires.

To assess whether this occurred, we estimate in Table 6 a set of job spell duration models performed separately by race. These estimates provide remarkably little evidence of disparate impacts. The point estimate for the effect of testing on mean job duration is 20 days for whites, 23 days for blacks, 19 days for males and 20 days for females. All are significant. Only for Hispanic hires (the smallest sub-group in our sample) is the point estimate of differing magnitude: 8 days, and insignificant.[39]

The second panel of Table 6 presents analogous IV models for job spell duration by race where tested status is instrumented with a dummy variable indicating the store has adopted job testing. As with earlier models, the instrumental variables point estimates are about 80 percent as large as comparable OLS estimates and are only slightly less precisely estimated. In this case, gains in job duration for whites are estimated to be slightly larger than for blacks. In summary, these results provide little evidence that testing had a disparate impact on the productivity of minority relative to non-minority hires.[40]

[37] We do not estimate equation (5) for job spell durations of *tested* hires since selection on the test score would be expected to attenuate estimates of β_4. As shown in Appendix 1, this relationship is positive in the tested sample ($\hat{\beta}_4 > 0$). But, as expected, it is substantially attenuated relative to the non-tested sample.

[38] As shown in Appendix Table 3, this relationship is also robust to inclusion of other demographic and regional controls, including log median income and minority resident share in the store zip code.

[39] A potential explanation for why the gains were smaller for Hispanic hires than other groups is that the test was initially only offered in English.

[40] The overall rise in tenure of 19 to 22 days (Table 3, columns 5 and 6) implies that the use of test-based screening was equivalent to a rise of 7 to 8 points in the average test scores of hires. If the firm was initially hiring unsystematically,

4.2 A second productivity measure: Firing for cause

To supplement the job duration evidence above, we explore a second dimension of worker productivity: firing for cause. Using linked personnel records, we distinguish terminations for cause – theft, job abandonment, insubordination – from neutral or positive terminations, such as return to school, relocation, or new employment. To provide an outcome measure that is uniformly defined across workers at different points in their employment spells, we measure employment status at 180 days following hire. We code three mutually exclusive categories: employed, neutral termination, and terminated for cause.[41] As shown in the first panel of Figure 4, two-thirds of job spells have ended at 180 days following hire, and 22 percent of spells have resulted in termination for cause.

To compare termination outcomes of tested and non-tested workers, we estimate the following linear probability model for employment status at 180 days:

$$E\left[1\left\{O_{ijt}^{180}=k\right\}\right]=\alpha+X_{ijt}\beta_5^k+\beta_6^k T_{ijt}+\psi_t^k+\varphi_j^k, \tag{6}$$

where $1\{\cdot\}$ is the indicator function and k corresponds to each of the three potential employment outcomes (O): employed, neutral termination, termination for cause. So that coefficients may be read as percentage points, the dependent variable is multiplied by 100. The coefficient of interest, β_6, estimates the conditional mean difference in the probability of each outcome for tested relative to non-tested hires.

Table 7 contains estimates. The first specification, which excludes the job testing dummy variable, indicates that 180 days after hire, minority workers are substantially more likely than non-minorities to have been fired for cause. As with the racial differences in mean tenure, these discrepancies are large. Relative to whites, black and Hispanic workers are, respectively, 9 and 3 percentage points (47 percent and 15 percent) more likely to have been terminated for cause within the first 180 days of hire.

Column 2 contrasts employment outcomes of tested relative to non-tested hires. At 180 days following hire, tested workers are 4.4 percentage points (14 percent) more likely than are non-tested workers to remain employed, 3.1 percentage points (6.7 percent) less likely to have received a neutral termination, and 1.4 percentage points (6.5 percent) less likely to have been terminated for cause. The first two point estimates are highly significant; the third is marginally significant ($t = 1.5$). As shown in Column 3, instrumental variables estimates for these models (using store test-adoption as an instrument) show comparable effects. Hence, tested hires appear to have better termination outcomes across the board.

this rise would have been fully 21 points (using the average scores of hires minus the average scores of applicants in Table 2). Clearly, testing improved screening, but screening was far from unsystematic initially.

[41] Results are similar if we use 120 or 240 days instead. Workers terminated for cause are ineligible for rehire.

The large racial differences in termination outcomes evident in Column 1 again underscore that job testing has the potential to generate disparate impacts by raising minority relative to non-minority productivity. We can benchmark the possible magnitude of these impacts using the procedure above. As shown in Panel B of Table 5, stores facing applicant pools with below average mean test scores had significantly higher rates of termination for cause: a one point lower mean applicant test score was associated with a 0.41 percentage point higher share of workers terminated for cause within 180 days. Under the 'unsystematic selection' benchmark, we calculate that the use of job testing would be expected to compress the black-white termination-for-cause gap by 1.6 percentage points (0.41×3.9), and the Hispanic-white termination-for-cause gap by 1.2 percentage points (0.41×2.9). These reductions are substantial, equal to 18 to 40 percent of the baseline difference in termination rates.

We find no evidence of a disparate impact of testing on terminations, however. As shown in panel B of Table 6, the point estimates imply that testing reduced termination rates – both neutral terminations and firings for cause – by roughly equal amounts for workers of all three race groups.

In net, our results indicate that job testing improved worker selection, leading to longer job spell durations and a reduction in the frequency of firing for cause. Most important for our analysis, we find no evidence of disparate impacts; productivity gains were uniformly large for minority and non-minority hires. In light of our theoretical framework, this suggests that firms may have held rational expectations in the pre-testing hiring regime – that is, they accurately accounted for expected productivity differences when selecting applicants. In this case, our model suggests that disparate impacts on minority hiring are likely to be small.

5 The impact of employment testing on hiring

5.1 Unsystematic selection baseline

We now assess whether testing had a disparate impact on minority hiring. Before doing so, we benchmark the potential magnitude of this impact. As shown in Table 2, there are significant differences in test scores among black, white and Hispanic job applicants. Figure 5, which plots locally weighted regressions of hiring rates on test scores (conditioning on store effects and application year $\times$ month), shows that, for applicants of all race groups, the probability of hire is strongly monotonically increasing in the test score. The overall hire rate is 8.9 percent, but applicants who score one standard deviation below the mean have essentially zero probability of hire, while those who score one standard deviation above the mean have a 12 to 15 percent probability of hire.[42] The importance of test scores

[42] We also estimated linear probability models for hiring odds as a function of test score, store effects, time effects, and race and gender. We estimate that a one standard deviation (20 point) increase in the test score raises an applicant's

for hiring is also visible in Figure 6, which plots the distribution of test scores for applicants who were subsequently hired. In contrast to the test score distributions for job applicants shown in Figure 1, the race difference in test scores among job hires is negligible. This suggests that race differences in test scores could have significant disparate impacts on hiring.

To benchmark these impacts, we again consider an unsystematic selection baseline. Using the data for white applicants exclusively, we estimate the following linear probability model for hiring:

$$E\left(H_i\right) = \sum_{n=1}^{100} \pi_n \times 1\left\{S_i = n\right\}. \tag{7}$$

Here, the dependent variable is an indicator equal to 1 if applicant i was hired, S is the applicant's test score and $1\{\cdot\}$ is the indicator function. The coefficients, π_n, estimate the hire rates for white applicants at each test score percentile.[43] We can apply this coefficient vector to the test score distribution for each race group to calculate predicted hiring rates on the assumption that firms use the same selection rules for all applicants. These predicted rates are 10.2 percent for white applicants (equal to the white mean by construction), 8.8 percent for black applicants and 9.3 for Hispanic applicants.[44] These race gaps in predicted hiring rates are sizable. If hiring was initially uncorrelated with the test, testing would cause the black hire rate to fall by 2.5 percentage points (25 percent) and the Hispanic hire rate by 1 percentage point (10 percent). As we show below, disparate impacts of this magnitude are detectable in our sample. We now assess if they occurred.

5.2 Evidence on disparate hiring impacts

As shown in Panel A of Table 1, simple mean comparisons of minority employment before and after the use of testing suggest that job testing had little effect on minority hiring. In fact, the employment share of white workers fell roughly 4.5 percentage points in the year following the introduction of testing. This uncontrolled comparison could potentially mask within-store shifts against minority hiring, however.

To rigorously assess the effect of testing on racial composition, it is useful to derive a link between the hiring rates observed in the data and the underlying parameters of interest, which is the effect of

hiring probability by 4.6 percentage points ($t = 67$). Given a baseline hiring rate of 9 percent, this is a large effect. A table of estimates is available from the authors.

[43] When estimating π, we also control for site effects. This has little effect on the results.

[44] As is visible in Table 2 panel C, observed hiring rates for tested black and Hispanic applicants are in fact *lower* than the predicted rates. This discrepancy is also suggested by Figure 5 where, conditional on test scores, minority applicants are generally less likely to be hired than non-minorities. Although this discrepancy could potentially be explained by taste-based discrimination, our model also predicts this pattern. During job interviews, firms will observe applicant characteristics that are not visible in our data, such as dress, comportment, and maturity. These observables are represented by $\tilde{\eta}$ in our model. Provided that $\tilde{\eta}$ is unbiased, our model immediately implies that minority applicants will have weaker observables than non-minority applicants conditional on their test scores: $E(\tilde{\eta}|\hat{\eta} = k, x = b) < E(\tilde{\eta}|\hat{\eta} = k, x = a)$. (A proof is available on request.) Hence, our model implies that minority applicants will have a lower hire rate than non-minorities conditional on their scores, which is what we observe in Figure 5.

testing on hiring odds for minority applicants. The data allow us to observe the race of new hires, which we express as $\Pr(B|H,A)$, that is the probability that a new worker is black given that he applied (A) and was hired (H). Using Bayes rule, we can write the following identity for the black/non-black (B/NB) hiring odds ratio:

$$\ln\left(\frac{\Pr(B|H,A)}{\Pr(NB|H,A)}\right) = \ln\left(\frac{\Pr(H|B,A)\cdot\Pr(B|A)}{\Pr(H|A)}\right) - \ln\left(\frac{\Pr(H|NB,A)\cdot\Pr(NB|A)}{\Pr(H|A)}\right), \quad (8)$$

Rearranging, we obtain,

$$\ln\left(\frac{\Pr(B|H,A)}{\Pr(NB|H,A)}\right) = \ln\left(\frac{\Pr H|B,A}{\Pr(H|NB,A)}\right) - \ln\left(\frac{\Pr(B|A)}{\Pr(NB|A)}\right). \quad (9)$$

This equation indicates that the odds that a newly hired worker is a minority depend on the hiring odds for minority versus non-minority applicants and the relative application rates of minorities and non-minorities.

Our empirical question concerns how testing affects the hiring odds for minorities. The second term in equation (9) – the minority application rate – is a confounding variable that we would like to eliminate. The lack of data on the composition of job applicants prior to the introduction of testing is therefore a point of some concern. Although we have no evidence suggesting that testing altered the racial composition of applicants, we also cannot offer evidence against this hypothesis.[45] One might speculate, for example, that because the computerized application requires applicants to submit a social security number and authorize a criminal background check, this could differentially discourage minority applicants.[46] If so, this would bias our results towards finding that job testing reduced minority hiring – which is not what we find.

As an empirical analog to equation (9), consider the following conditional ('fixed-effects') logit model:

$$E\left(B_{ijt}|H_{ijt},A_{ijt},T_{ijt},\psi_t,\varphi_j\right) = F\left(\psi_t+\varphi_j+\beta_7 T_{ijt}\right), \quad (10)$$

where B indicates that a hired worker is black, the vectors φ and ψ contain a complete set of store and month-by-year of hire dummies, and $F(.)$ and is the cumulative logistic function. The coefficient, β_7, measures the impact of job testing on the log odds that a newly hired worker is black. Without further assumptions, β_7 captures the combined impact of testing on both relative application rates *and* hiring odds by race. If we assume that minority application rates are roughly constant within stores, these will be eliminated by the store fixed effects, φ. In this case, β_7 captures the impact of testing on hiring odds by race, which is the parameter of interest.

[45] Unicru personnel interviewed for this research believe that application kiosks are enjoyable to use and hence yield more applicants.

[46] Petit and Western (2004) estimate that, among men born between 1965 and 1969, 3 percent of whites and 20 percent of blacks had served time in prison by their early thirties.

To avoid the incidental parameters problem that arises when estimating a maximum likelihood model with a very large number of fixed effects $(1,363)$, we estimate equation (9) using a conditional logit model. This estimator effectively 'conditions out' time-invariant store-specific factors, which include, by assumption, relative minority/non-minority application rates.

The top panel of Table 8 reports estimates of equation (10) for the hiring of white, black and Hispanic workers. These models yield no evidence that employment testing affected relative hiring odds by race. In all specifications, the logit coefficient on the job testing dummy variable is small relative to its standard error ($z < 1$), and its magnitude is economically insignificant. The estimated impact of testing on the hiring probability of blacks and Hispanics is -0.3 and -0.2 percentage points, respectively.[47]

As a robustness test for the conditional logit estimates, we also fit a simple fixed-effects, linear probability model of the form:

$$E\left(B_{ijt}|H_{ijt}, A_{ijt}\right) = \alpha + \beta_8 T_{ijt} + \psi_t + \varphi_j. \tag{11}$$

This model contrasts the share of hires by race at each store among tested and non-tested hires. Although the linear model is technically misspecified for this problem, it may provide more power to detect a small change in the racial composition of hires.

Panel B of Table 8 contains estimates of equation (11) where the dependent variable is multiplied by 100 so that coefficients may be read as percentage points. In all cases, the impact of testing on hiring rates by race is precisely estimated and close to zero. The point estimates imply that testing raised white hire rates by 0.5 percentage points and reduced black and Hispanic hiring rates by 0.2 and 0.1 percentage points.[48] None of these effects are significant. The third panel of Table 8 performs instrumental variable versions of these estimates, using stores' adoption of testing as an instrument for applicants' tested status. These IV estimates are similar to the corresponding OLS models.

Earlier, we calculated that testing could potentially lower the hiring rate of black and Hispanic applicants by 2.5 and 1.0 percentage points respectively. Table 8 strongly suggests that this did not occur: we can reject disparate impacts of this magnitude with well over 99 percent confidence.

5.3 Disparate hiring impacts: A second test

Since these results are central to our conclusions, we test their robustness by analyzing a complementary source of variation. As we show below, there is a tight link between the neighborhoods in which

[47] Marginal effects are calculated as $\partial \Pr\left(H\right)/\partial T = \Pr\left(H\right) \cdot \left(1 - \Pr(H) \cdot \beta_7\right.$.

[48] Point estimates for these three categories do not sum to zero since there is a small number of 'other' race workers in the sample.

stores operate and the race of workers that they hire: stores in minority and low-income zip codes hire a disproportionate share of minority workers. We can use this link to explore whether the introduction of testing systematically changed the relationship between stores' neighborhood demographics and the race of hires. Specifically, we estimate a version of equation (11) augmented with measures of the minority share or median income of residents in the store's zip code, calculated from the 2000 U.S. Census. We first estimate this model separately for tested and non-tested hires at each store (excluding site effects) to assess the cross-sectional relationship between zip code characteristics and the race of hires. We next test formally if job testing changed this relationship.

Table 9 contains estimates. Column 1 of the first panel documents a close correspondence between the race of neighborhood residents and the race of hires. The coefficient of -86.8 ($t = 38$) on the non-white residents variable indicates that, prior to the use of testing, a store situated in an entirely non-white zip code would be expected to have 88 percent non-white hires. Column 2 shows the analogous estimate for tested hires. The point estimate of -85.6 indicates that the relationship between store location and worker race was little changed by employment testing.

Columns 3 and 4 make this point formally. When we pool tested and non-tested hires and add an interaction between the test dummy and the share of non-white residents in the zip code, the interaction term is close to zero and insignificant. When site dummies are added in column 4 – thus absorbing the main effect of zip code share non-white residents while retaining the interaction term – the interaction term is again close to zero. Subsequent columns, which repeat this exercise for black and Hispanic hires, confirm these patterns.

Panel B performs analogous estimates for the racial composition of hires using neighborhood household income in place of zip code minority share. In the pre-testing period, stores in more affluent zip codes had a substantially larger share of white employees; 10 additional log points in neighborhood household income is associated with a 3.2 percentage point higher share of white hires. Employment testing does not appear to have altered this link. For all race groups, and for both measures of neighborhood demographics, the pre-post change in the relationship between neighborhood characteristics and the race of hires is insignificant.

In net, despite sizable racial differences in test scores, we find no evidence that job testing had disparate racial impacts on hiring at the $1,363$ stores in our sample. This evidence concords with our earlier finding that testing did not differentially raise productivity of minority hires. As underscored by our model, if prior to testing, screening was blind to the information revealed by the test, disparate impacts on both hiring and productivity are likely. The fact that neither type of disparate impact occurred strongly suggests that prior to testing, firms in our sample had 'rational expectations' – that is, they statistically discriminated. The fact that firms had rational expectations does not imply,

however, that the screening provided by the test was redundant; the fact that productivity rose proves otherwise. Rather, it suggests that testing raised productivity by improving selection within observable race groups. Between group differences – while sizable – were already implicitly taken into account by the informal screen.

6 Conclusion

An influential body of research concludes that the use of standardized tests for employment screening poses an intrinsic equity-efficiency trade-off; raising productivity through better selection comes at a cost of screening out minority applicants. This inference rests on the presumption that in the absence of standardized tests, employers do not already account for expected productivity differences among applicants from different demographic groups. Accordingly, a test that reveals these differences will disproportionately reduce hiring (and improve productivity) of workers from low-scoring groups. In a competitive hiring environment, however, this may not be the most relevant case. If, absent testing, employers already account imperfectly for expected productivity differences among applicant groups, it is possible for employment testing to improve selection without adversely affecting equity. The reason is that the gains from testing may primarily accrue from selecting better candidates within applicant groups rather than from reducing hiring of groups with lower average scores.

We studied the evidence for an equality-efficiency trade-off in employment testing at a large, geographically dispersed retail firm whose $1,363$ stores switched over the course of 12 months from informal, paper-based hiring to a computer-supported screening process that relies heavily on a standardized personality test. We found that the move to employment testing increased productivity at treated stores, raising mean and median employee tenure by 10 percent, and slightly lowering the frequency of terminations for cause. Consistent with expectations, minority applicants performed significantly worse on the employment test. Had the pre-testing hiring screen been 'blind' to the expected productivity differences revealed by the test, we calculated that employment testing would have reduced minority hiring by approximately 10 to 25 percent. This did not occur. We found no evidence that employment testing changed the racial composition of hiring at this firm's $1,363$ sites. Moreover, productivity gains were equally large among minority and non-minority hires. The combination of uniform productivity gains and no disparate hiring impacts suggests that employers were effectively statistically discriminating prior to the introduction of employment testing. Consequently, the gain in improved selection came at no measurable cost in equity.

Several caveats apply to these results. First, our data are from only one large retailer. Since retail firms in the U.S. operate in a competitive environment, we might anticipate that other firms would respond similarly. However, analysis of other cases is needed before general conclusions can be drawn.

A second caveat is that the between group differences found by the employment test used at this firm are not as large as differences found on other standard ability tests, such as the Armed Forces Qualification Test. An alternative employment test that revealed larger group productivity differences might potentially generate disparate impacts. Although we do not discount this possibility, there are two reasons to believe it is not a first order concern. First, we generally expect that employers *will* account for expected group productivity differences; hence, a test that reveals large disparities on some measure should not necessarily generate large surprises. Second, employment testing guidelines issued by the Equal Employment Opportunity Commission make it difficult, and potentially risky, for firms to utilize employment tests that 'pass' minority applicants at less than 80 percent of the pass-rate of non-minority applicants.[49] We therefore do not expect typical employment tests to show greater group differences than those found here.

A final caveat in interpreting our results is that they speak only to firms' *private* gains from improved worker selection. The extent to which these private gains translate into social benefits depends largely on the mechanism by which testing improves selection. If testing improves the quality of matches between workers and firms, the attendant gains in allocative efficiency are likely to raise social welfare. By contrast, if testing primarily redistributes 'desirable' workers among competing firms where they would have comparable marginal products, social benefits will be decidedly smaller than private benefits (cf. Stiglitz, 1975; Lazear, 1986). Moreover, since testing is itself costly, the net social benefits in the pure screening case could well be negative. Though our results provide little guidance as to which of these scenarios is more relevant, it appears unlikely that social benefits from testing exceed the private benefits. Quantifying these social benefits remains an important topic for future work.

7 Appendix 1: Proofs of propositions 4 and 8

Proposition 4 *(Naive selection case) Testing unambiguously raises the productivity of both a and b hires.*

The expected productivity of hires at firms using naive selection is

$$E(\eta|H,x) = \bar{\eta}_x + \gamma\sigma_\eta\lambda\left(\frac{\gamma\left(\kappa_n\left(\gamma\right) - \bar{\eta}_x\right)}{\sigma_\eta}\right).$$

Introduction of testing is equivalent to a rise in screening precision. The impact of screening precision

[49] This is referred to by the EEOC's Uniform Guidelines on Employee Selection Criteria (1978) as the "Four Fifths" rule. The test used at this firm was evaluated for "Fourth Fifths" compliance. Had it failed, it would likely have been modified.

on the productivity of hires is:

$$\frac{\partial E(\eta|H,x)}{\partial\gamma}=\sigma_\eta\lambda\left(\frac{\gamma\left(\kappa_n\left(\gamma\right)-\bar{\eta}_x\right)}{\sigma_\eta}\right)+\gamma\sigma_\eta\lambda'\left(\frac{\gamma\left(\kappa_n\left(\gamma\right)-\bar{\eta}_x\right)}{\sigma_\eta}\right)\left(\frac{\kappa_n\left(\gamma\right)-\bar{\eta}_x+\gamma\eta_n'\left(\gamma\right)}{\sigma_\eta}\right). \quad (12)$$

As shown in the text: $\eta_n'\left(\gamma\right)<0$, $\kappa_n\left(\gamma\right)-\bar{\eta}_b+\gamma\eta_n'\left(\gamma\right)>0$, $\kappa_n\left(\gamma\right)-\bar{\eta}_a+\gamma\eta_n'\left(\gamma\right)<0$. Noting that $\lambda\left(\cdot\right),\lambda'\left(\cdot\right)>0$, equation (12) is positive for b hires. Hence, b productivity rises. To show that equation (12) is also positive for a hires, we use the fact that $\gamma\eta_n'\left(\gamma\right)>\bar{\eta}_b-\kappa_n\left(\gamma\right)$, and substitute into equation (12):

$$\begin{aligned}\frac{\partial E(\eta|H,a)}{\partial\gamma} &> \sigma_\eta\left[\lambda\left(\frac{\gamma\left(\kappa_n\left(\gamma\right)-\bar{\eta}_a\right)}{\sigma_\eta}\right)+\lambda'\left(\frac{\gamma\left(\kappa_n\left(\gamma\right)-\bar{\eta}_a\right)}{\sigma_\eta}\right)\left(\frac{\gamma\left(\kappa_n\left(\gamma\right)-\bar{\eta}_a+\kappa_n\left(\gamma\right)+\bar{\eta}_b\right)}{\sigma_\eta}\right)\right]\\ &= \sigma_\eta\left[\lambda\left(\frac{\gamma\left(\kappa_n\left(\gamma\right)-\bar{\eta}_a\right)}{\sigma_\eta}\right)+\lambda'\left(\frac{\gamma\left(\kappa_n\left(\gamma\right)-\bar{\eta}_a\right)}{\sigma_\eta}\right)\left(\frac{\gamma\left(\bar{\eta}_b-\bar{\eta}_a\right)}{\sigma_\eta}\right)\right].\end{aligned}$$

Since $\gamma\left(\kappa_n\left(\gamma\right)-\bar{\eta}_a\right)>\gamma\left(\bar{\eta}_b-\bar{\eta}_a\right)$ and (using the Inverse Mills Ratio) $\lambda\left(x\right)\geq\lambda'\left(x\right)x$ for $x>0$, the right hand side of this equation is weakly positive, which establishes that $\partial E(\eta|H,a)/\partial\gamma>0$.

Proposition 8 *(Statistical discrimination case) Testing unambiguously raises the productivity of both a and b hires.*

The expected productivity of hires at firms using statistical discrimination is

$$E(\eta|H,x)=\bar{\eta}_x+\gamma\sigma_\eta\lambda\left(\frac{\left(\kappa_s\left(\gamma\right)-\bar{\eta}_x\right)}{\gamma\sigma_\eta}\right).$$

The effect of screening precision on the productivity of hires is:

$$\frac{\partial E(\eta|H,x)}{\partial\gamma}=\sigma_\eta\lambda\left(\frac{\kappa_s\left(\gamma\right)-\bar{\eta}_x}{\gamma\sigma_\eta}\right)+\sigma_\eta\lambda'\left(\frac{\kappa_s\left(\gamma\right)-\bar{\eta}_x}{\gamma\sigma_\eta}\right)\left(\frac{\bar{\eta}_x-\kappa_s\left(\gamma\right)+\gamma\kappa_s\left(\gamma\right)}{\gamma\sigma_\eta}\right). \quad (13)$$

As shown in the text: $\eta_s'\left(\gamma\right)>0$, $\gamma'\kappa_s\left(\gamma\right)<\kappa_s\left(\gamma\right)-\bar{\eta}_b$, $\gamma'\kappa_s\left(\gamma\right)>\kappa_s\left(\gamma\right)-\bar{\eta}_a$. Equation (13) is positive for a hires; a productivity rises. To show that equation (13) is also positive for b hires, we substitute the second of these inequalities ($\gamma'\kappa_s\left(\gamma\right)>\kappa_s\left(\gamma\right)-\bar{\eta}_a$) for $\gamma\kappa_s\left(\gamma\right)$ in equation (13):

$$\frac{\partial E(\eta|H,b)}{\partial\gamma}>\sigma_\eta\left[\lambda\left(\frac{\kappa_s\left(\gamma\right)-\bar{\eta}_b}{\gamma\sigma_\eta}\right)-\lambda'\left(\frac{\kappa_s\left(\gamma\right)-\bar{\eta}_x}{\gamma\sigma_\eta}\right)\left(\frac{\bar{\eta}_a-\bar{\eta}_b}{\gamma\sigma_\eta}\right)\right].$$

Since $\kappa_s\left(\gamma\right)-\bar{\eta}_b>\bar{\eta}_a-\bar{\eta}_b$ and $\lambda\left(x\right)\geq\lambda'\left(x\right)x$ for $x>0$, the right hand side of this equation is weakly positive, which establishes that $\partial E(\eta|H,b)/\partial\gamma>0$.

8 Appendix 2: The relationship between average applicant test scores and store level productivity

Our analysis of applicant test scores in sections (4) and (5) draws on a database of 214,688 applications submitted to the 1,363 stores in our sample during the year *after* the rollout of employment testing.

If these applications are not representative of applications submitted during the time-frame of our employment sample, we might either under- or overstate the *expected* effect of employment testing on productivity and hiring (though this would have no bearing on our estimation of the actual effect of testing on productivity or hiring in Tables 3 - 9).

To explore this concern, we estimate in Appendix Table 3 a set of models for the relationship between the mean employment test score of a store's applicants and the job spell durations of workers hired at that store:

$$D_{ijt} = \alpha + X_{ijt}\beta_9 + \beta_{10}\bar{S}_j + \beta_{11}T_{jt} + \beta_{12}\bar{S}_j \times T_{jt} + \psi_t + \varphi_j + e_{ijt}. \quad (14)$$

Here, the dependent variable is the completed job spell duration of workers hired at each store j, and $\bar{S}_j$ is the average test score of store $j's$ applicants. All models include either state effects or site effects and control for gender, race, year-month of hire and in some specifications, zip-code demographic variables (as in Table 9). This model is identical to equation (5) in the text, except that it is estimated with outcome variables for both tested and non-tested hires and includes interactions between tested status and mean store-level test scores.

If our applicant database accurately captures the characteristics of stores' applicants pools before and after the use of testing, we should expect two relationships: stores with lower average test scores should have lower productivity hires (that is, shorter job durations) ($\beta_{12} > 0$); and productivity gains from employment testing should be larger for stores with lower average test scores since, absent the test, a greater share of hires at these stores would be expected to be of low productivity ($\beta_{13} < 0$).

Repeating column 1 of Table 5, the first column of the appendix table shows a sizable, positive relationship between the test scores of applicants and the quality of hires in the pre-testing regime. The coefficient of 2.73 ($t = 5.0$) on the mean test score variable indicates that, conditional on race, gender, time and state effects, a 1 point higher average test score among a store's applicants predicts 2.7 additional days of job duration for the store's non-tested hires. Controlling for minority resident share and median household income in the store's zip code raises the coefficient on the mean test slightly to 3.2 days ($t = 4.5$). Hence, a one-standard deviation (3.7 point) difference in average store-level test scores predicts a 12 day difference in mean job duration.

In columns 3 and 4, we estimate equation (5) for the sample of workers hired using the employment test. Because this group of hires was selected using the test, we expect to find a weaker test-tenure relationship here. This expectation is confirmed. The coefficient on the average applicant test score is only half as large for the tested relative to non-tested sample, and it is insignificant. When we pool all hires and add an interaction term between the store's mean applicant test score and a dummy variable indicating whether a worker was hired using employment testing, we find (column 6) that

mean applicant test scores are much less predictive of productivity for the sample of workers hired using the test than those hired without: 3.3 versus 1.7 days tenure gain per 1 additional test point.

In column 7, we add site fixed effects. These absorb the main effect of applicant test scores but identify the interaction term. Consistent with prior columns, the gains to testing depend upon baseline applicant characteristics. While the (employment-weighted) mean store in our sample gains 18.7 days of tenure from employment testing, a store whose applicants are 5 percentage points below average gains 25.0 days of tenure and a store whose applicants are 5 percentage points above average gains 12.5 days of tenure. Hence, where applicants are of lower average quality, employment testing has greater potential to add value by screening out unproductive hires.

These findings – stores with higher applicant test scores had substantially higher productivity before the adoption of employment testing and stores with weaker applicant pools experienced greater productivity gains – suggest that the applicant database used for our analysis may provide a reasonable characterization of applicant characteristics in the period when employment testing was adopted.

9 References

Aberdeen Group. 2001. "Hourly Hiring Management Systems: Improving the Bottom Line for Hourly Worker-Centric Enterprises." Boston: Aberdeen Group.

Aigner, Dennis J. and Glen C. Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review*, 30(2), 175-187.

Altonji, Joseph and Charles Pierret. 2001. "Employer Learning and Statistical Discrimination," *Quarterly Journal of Economics*, 116(1), 313-350.

Angrist, Joshua D. 1993. "The "Misnorming" of the U.S. Military's Entrance Examination and its Effect on Minority Enlistments." University of Wisconsin-Madison: Institute for Research on Poverty Discuss Paper 1017-93, September.

Barrick, Murray R. and Michael K. Mount. 1991. "The Big Five Personality Dimensions and Job Performance: A Meta-Analysis." *Personnel Psychology*, 44(1), 1-26.

Becker, Gary S. 1957. *The Economics of Discrimination.* Chicago: University of Chicago Press.

Bertrand, Marriane and Sendhil Mullainathan. Forthcoming. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review.*

Bureau of National Affairs. 1983. *Employee Selection Procedures.* Washington, DC: Bureau of National Affairs.

Bureau of National Affairs. 1988. *Recruiting and Selection Procedures* (Personnel Policies Forum Survey No. 146). Washington, DC: Bureau of National Affairs.

Coate, Stephen and Glenn C. Loury. 1993. "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review*, 83(5), December, 1220-1240.

Digman, John M. 1990. "Personality Structure: The Emergence of the Five-Factor Model." *Annual Review of Psychology*, 41, 417-440.

Eitelberg, Mark J., Janice H. Laurence, Brian K. Waters, with Linda S. Perelman. 1984. "Screening for Service: Aptitude and Education Criteria for Military Entry." Washington, DC: United States Department of Defense.

Farber, Henry S. and Robert Gibbons. 1998. "Learning and Wage Dynamics." *Quarterly Journal of Economics*, 111(4), 1007-1047.

Fernandez, Roberto M. and Isabel Fernandez-Mateo. 2004. "Networks, Race, and Hiring." Mimeograph, MIT Sloan Schol of Management, May.

Giuliano, Laura. 2003. "Race, Gender, and Hiring Patterns: Evidence from a Large Service-Sector Employer." Mimeograph, Institute of Industrial Relations, University of California at Berkeley, October.

Goldberg, Lewis R., Dennis Sweeney, Peter F. Merenda and John Edward Hughes, Jr. 1998. "Demographic Variables and Personality: The Effects of Gender, Age, Education, and Ethnic/Racial Status on Self-Descriptions of Personality Attributes." *Personality and Individual Differences*, 24(3), 393-403.

Goodstein, Leonard D. and Richard I. Lanyon. 1999. "Applications of Personality Assessment to the Workplace: A Review." *Journal of Business and Psychology*, 13(3), 291-322.

Hartigan, John, and Alexandra Wigdor, eds. 1989. *Fairness in Employment Testing: Validity, Generalization, Minority Issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Holzer, Harry J., Steven Raphael and Michael A. Stoll. 2002. "Perceived Criminality, Criminal Background Checks, and the Racial Hiring Practices of Employers." Institute for Research on Poverty, Discussion Paper No. 1254-02, June.

Hunter, John E., and Frank L. Schmidt. 1982. "Fitting People to Jobs: The Impact of Personnel Selection on National Productivity," in Marvin D. Dunnette and Edwin A. Flesihman (eds.), *Human Performance and Productivity: Vol. 1., Human Capability Assessment* Hillsdale, NJ: Erlbaum.

Lazear, Edward P. 1986. "Salaries and Piece Rates." *Journal of Business*, 59(3), 405-431.

Levine, David I., Jonathan Leonard and Laura Giuliano. 2003. "Manager-Employee Similarity and Turnover: An Analysis of Quits and Dismissals at a Large Service-Sector Employer." Mimeograph, Haas School of Business, October.

Lundberg, Shelly J. and Richard Startz. 1983. "Private Discrimination and Social Intervention in

Competitive Labor Markets." *American Economic Review*, 73(3), June, 340-347.

Masters, Adrian. 2004."Antidiscrimination policy with culturally biased testing."Mimeograph, SUNY Albany, July.

Jacoby, Sanford M. 1985. "Employing Bureaucracy: Managers, Unions, and the Transformation of Work in American Industry, 1900-1945." New York: Columbia University Press.

Jencks, Christopher and Meredith Phillips, eds. 1998. *The Black-White Test Score Gap*, Washington, DC: Brookings Institution Press.

Montgomery, James D. 1991. "Social Networks and Labor-Market Outcomes: Towards an Economic Analysis" *American Economic Review*, 81(5), 1408–1418.

Murnane, Richard J. and Frank Levy. 1996. *Teaching the New Basic Skills.* New York: The Free Press.

Neal, Derek A. and William R. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy*, 104(5), 869-895.

Petit, Becky and Bruce Western. 2004. "Mass Imprisonment and the Life Course: Race and Class Inequality in U.S. Incarceration." *American Sociological Review*, 69(2), April, 151-169.

Phelps, Edmund S. 1972. "The Statistical Theory of Discrimination." *American Economic Review*, 62(4), 659-661.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature*, 37(1), March, 7-63.

Salop, Joanne and Steven Salop, "Self-Selection and Turnover in the Labor Market," *Quarterly Journal of Economics*, 60, 619-627.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics*, 87, 355-374.

Stiglitz, Joseph. 1975. "The Theory of "Screening," Education, and the Distribution of Income." *American Economic Review*, 65(3), June, 283-300.

Tett, Robert P., Douglas N. Jackson and Mitchell Rothstein. 1991. "Personality Measures as Predictors of Job Performance: A Meta-Analytic Review." *Personnel Psychology*, 44(4), 703-742.

U.S. Census Bureau. 2001. "Census 2000 Summary File 1: Census of Population and Housing." Washington, DC.

U.S. Census Bureau. 2003. "Census 2000 Summary File 3: Census of Population and Housing." DVD V1-D00S3ST-08-US1, Washington, DC.

U.S. Department of Labor, Equal Employment Opportunity Commission. 1978. "Uniform Guidelines on Employee Selection Procedures." 41CFR60-3.

Wigdor, Alexandra and Bert F. Green, Jr., eds. 1991. *Performance Assessment for the Workplace. Volume I.* Washington, DC: National Academy Press.

Wiggins, Jerry S. (editor). 1996. *The Five-Factor Model of Personality: Theoretical Perspectives.* New York: The Guilford Press.

Wilk, Stephanie L. and Peter Cappelli. 2003. "Understanding the Determinants of Employer Use of Selection Methods," *Personnel Psychology*, 56(1), Spring, 103-124.

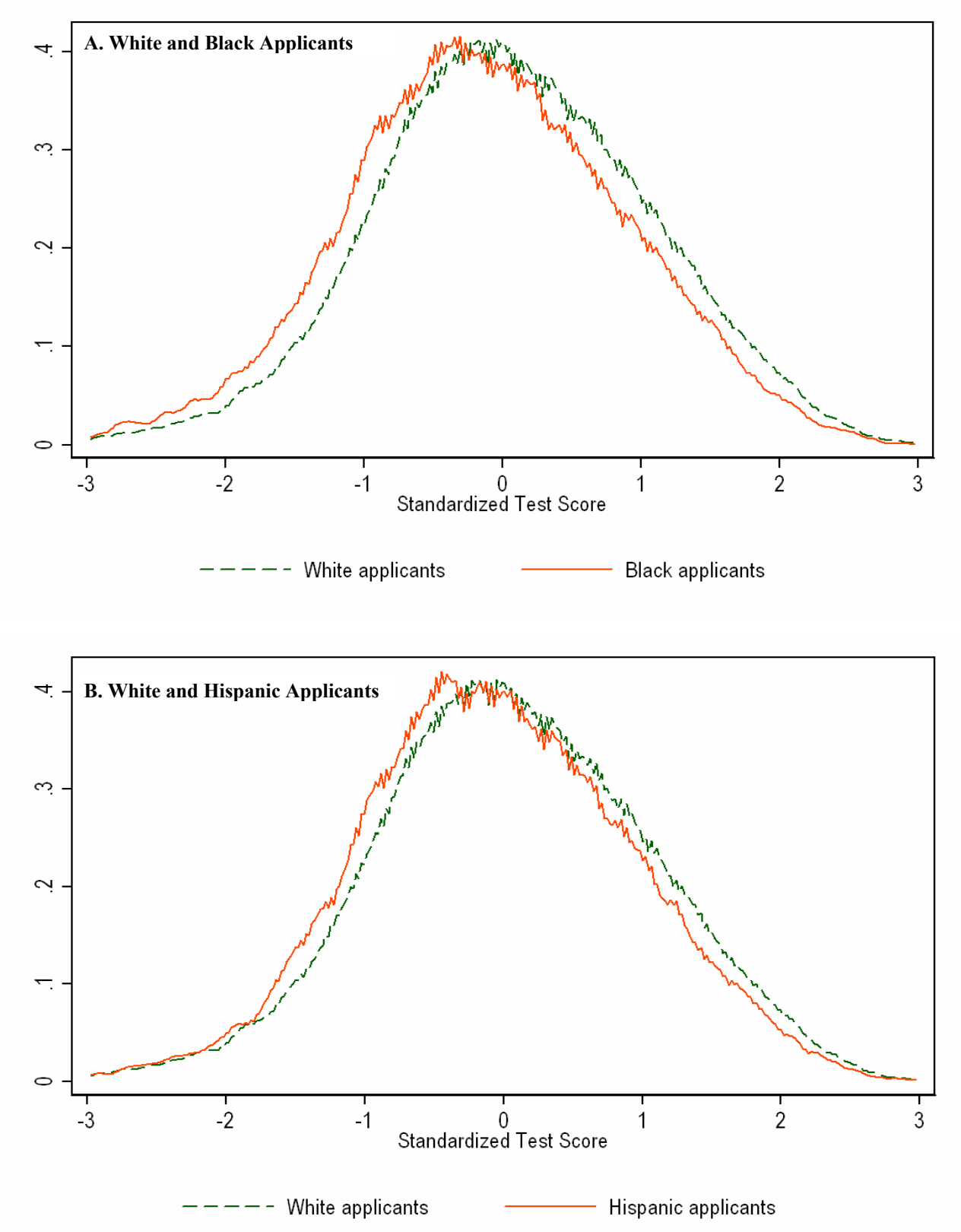

Figure 1. Density of Applicant Test Scores
Sample: All white, black and Hispanic applicants, June 2000 - May 2001 (n=189,067)

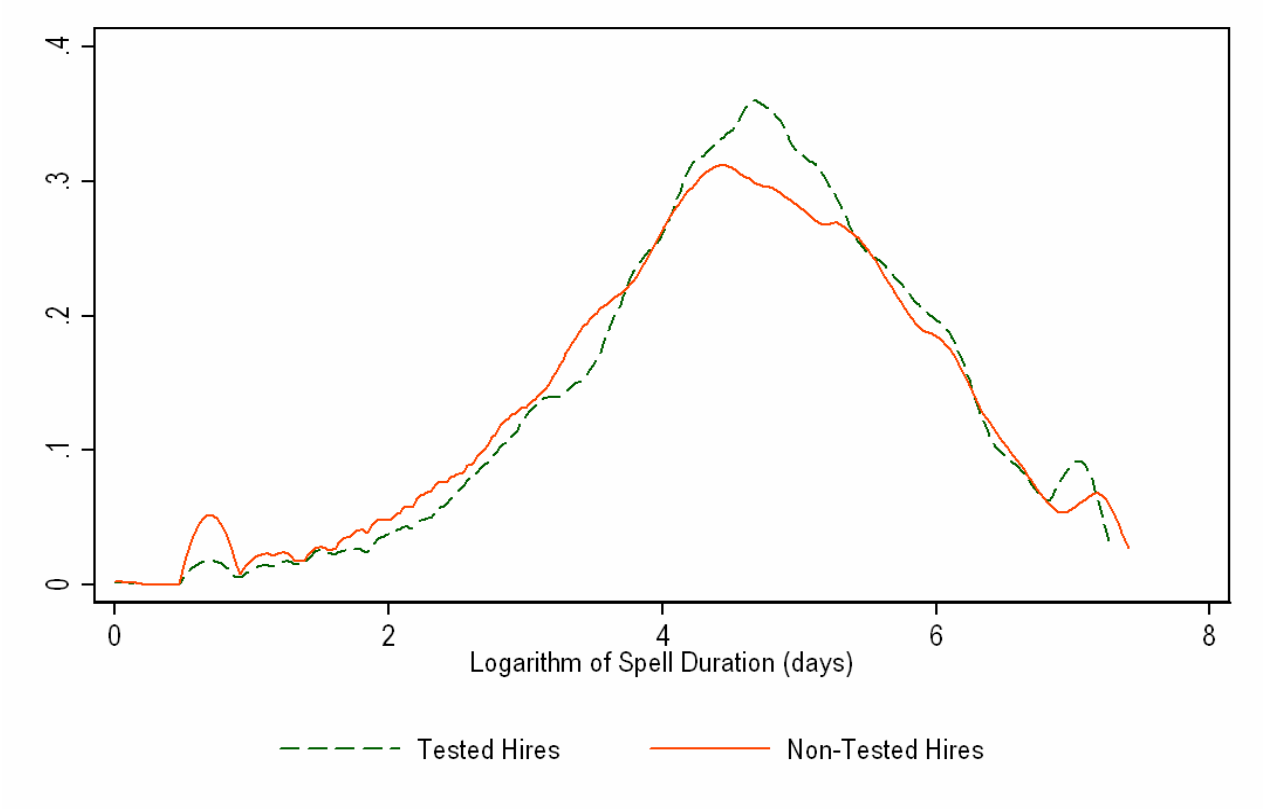

Figure 3. Density of Completed Job Spell Durations of Tested and Non-Tested Hires. Sample: All workers hired January 1999 - May 2000 (n=34,247).

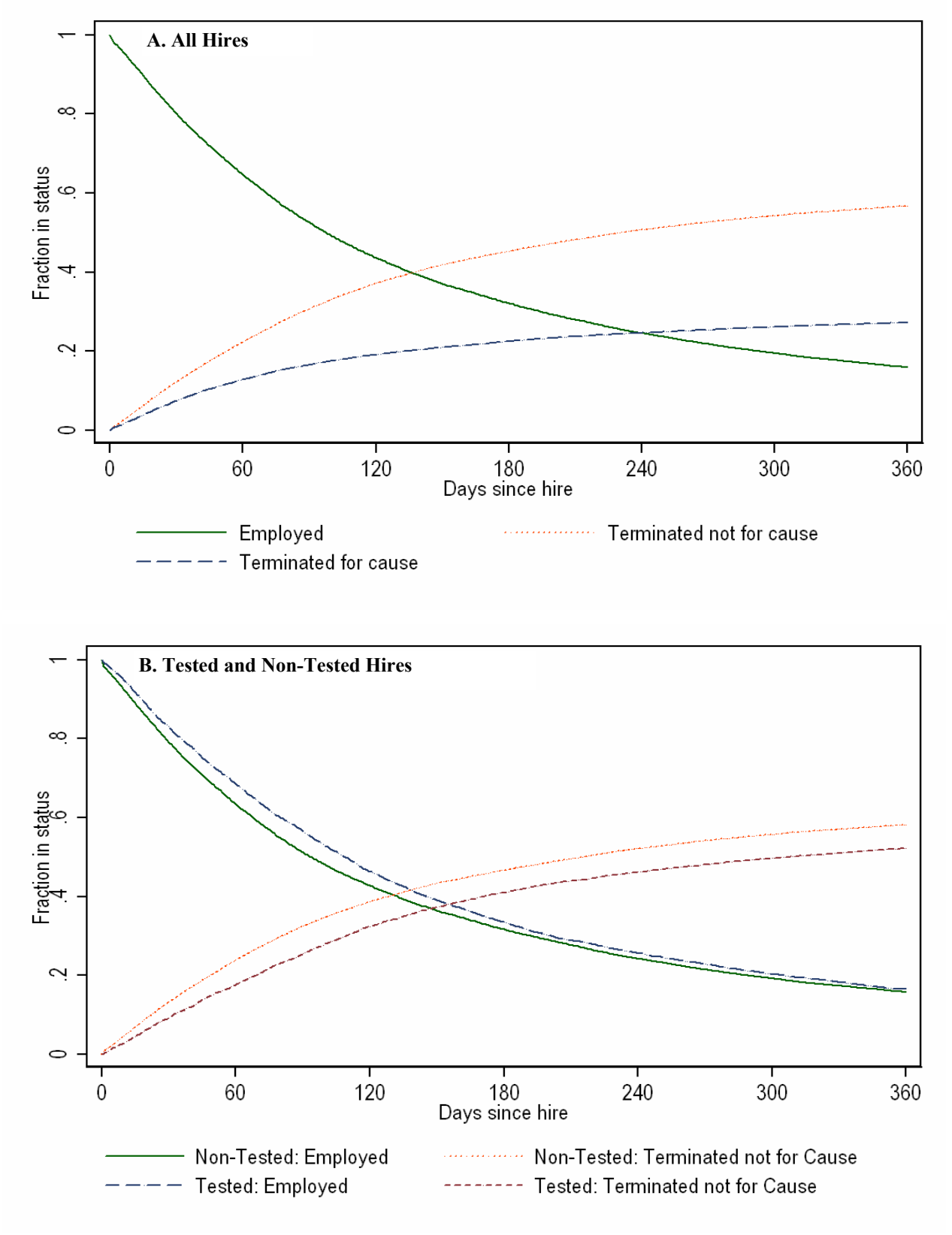

Figure 4. Employment Status of Workers during First 360 Days Following Hire.
Sample: Hires June 2000 - May 2001 with Valid Outcome Data (n=33,411)

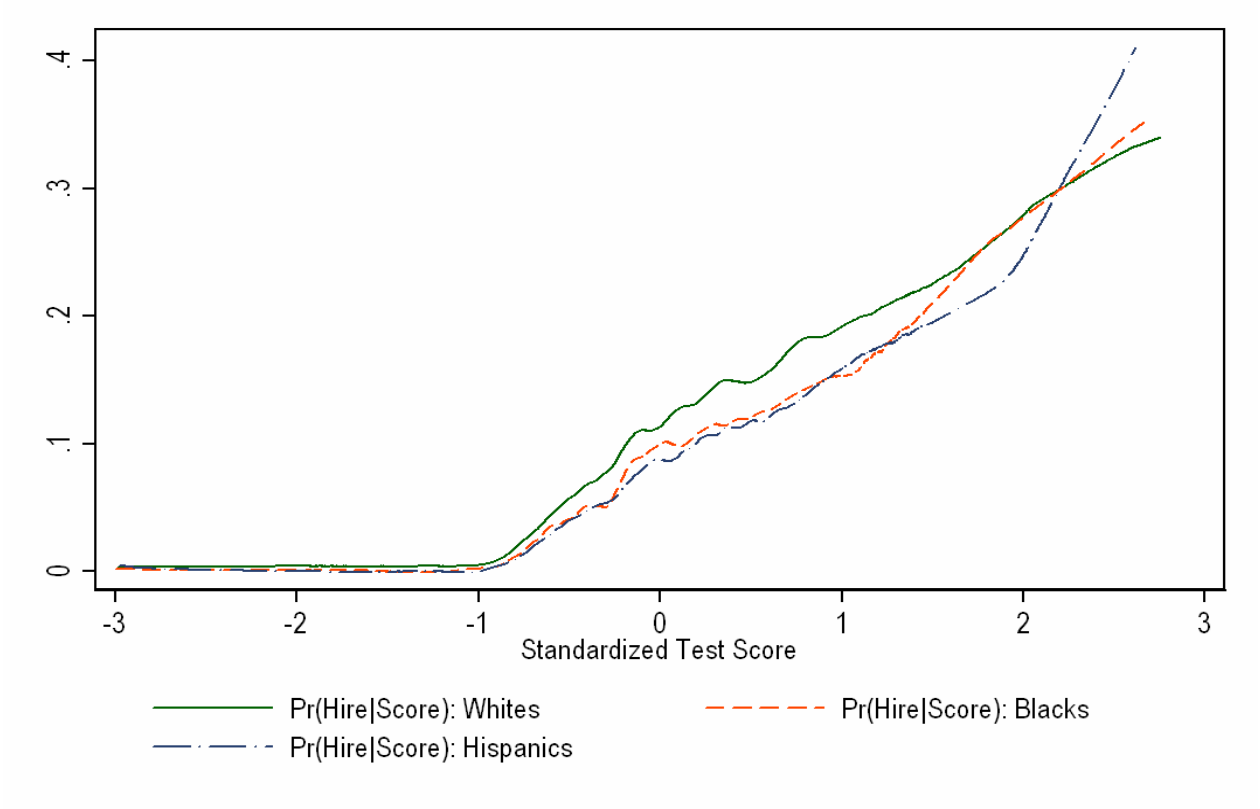

Figure 5. Conditional Probability of Hire as a Function of Test Score by Race: Locally Weighted Regressions

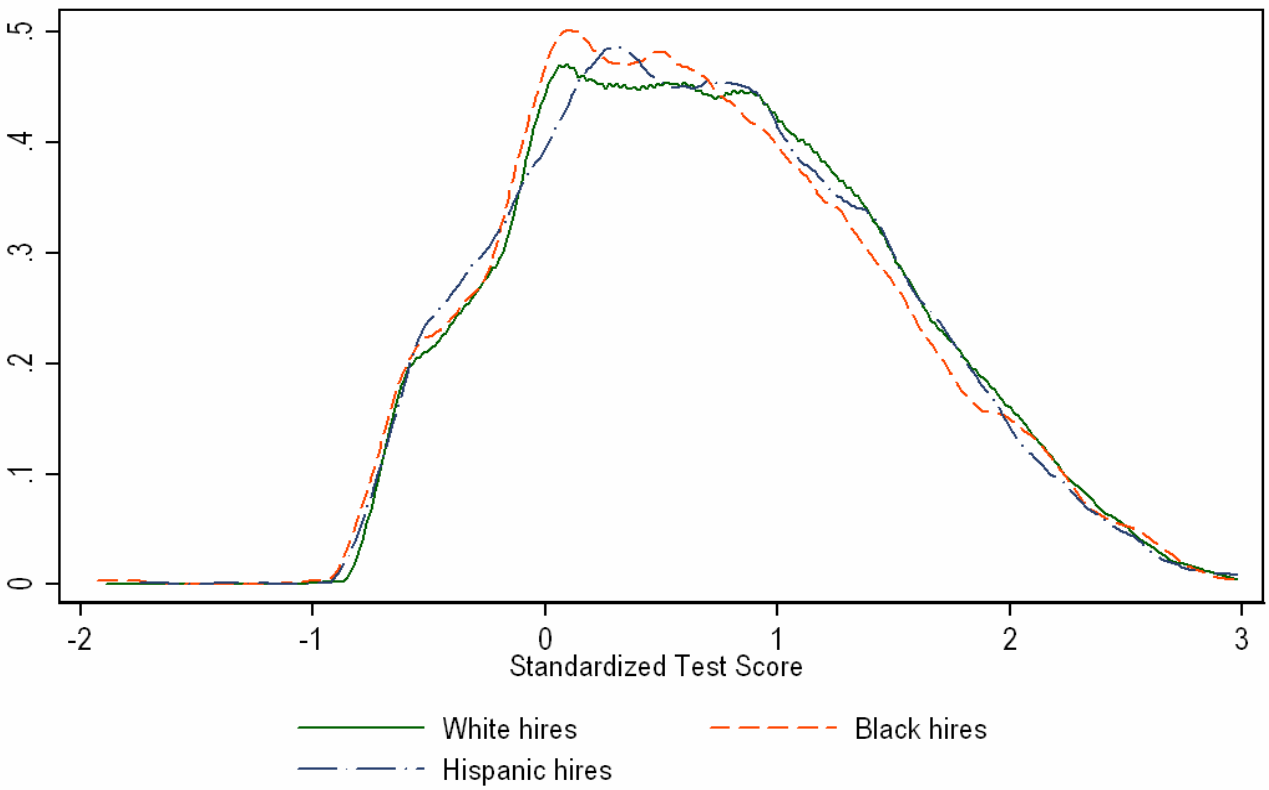

Figure 6. Test Score Densities of Hired Workers by Race

Table 1. Race and Gender Characteristics of Tested and Non-Tested Hires

Panel A: Frequencies

	Full Sample		Non-Tested Hires		Tested Hires	
	Frequency	% of Total	Frequency	% of Total	Frequency	% of Total
All	34,247	100.0	25,820	100.0	8,427	
White	23,560	68.8	18,057	69.9	5,503	65.3
Black	6,262	18.3	4,591	17.8	1,671	19.8
Hispanic	4,102	12.0	2,913	11.3	1,189	14.1
Male	17,604	51.4	13,135	50.9	4,469	53.0
Female	16,643	48.6	12,685	49.1	3,958	47.0

Panel B: Employment spell duration (days)

	Full Sample		Non-Tested Hires		Tested Hires	
	Mean	Median	Mean	Median	Mean	Median
All	173.7	99	173.3	96	174.9	107
	(1.9)	[97, 100]	(2.1)	[94, 98]	(3.0)	[104, 110]
White	184.0	106	183.0	102	187.1	115
	(2.1)	[103, 108]	(2.3)	[100, 105]	(3.6)	[112, 119]
Black	140.1	77	138.1	74	145.7	87
	(3.0)	[75, 80]	(3.5)	[71, 77]	(4.8)	[82, 92]
Hispanic	166.4	98	169.3	98	159.5	99
	(4.6)	[93, 103]	(5.4)	[92, 104]	(6.4)	[90, 106]

Panel C: Percent still working and terminated for cause after 180 days

	Full Sample		Non-Tested Hires		Tested Hires	
	Working	Term for Cause	Working	Term for cause	Working	Term for cause
All	32.6	22.4	32.2	21.5	34.0	25.2
	(0.4)	(0.4)	(0.5)	(0.4)	(0.7)	(0.7)
White	34.9	19.4	34.3	18.7	36.9	21.5
	(0.5)	(0.4)	(0.5)	(0.4)	(0.8)	(0.7)
Black	25.0	32.5	24.4	31.5	26.9	35.6
	(0.7)	(0.8)	(0.8)	(0.9)	(1.3)	(1.5)
Hispanic	31.2	24.0	31.3	22.4	31.1	27.9
	(1.0)	(0.8)	(1.1)	(0.9)	(1.7)	(1.6)

Table Notes:

-Sample includes workers hired between Jan 1999 and May 2000.

-Mean tenures include only completed spells (98% spells completed). Median tenures include complete and incomplete spells.

-Standard errors in parentheses account for correlation between observations from the same site (1,363 sites total). 95 percent confidence intervals for medians given in brackets.

-In Panel C, omitted outcome category is Terminated not for Cause, equal to one - [fraction still working + fraction term for cause].

Table 2. Test Scores and Hire Rates by Race and Gender for Tested Subsample

A. Test Scores of Applicants (range 0 to 100)

			Percent in each category		
	Mean	SD	Red	Yellow	Green
All	51.3	28.8	23.2	24.8	52.0
White	53.1	28.6	20.9	24.5	54.6
Black	47.7	29.0	27.8	25.2	47.1
Hispanic	49.6	28.6	24.9	25.6	49.6
Male	50.8	29.3	24.4	24.3	51.3
Female	51.8	28.1	21.6	25.5	52.9

B. Test Scores of Hires (range 0 to 100)

			Percent in each category		
	Mean	SD	Red	Yellow	Green
All	71.9	20.7	0.2	16.3	83.5
White	72.3	20.4	0.1	15.7	84.2
Black	70.8	20.8	0.4	16.4	83.2
Hispanic	71.7	20.6	0.1	17.3	82.6
Male	71.0	20.7	0.2	15.0	84.7
Female	72.9	20.6	0.2	17.5	82.3

C. Hire Rates by Applicant Group

By Race and Gender

Race/Sex	% Hired	Obs
All	8.90	214,688
White	10.16	113,354
Black	7.17	43,314
Hispanic	7.12	32,399
Male	8.57	112,669
Female	9.27	102,019

By Test Score Decile

Decile	% Hired	Obs
1	0.09	21,784
2	0.09	21,977
3	3.38	20,836
4	5.60	24,198
5	7.99	21,589
6	11.01	20,471
7	11.62	21,096
8	13.74	20,214
9	16.11	21,814
10	20.72	20,709

Table Notes:
- N=214,688 applicants and 19,107 hires at 1,363 sites.
- Sample includes all applicants and hires between June 2000 and May 2001 at sites used in treatment sample.

Table 3. OLS and IV Estimates of the Effect of Job Testing on the Job Spell Duration of Hires

Dependent Variable: Length of completed employment spell (days)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			A. OLS Estimates				B. IV Estimates			
Employment test			8.8	18.8	18.7	22.1	6.2	15.0	14.9	18.1
			(4.5)	(4.0)	(4.0)	(4.3)	(5.1)	(4.6)	(4.6)	(5.0)
Black	-43.4	-25.6			-25.6	-25.5			-25.6	-25.5
	(3.2)	(3.4)			(3.4)	(3.4)			(3.4)	(3.4)
Hispanic	-17.5	-11.9			-11.9	-11.9			-11.9	-11.9
	(4.4)	(4.1)			(4.1)	(4.1)			(4.1)	(4.1)
Male	-4.1	-1.9			-1.9	-1.8			-1.9	-1.8
	(2.4)	(2.4)			(2.4)	(2.4)			(2.4)	(2.4)
Site effects	No	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
State trends	No	No	No	No	No	Yes	No	No	No	Yes
R-squared	0.011	0.109	0.005	0.108	0.109	0.111				

Table Notes:

-N=33,588

-Robust standard errors in parentheses account for correlation between observations from the same site hired under each screening method (testing or no testing).

-All models include controls for month-year of hire.

-Sample includes workers hired Jan 1999 through May 2000 at 1,363 sites.

-Instrument for worker receiving employment test in columns 7 - 10 is an indicator variable equal to one if site has begun testing.

Table 4. Quantile Regression Estimates of the Effect of Job Testing on Job Spell Duration
Dependent Variable: Length of employment spell (days)

	(1)	(2)	(3)	(1)	(2)	(3)	(4)	(5)	(6)
	A. All Spells			B. Completed Spells					
	Median			Median		10th	25th	75th	90th
Employment test		9.0	8.0		9.8	3.0	5.0	16.0	-1.8
		(2.1)	(2.1)		(2.3)	(1.3)	(1.8)	(6.8)	(12.8)
Male	2.0		2.0	3.0	3.0	2.0	3.0	-7.5	-12.5
	(1.2)		(1.2)	(1.4)	(1.3)	(0.7)	(1.1)	(4.0)	(7.5)
Black	-24.0		-24.0	-22.3	-22.2	-2.0	-7.0	-56.1	-102.8
	(1.7)		(1.7)	(1.9)	(1.8)	(1.0)	(1.5)	(5.4)	(10.1)
Hispanic	-10.0		-10.0	-9.3	-9.5	-1.0	-4.0	-20.8	-38.7
	(2.0)		(2.0)	(2.3)	(2.2)	(1.2)	(1.7)	(6.4)	(12.1)
Obs	34,200	34,200	34,200	33,588	33,588	33,588	33,588	33,588	33,588

Table Notes:
-Standard errors in parentheses.
-All models include dummies for state and month-year of hire (not shown).
-Sample includes workers hired Jan 1999 through May 2000.
-Columns 5 through 10 present results only for completed spells. Columns 1 - 4 also include incomplete spells.

Table 5. The Relationship between Site-Level Applicant Mean Test Scores and the Job Spell Duration and Dismissal Status of Hired Workers.

	A. Job Spell Duration (days)	B. Employment Status at 180 Days		
		Employed	Neutral Termination	Termination for Cause
Mean applicant test score at site	2.73	0.31	0.11	-0.41
	(0.55)	(0.09)	(0.13)	(0.12)
Black	-33.32	-4.97	-7.08	12.05
	(3.99)	(0.63)	(0.97)	(0.99)
Hispanic	-6.85	-1.91	-2.54	4.44
	(5.48)	(0.90)	(1.15)	(1.00)
Male	-5.79	-1.47	-3.17	4.64
	(2.81)	(0.48)	(0.64)	(0.56)
State effects	Yes	Yes	Yes	Yes
Month x year of hire effects	Yes	Yes	Yes	Yes
R-squared	0.024	0.017	0.016	0.036
n	25,347	25,252		

Robust standard errors in parentheses account for error correlations between observations from the same site (n = 1,363). Sample is workers hired at each site prior to rollout of testing. Hire dates span January 1999 - May 2001. Mean applicant test scores by store are calculated for sample of all job applications submitted to sites during June 2000 - May 2001 (n = 214,488)

Table 6. OLS and IV Estimates of the Effect of Job Testing on Job Spell Duration by Race and Gender

Dependent Variable: Length of employment spell (days)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Whites		Blacks		Hispanics		Males		Females	
					A. OLS Estimates					
Employment test	20.4	24.4	22.8	21.0	8.2	15.3	18.7	21.6	20.1	25.2
	(5.2)	(5.5)	(9.3)	(10.1)	(13.1)	(13.7)	(5.8)	(6.2)	(6.0)	(6.4)
R-squared	0.121	0.124	0.231	0.238	0.303	0.311	0.147	0.150	0.160	0.164
					B. Instrumental Variables Estimates					
Employment test	19.3	23.3	18.3	16.5	6.2	15.1	12.9	15.2	17.5	22.8
	(5.9)	(6.4)	(11.3)	(12.5)	(14.4)	(15.4)	(6.4)	(7.1)	(6.8)	(7.3)
State trends	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Obs	23,030	23,030	6,199	6,199	4,037	4,037	17,292	17,292	16,296	16,296

Table Notes:

-Robust standard errors in parentheses account for correlation between observations from the same site hired under each screening method (testing or no testing).

-All models include 1,363 site fixed effects and controls month-year of hire, gender, and, in columns 7 - 10, race.

-Sample includes workers hired Jan 1999 through May 2000.

-Instrument for worker receiving employment test is an indicator variable equal to one if site has begun testing.

Table 7. OLS and IV Linear Probability Models for The Effect of Job Testing on Employment Status 180 Days Following Hire

Dependent Variable: Dichotomous variable equal to 100 if worker has indicated status

	(1)			(2)			(3)		
	Em-ployed	Term not for cause	Term for cause	Em-ployed	Term not for cause	Term for cause	Em-ployed	Term not for cause	Term for cause
Panel A: All Observations									
		OLS			OLS			IV	
Employment Test				4.44	-3.05	-1.39	3.73	-2.91	-0.82
				(0.97)	(1.08)	(0.95)	(1.12)	(1.21)	(1.09)
Black	-5.68	-3.53	9.21	-5.66	-3.54	9.21	-5.67	-3.54	9.21
	(0.82)	(0.89)	(0.83)	(0.82)	(0.89)	(0.83)	(0.82)	(0.89)	(0.83)
Hispanic	-2.05	-0.95	3.00	-2.05	-0.95	3.00	-2.05	-0.95	3.00
	(0.97)	(1.05)	(0.88)	(0.97)	(1.05)	(0.88)	(0.97)	(1.05)	(0.88)
Male	-0.36	-3.58	3.94	-0.36	-3.58	3.94	-0.36	-3.58	3.94
	(0.54)	(0.59)	(0.48)	(0.54)	(0.59)	(0.48)	(0.54)	(0.59)	(0.48)
R-squared	0.100	0.079	0.108	0.100	0.079	0.108	0.100	0.079	0.108
Obs		33,250			33,250			33,250	
Panel B: Effects by Worker Race									
		White Hires			Black Hires			Hispanic Hires	
OLS estimate	5.44	-3.73	-1.72	4.01	-2.58	-1.44	3.40	-1.44	-1.96
	(1.20)	(1.32)	(1.08)	(2.29)	(2.79)	(2.74)	(3.20)	(3.51)	(2.95)
IV estimate	5.11	-4.16	-0.95	4.20	-1.11	-3.08	2.50	-1.84	-0.67
	(1.40)	(1.50)	(1.26)	(2.74)	(3.24)	(3.16)	(3.43)	(4.04)	(3.47)
Obs		22,871			6,070			3,992	

Table Notes:

-Robust standard errors in parentheses account for correlation between observations from the same site hired under each screening method (testing or no testing).

-All models include 1,363 site fixed effects and controls for month-year of hire.

-Sample includes workers hired Jan 1999 through May 2000.

-Instrument for worker receiving employment test is an indicator variable equal to one if site has begun testing.

Table 8. Conditional Logit and Linear Probability Models of The Effect of Job Testing on Applicant Hiring Odds by Race

Dependent Variable: An indicator variable equal to 100 if hired worker is of given race

	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: Fixed Effects Logit Estimates			
	White	White	Black	Black	Hispanic	Hispanic
Employment test (logit coefficient)	0.035	0.028	-0.017	0.007	-0.017	-0.049
	(0.055)	(0.058)	(0.067)	(0.071)	(0.073)	(0.076)
State trends	No	Yes	No	Yes	No	Yes
Obs	31,595	31,595	27,288	27,288	22,689	22,689
			Panel B: OLS Estimates			
	White	White	Black	Black	Hispanic	Hispanic
Employment test (OLS coefficient)	0.52	0.39	-0.21	0.04	-0.08	-0.13
	(0.85)	(0.90)	(0.68)	(0.71)	(0.61)	(0.66)
State trends	No	Yes	No	Yes	No	Yes
Obs	34,247	34,247	34,247	34,247	34,247	34,247
			Panel C: Instrumental Variables Estimates			
	White	White	Black	Black	Hispanic	Hispanic
Employment test (IV coefficient)	0.89	0.82	-0.11	0.14	-0.57	-0.69
	(0.96)	(1.04)	(0.77)	(0.80)	(0.69)	(0.76)
State trends	No	Yes	No	Yes	No	Yes
Obs	34,247	34,247	34,247	34,247	34,247	34,247

Table Notes:

-Standard errors in parentheses. For OLS and IV models, robust standard errors in parentheses account for correlations between observations from the same site.

-Sample includes workers hired Jan 1999 through May 2000.

-All models include controls for month-year of hire and site fixed effects.

-Fixed effects logit models discard sites where all hires are of one race or where relevant race is not present.

Table 9: The Relationship Between Store Zip Code Demographics and Race of Hires Before and After Job Testing
Dependent Variable: An indicator variable equal to 100 if hired worker is of given race

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Panel A: Race of Hires and Racial Composition of Store Zip-Code											
	White				Black				Hispanic			
	Pre	Post	Both	Both	Pre	Post	Both	Both	Pre	Post	Both	Both
Share non-white in zip code	-86.8	-85.6	-87.0		56.1	56.4	56.1		30.7	29.2	30.9	
	(2.3)	(3.4)	(2.2)		(3.5)	(5.0)	(3.3)		(3.0)	(4.3)	(2.8)	
Share non-white in zip code x post			1.3	-0.2			1.1	1.4			-2.4	-1.2
			(3.3)	(1.8)			(4.9)	(1.7)			(4.5)	(1.6)
Site effects	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes
State effects	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No
R-squared	0.229	0.251	0.234	0.350	0.168	0.195	0.173	0.354	0.129	0.109	0.122	0.293
Obs	25,820	8,427	34,247	34,247	25,820	8,427	34,247	34,247	25,820	8,427	34,247	34,247
	Panel B: Race of Hires and Log Median Income in Store Zip-Code											
	White				Black				Hispanic			
	Pre	Post	Both	Both	Pre	Post	Both	Both	Pre	Post	Both	Both
Log median income in zip code	31.7	39.2	31.9		-19.8	-22.9	-19.8		-11.9	-16.3	-12.2	
	(2.5)	(3.1)	(2.4)		(2.5)	(3.2)	(2.4)		(1.6)	(2.5)	(1.6)	
Log median income in zip code x post			6.0	0.7			-3.1	-0.4			-2.9	-0.3
			(3.8)	(1.6)			(3.7)	(1.4)			(2.8)	(1.2)
Site effects	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes
State effects	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No
R-squared	0.116	0.153	0.123	0.350	0.099	0.128	0.104	0.354	0.101	0.094	0.097	0.293
Obs	25,820	8,427	34,247	34,247	25,820	8,427	34,247	34,247	25,820	8,427	34,247	34,247

Table Notes:
-Robust standard errors in parentheses account for correlations between observations from the same site (pre or post use of employment testing in models where both included).
-Sample includes workers hired Jan 1999 through May 2000.
-All models include controls for month-year of hire, and where indicated, 1,363 site fixed effects or state fixed effects.

Appendix Table 1. First Stage Models for Worker Receipt of Employment Test
Dependent Variable: Equal to one if hired worker received test

	(1)	(2)	(3)	(4)
Store has adopted test	0.888	0.862	0.863	0.852
	(0.008)	(0.010)	(0.007)	(0.008)
Male	0.000	0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.001)	(0.001)
Black	0.002	0.004	0.000	0.000
	(0.003)	(0.003)	(0.003)	(0.003)
Hispanic	0.008	0.006	0.003	0.003
	(0.003)	(0.003)	(0.003)	(0.003)
State trends	No	Yes	No	Yes
Site effects	No	No	Yes	Yes
R-squared	0.892	0.895	0.909	0.910

Table Notes:
-N=34,247 includes workers hired Jan 1999 through May 2000.
-Robust standard errors in parentheses account for correlation between observations from the same site hired under each screening method (testing or no testing).
-All models include controls for month-year of hire.

Appendix Table 2. The Effect of Job Testing on Job Spell Duration: Lead and Lag Specifications
Dependent Variable: Length of Completed Employment Spell (days)

Month relative to adoption of testing	(1)	(2)
5 months prior	6.3	5.6
	(6.2)	(6.2)
4 months prior	8.0	7.5
	(5.9)	(5.9)
3 months prior	-8.2	-7.8
	(5.9)	(5.9)
2 months prior	-6.9	-6.2
	(5.8)	(5.8)
1 month prior	8.0	8.8
	(6.6)	(6.7)
Month of rollout	14.1	16.7
	(6.6)	(6.6)
1 month post	28.3	31.8
	(7.9)	(8.0)
2 months post	25.8	29.5
	(8.3)	(8.5)
3 months post	18.6	24.4
	(9.4)	(9.8)
4+ months post	20.8	32.1
	(8.4)	(9.8)
State Trends	No	Yes
R-squared	0.110	0.112
Obs	33,588	33,588

Table Notes:
-Robust standard errors in parentheses account for correlation between observations from the same site.
-All models include controls for month-year of hire.
-Sample includes workers hired Jan 1999 through May 2000.

Appendix Table 3. The Relationship Between Job Spell Duration and Store Average Job Test Scores

Dependent Variable: Length of employment spell (days)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	No Pre-Test		Pre-Test		All		
Mean applicant test score	2.73	3.20	1.02	1.62	2.83	3.26	
	(0.55)	(0.72)	(0.82)	(1.04)	(0.60)	(0.67)	
Mean applicant test score x PT						-1.54	-1.25
						(0.74)	(0.62)
Worker received pre-employment test						7.98	18.68
						(4.79)	(4.03)
Share non-white in store zip code		-3.42		-7.60	-2.60	-3.75	
		(12.36)		(17.67)	(10.16)	(10.93)	
Log median income in store zip code		-15.40		-24.29	-17.14	-17.95	
		(7.32)		(11.25)	(6.16)	(6.63)	
State effects	Yes	Yes	No	Yes	Yes	Yes	No
Site effects	No	No	Yes	No	No	No	Yes
R-squared	0.024	0.024	0.025	0.026	0.022	0.022	0.109
Obs	25,347	25,347	8,241	8,241	33,588	33,588	33,588

Table Notes:

-Robust standard errors in parentheses account for correlation between observations from the same site (and, in columns 4 - 6, hired under each screening method: testing or no testing).

-Tenure sample includes 33,588 workers hired Jan 1999 through May 2000.

-All models include dummies for gender, race, and year-month of hire.

-Applicant test sample includes all applications submitted from June 2000 through May 2001 at treatment sites (214,588 applicants total).