NBER WORKING PAPER SERIES

WILL JOB TESTING HARM MINORITY WORKERS?

David H. Autor
David Scarborough

Working Paper 10763
http://www.nber.org/papers/w10763

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2004

We thank Daron Acemoglu, Joshua Angrist, David Card, John Donohue, Roland Fryer, Caroline Hoxby, Lawrence Katz, Edward Lazear, Michael Greenstone, Sendhil Mullainathan, Roberto Fernandez, numerous seminar participants, and especially Stacey Chen, Peter Schnabl and one incomparable referee for their contributions to the manuscript. Tal Gross provided stellar research assistance and Alan Baumbusch provided invaluable assistance with all data matters. Autor gratefully acknowledges financial support from the National Science Foundation (CAREER SES-0239538) and the Alfred P. Sloan foundation. Any opinions, findings and conclusions are those of the authors and do not necessarily reflect the views of the National Science Foundation or Kronos Incorporated, or those of the National Bureau of Economic Research.

Will Job Testing Harm Minority Workers?
David H. Autor and David Scarborough
NBER Working Paper No. 10763
September 2004, Revised January 2007
JEL No. D63,D81,J15,J71

ABSTRACT

Because minorities typically fare poorly on standardized tests, job testing is thought to pose an equity-efficiency trade-off: testing improves selection but reduces minority hiring. We develop a conceptual framework to assess when this tradeoff is likely to apply and evaluate the evidence for such a trade-off using data from a national retail firm whose 1,363 stores switched from informal to test-based worker screening over the course of on year. We document that testing yielded more productive hires at this firm -- raising median tenure by 10-plus percent. Consistent with prior research, minorities performed worse on the test. Yet, testing had no measurable impact on minority hiring, and productivity gains were uniformly large among minorities and non-minorities. These results suggest that job testing raised the precision of screening without introducing additional negative information about minority applicants, most plausibly because both the job test and the informal screen that preceded it were unbiased.

David H. Autor
Department of Economics
MIT, E52-371
50 Memorial Drive
Cambridge, MA 02142-1347
and NBER
dautor@mit.edu

David Scarborough
Black Hills State University
College of Business and Technology
1200 University Street, Unit 9551
Spearfish, South Dakota 57799-9551
and Kronos, Inc.
dscarborough@unicru.com

I Introduction

In the early twentieth 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]. 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 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 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 equality; 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]. This concern is forcefully articulated by Hartigan and Wigdor 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?"

Nor is this expression of concern merely rhetorical. Hartigan and Wigdor recommend that the U.S. Employment Service apply race-conscious score adjustments to the General Aptitude Testing Battery (GATB) to limit harm to minorities—despite their conclusion that the GATB is *not* race biased.

This presumed trade-off between efficiency and equality has garnered substantial academic, legal and regulatory attention, including specific provisions in Title VII of the Civil Rights Act of 1964 gov-

erning the use of employment tests,[1] several Equal Employment Opportunity Commission guidelines regulating employee selection procedures [U.S. Department of Labor, 1978],[2] and two National Academy of Sciences studies evaluating the efficacy and fairness of job testing [Hartigan and Wigdor, 1989; Wigdor and Green, 1991]. Despite this substantial body of research and policy, the case for a trade-off between equality and efficiency in the use of job testing is not well-established empirically—nor, as this paper argues, is it well-grounded conceptually.

We start from the presumption that competitive employers face a strong incentive to assess worker productivity accurately, but such assessments are inevitably imperfect. In our discussion and conceptual model, we consider two distinct—and not mutually exclusive—channels by which job testing may affect worker assessment. The first is to raise the precision of screening, which occurs if testing improves the accuracy of firms' assessments of applicant productivity. A large body of research demonstrates the efficacy of job testing for improving precision, so we view this channel as well-established.[3] The second is to 'change beliefs'—that is, to introduce information that systematically deviates from firms' assessments of applicant productivity based on informal interviews. This occurs if either the job test is biased *or* if the informal screen that precedes it is biased—or, potentially, if both are biased, albeit differently.

To see the relevance of these distinctions, consider a firm that is initially screening informally for worker productivity and which introduces a formal job test that improves the precision of screening. Assuming that minority applicants perform significantly worse than majority applicants on this test, will the gain in screening precision come at a cost of reduced minority hiring? As we show below, the answer will generally be *no* if both the informal screen and the formal test provide *unbiased* measures of applicant productivity. In this case, the main effect of testing will be to raise the precision of screening within each applicant group; shifts in hiring for or against minority applicants are likely to be small and will favor minorities. Notably, this result does not require that both the test and informal screen are unbiased. Our model below suggests that the harm or benefit to minority workers from testing depends primarily on the *relative* biases of the formal and informal screens. So long as

[1] See Title VII of the Civil Rights Act of 1964, 42 U.S.C. §§ 2000e-2, Section 703(h).

[2] The EEOC's Uniform Guidelines on Employee Selection Criteria [1978] introduces the "Four Fifths" rule, which states (Section 4d), "A selection rate for any race, sex, or ethnic group which is less than four-firths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact."

[3] 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. Most relevant to this study, the consensus of the personnel psychology literature is 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].

the information provided by job tests about minority applicants is not systematically more negative than firms' beliefs derived from informal screens, job testing has the potential to raise productivity without a disparate impact on minority hiring. This result makes it immediately apparent why the presumption that job testing will harm minority workers is suspect: there is little reason to expect that job testing is *more* minority-biased than informal hiring practices.[4]

This discussion, and our conceptual model, suggest that the presumed trade-off between efficiency and equality in hiring is an empirical possibility rather than a theoretical certainty. Evaulation of the evidence for this trade-off requires a comparison of the hiring and productivity of similar workers hired by comparable employers with and without the use of employment testing. There is to our knowledge no prior research that performs this comparison.[5] In this paper, we empirically evaluate the consequences of private sector job testing for minority employment and productivity by studying the experience of a large, geographically dispersed retail firm whose 1,363 establishments switched from an informal, paper-based screening process to a computer-supported, test-based screening process over a one year period. Both hiring methods use face-to-face interviews, while the test-based method also places substantial weight on a personality test that is administered and scored by computer. 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 the date that employment tests were introduced at their sites.

We find strong evidence that testing yielded more productive hires—increasing mean and median employee tenure by 10 to 12 percent. Consistent with a large body of work, we find that minority applicants performed significantly worse than majority applicants on the employment test. Had the test changed employers' beliefs about the average productivity of minority relative to majority applicants, our model suggests that testing would have raised White hiring at the expense of Black hiring and reduced the substantial productivity gap between Black and White workers. Neither of these effects occurred; the racial composition of hires was unchanged by testing and productivity gains were uniformly large among both minority and majority hires.

In light of our theoretical model, these results imply that the job test was unbiased *relative to the informal screen* it supplemented. Testing therefore raised productivity by improving selection within minority and majority applicant pools rather than by shifting the distribution of employment towards the higher scoring group (White applicants). By performing a parametric simulation of the conceptual

[4] In practice, there is considerable evidence that employers favor majority over minority workers when interviewing and hiring, suggesting the presence of taste-based or statistical discrimination or both [Altonji and Blank, 1999; Goldin and Rouse, 2000; Bertrand and Mullainathan, 2004].

[5] All prior studies of which we are aware compare anticipated or actual hiring outcomes using an employment test to a *hypothetical* case in which, absent testing, firms do not already account for majority/minority productivity differences.

model using the observed data on test scores, hiring and productivity, we reach a stronger conclusion: the lack of *relative bias* demonstrated by our results is most plausibly explained by a lack of *absolute* bias. That is, we accept the hypothesis that both the informal screen and job test were *unbiased*.

Our research contributes to the influential literature on testing, race and employment in three respects. First, despite substantial regulatory concern about the possible adverse consequences of job testing for minority hiring, we are unaware of any work that empirically evaluates whether use of job testing in a competitive hiring environment harms (or benefits) minority workers. Second, whereas the bulk of the prior literature on job testing focuses on the U.S. military and other public sector agencies, we study the experience and personnel data of a large, for-profit retail enterprise as it introduces job testing. Since incentives and constraints are likely to differ between public and private sector employers, we believe this makes the findings particularly useful. A final unusual feature of our research is that we look beyond the hiring impacts of job testing to evaluate its consequences for the productivity of hires (as measured by job spell durations), both overall and by race. As our conceptual model underscores, these two outcomes—hiring and productivity—are theoretically closely linked and hence provide complementary evidence on the consequences of job testing for employee selection.

Most closely related to our work is a study by Holzer, Raphael and Stoll [2006], which finds that employers that initiate criminal background checks for job applicants are more likely to hire minority workers than those who do not. Holzer et al. conclude that absent criminal background checks, employers statistically discriminate against Black applicants. Like these authors, we find that improved job applicant information—here, job tests—did not harm minority applicants, despite the fact that minority applicants perform worse than majority applicants on the hiring screen. Relative to prior work, a key virtue of our study is that it exploits the phased rollout of job testing across numerous sites belonging to the same firm to potentially eliminate any unmeasured confounds between sites' screening practices and their preferences towards hiring minority workers. The analysis is therefore likely to provide a credible test of the ceteris paribus impact of testing on minority hiring and productivity.[6]

Prior to the analysis, it is important to clarify the provenance of our data and address questions about the constraints under which the research was performed. The data analyzed for this study were provided to the first author by Unicru, Inc. (now a subsidiary of Kronos, Inc.) under a non-disclosure agreement with the Massachusetts Institute of Technology. This agreement placed no constraints on the conclusions of the analysis except that they be factually accurate. Among numerous potential firms available for analysis, the firm studied in this article was selected by the first author because its phased rollout of job testing across company sites offered a compelling research design. Unicru

[6] Close in spirit to our study, though answering a distinct question, is Angrist [1993], which demonstrates that successive increases in the military's AFQT qualification standard reduced minority enlistment.

personnel had not previously analyzed the firm's data to evaluate the effect of job testing on the racial distribution of hiring or productivity. All data used for the analysis were obtained from Unicru's data warehouse, which contains archives of personnel data for client firms. Consent for use of these data was not required or requested from the firm studied. After the analysis was complete and the first draft of the paper was in circulation in January 2004, personnel managers of the firm were briefed on the study and interviewed about the firm's personnel policies before and after the implementation of job testing.

The paper proceeds as follows. The subsequent section describes our data and details the firm's hiring procedures before and after the introduction of testing. Section III offers a theoretical model that illustrates how the possible disparate impacts of job testing depend critically on both the test's precision and its *bias* relative to the informal screens it supplements. Sections IV and V provide our empirical analysis of the consequences of testing for productivity and hiring. Section VI synthesizes and interprets these findings by benchmarking them against a parametric simulation of the theoretical model applied to the observed applicant, hiring and productivity database. Section VII concludes.

II 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 include 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 through 30, and many hold their jobs for short durations. As is shown in the first panel of Table I, approximately 70 percent of hires are White, 19 percent are Black (non-Hispanic), and 12 percent are Hispanic.[7] Median duration of completed job spells of line workers is 99 days, and the corresponding mean is 174 days (panel B).

A Worker screening before and after use of job tests

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 application forms, available from store

[7] These figures pertain to the flow of workers. Since Whites at this firm typically have longer job spells than minorities, they will be over-represented among the *stock* of workers relative to the flow of hires.

employees. If the store had an opening or a potential hiring need, the store manager would typically phone the applicant for a job interview and make a hiring decision shortly thereafter.

Commencing in June of 1999, the firm began rolling out electronic application kiosks provided by Unicru, Inc. By June of 2000, all stores in our sample were equipped with the technology. At the kiosk, applicants complete a questionnaire administered by a screen-phone or computer terminal, or in a minority of cases, by a web-based application. Like the paper application form, the electronic questionnaire gathers information on demographics and prior experience. In addition, applicants sign a release authorizing a criminal background check and a search of records in a commercial retail offender database.

A major component of the electronic application process is a computer-administered personality test, which contains 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 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].[8]

Once the electronic application is completed, the data are sent to Unicru for automated processing. The results are transmitted 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 ranking, along with a color code denoting the following score ranges: lowest quartile ('red'), second-to-lowest quartile ('yellow'), and two highest quartiles ('green').

Following the employment test, 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 first quartile ('red') applicants, and, as is shown in Table II, fewer than 1 percent of hires are from this group. Figure I shows that hiring rates

[8] An identical paper and pencil personality test could have been used prior to the introduction of computer-supported applicant screening. The cost of administering and scoring this paper and pencil test may have made it unattractive, however.

are strongly increasing in the test score in all deciles (see also panel c of Table II).[9] The low hiring rate observed in the data—only one in 11 applicants is hired—in part reflects the fact that applications are submitted continually while vacancies open only occasionally. For the typical store in our sample with 15 line positions and mean tenure of 173 days, we would expect approximately 28 job applications per month for 2 to 3 vacancies.

B Hiring and termination data

The primary data source for our analysis is company personnel records containing worker demographics (gender, race), date of hire, and termination date and termination reason for each worker hired during the sample frame. We use these data to measure pre-post changes in hiring and productivity by store following the introduction of testing. We calculate length of service for all employment spells in our sample, 98 percent of which are completed by the close of the sample. In addition, we utilize data on applicant's self-reported gender, race (White, Black, Hispanic), zip code of self-reported residence and zip code of the store to which they applied.[10] 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 of each applicant's residence and each store's location.

A critical feature of our research database is that employment (but not application) records are available for workers hired prior to implementation of the Unicru system at each store.[11] 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 $33,924$ workers hired into line positions, $25,561$ of whom were hired without use of testing and $8,363$ of whom were hired after receiving the test.[12]

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

[9] Our data do not distinguish between an applicant who is not offered a job and an applicant who declines a job offer. The observed hiring rate is therefore a lower bound on the offer rate.

[10] A small share of workers (0.9 percent) is classified as 'other' race. We exclude these workers because of a concern that the 'other' race category was not consistently coded before and after the introduction of job testing. The working paper version of this manuscript [Autor and Scarborough, 2004] contains complete results that include the 'other' race category. These results are nearly identical.

[11] Unicru imports its clients' personnel data into its own computer systems to produce performance evaluations. A by-product of this practice is that Unicru often obtains personnel data for workers hired prior to the implementation of the Unicru system. This is the case with the firm studied here.

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

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 that 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 unobserved wage variation in the data.

C Applicant test scores

To analyze test score differences in our sample, we draw on an applicant database containing all White, Black and Hispanic applications (189,067 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). This secondary data source is not linked to our primary research sample and is collected exclusively after the rollout of testing was completed. Although we would ideally analyze paper and electronic applications submitted during the rollout, these were not retained. In Section IV, we demonstrate that test scores from the applicant database are strongly correlated with the productivity of workers hired at each store *before* the introduction of employment testing, suggesting that this database provides an informative characterization of workers applying for job during the rollout period.

Table II shows that there are marked differences in the distribution of standardized (i.e., mean zero, variance one) test scores among White, Black and Hispanic applicants. Kernel density comparisons of test scores in Figure II underscore the pervasiveness of these differences. Relative to the White test score distribution, the Black and Hispanic test score densities are visibly left-shifted. These racial gaps, equal to 0.19 and 0.12 standard deviations, accord closely with the representative test data reported by Goldberg et al. [1998].[13] To examine whether these race differences are explained by other observable, non-race attributes of test-takers, we report in Table III a set of descriptive OLS models in which individual test scores of job applicants are regressed on all the major covariates in our database including race, gender, year and month of application, indicator variables for each of the 1,363 stores in the sample, state specific time trends, and measures of the median log income and percent non-White in the applicant's zip-code of residence. Conditional on these detailed control variables, race gaps in test scores are 60 to 75 percent as large as the unconditional differences and are highly significant, suggesting that the job test conveys information about applicants that is not fully proxied by observable characteristics.

[13] Using a representative sample of the U.S. workforce, Goldberg et al. [1998] 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.

III When does job testing pose an equality-efficiency trade-off?

Prior to analyzing the effect of testing on hiring and productivity at this firm, we provide a conceptual framework to explore when job testing is likely to pose an equality-efficiency trade-off. We define an equality-efficiency trade-off as a case where the productivity gains from testing come at a cost of reduced minority hiring. Our conceptual framework is related to well known models of statistical discrimination by Phelps [1972], Aigner and Cain [1977], Lundberg and Startz [1984], Coate and Loury [1993] and Altonji and Pierret [2001]. The contribution of our analysis is to explore how the impact of job testing on the employment opportunities and productivity (conditional on hire) of minority and majority workers depends on the discrepancy between firms' prior information about population parameters—based on job interviews and other established hiring practices (briefly 'interviews')—and the information provided by job tests. We refer to the discrepancy between tests and interviews as the *relative bias* of tests.

Our analysis yields three main results: (1) if *job tests are relatively unbiased* (i.e., relative to interviews), they *do not* pose an equality-efficiency trade-off; that is, the efficiency gains from testing come at no cost in reduced minority hiring; (2) if *job tests are bias-reducing*—that is, they mitigate existing biases—efficiency gains accrue in part from reduced hiring of groups favored by *pre-existing* (i.e., interview) biases; thus, minority hiring is reduced if pre-existing biases favor minorities, and so an equality-efficiency trade-off is present; (3) *if job tests are bias-enhancing*—that is, they increase the extent of bias—testing raises the hiring of groups favored by the test but does not necessarily increase efficiency. We present the model and its main conclusions below. Proofs of selected propositions are found in the Appendix, with detailed proofs of all propositions available from the authors.

A The environment, timing and distributional assumptions

There are many firms facing numerous job applicants from two identifiable demographic groups, x_1 and x_2, corresponding to a majority and minority group. For simplicity, assume that each group comprises half of the applicant population (thus, 'minority' refers to historical circumstances rather than population frequency).

The ability (Y) of job candidates is distributed as

$$Y \sim N\left(\mu_0\left(x\right), 1/h_0\right).$$

The mean parameter $\mu_0\left(x\right)$ may depend on x. Assume that h_0, equal to the inverse of the population variance σ_0^2, is constant, independent of x.[14] Let the ability of each applicant, y, be a random draw from the population distribution for the relevant demographic group (x_1 or x_2). The firm treats the

[14] The assumption that σ_0^2 is independent of x stands in contrast to several models of statistical discrimination in

population parameters as known. Thus, the firm's prior distribution for a draw y is the population distribution.

Firms have a linear, constant returns to scale production technology and are risk neutral. Workers produce output, $f(y) = y$. Hence, ability and productivity are synonymous. Job spell durations are independent of y and wages are fixed,[15] so firms strictly prefer to hire more productive workers.

Job applicants are drawn at random from the pooled distribution of x_1 and x_2 workers. Firms hire applicants using a screening threshold where applicants whose expected productivity exceeds a specified value are hired. In a fully elaborated search framework, this screening threshold would depend on technology and labor market conditions. In our reduced form setup, the screening threshold is chosen so that the aggregate hiring rate is held constant at $K \in (0, 0.5)$. This simplification focuses our analysis on the first-order impacts of job testing on the distribution of hiring across demographic groups, holding total employment fixed. We additionally assume that the hiring rate of each demographic group is below 50 percent, so selection is from the right-hand tail of each applicant distribution.

Initially, applicants are screened using interviews. Each interview generates an *interview signal,* η. When testing is introduced, applicants are screened using both interviews and tests. The test score is denoted by s.

Suppose that there is no bias in interviews. Then the distribution of interview signals will be centered on the true productivity of each applicant. Precisely,

$$\eta \sim N(y, 1/h_\eta), \tag{1}$$

where h_η is the inverse of the variance of the interview signal (a measure of accuracy of the interview). Assume h_η does not depend on x.

Conditional on perceived productivity $\mu_0(x)$ for group x and the interview signal η, the firm updates its assessment of the expected productivity of the applicant:

$$m(x, \eta) \equiv y|_{x,\eta} \sim N(\mu(x, \eta), 1/h_I), \tag{2}$$

where the updated degree of precision equals $h_I \equiv h_\eta + h_0$, and the updated mean equals $\mu(x, \eta) \equiv [\eta h_\eta + \mu_0(x) h_0] / h_I$.

which testing is differentially informative (or uninformative) for minority groups due to their higher (lower) underlying productivity variance, e.g., Aigner and Cain [1977], Lundberg and Startz [1984], and Masters [2006]. We believe that the evidence supports our assumption. Analysis in Hartigan and Wigdor [1989], Wigdor and Green [1991] and Jencks and Philips [1989, chapter 2] all suggest that while tests commonly used for employee selection show marked mean differences by race, the by-race variances are comparable and, moreover, these tests are about equally predictive of job performance for minorities and non-minorities. As shown in Figure II and Table II, *mean* test scores in our sample also differ significantly among White, Black and Hispanic applicant groups but the variances of test scores are nearly identical for all three groups.

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

Suppose that there is no bias in testing. Then the distribution of test signals will be centered on the true productivity of each applicant. Precisely,

$$s \sim N(y, 1/h_S), \tag{3}$$

where h_S is the inverse of the variance of the interview signal (a measure of accuracy of the interview). Assume h_S does not depend on x. This generates a *posterior* for the firm's perception of the applicant's productivity

$$m(x, \eta, s) \equiv y|_{x,\eta,s} \sim N(\mu(x, \eta, s), 1/h_T), \tag{4}$$

where the degree of accuracy for the posterior (based on both testing and interviews) is $h_T \equiv h_S + h_I$; and the updated mean equals $\mu(x, \eta, s) \equiv [sh_S + \mu(x, \eta)h_I]/h_T$. Note that $h_T > h_I$.

B First outcome of interest: Hiring rates

To assess when testing poses an *equality-efficiency trade-off*, we study two outcomes. The first is the *hiring gap*, defined as the hiring rate of majority workers minus the hiring rate of minority workers.

Denote the hiring decision as $Hire = 0, 1$ for the firm. If there is no testing, the hiring decision will completely depend upon the firm's prior and the results of interviews: $Hire = I\{\mu(x, \eta) > \kappa_I\}$, where κ_I is the screening threshold that yields a total hiring rate of K using interviews and $I\{\cdot\}$ is the indicator function. The expected hiring rate of group x applicants who have received the interview is

$$E_\eta[Hire|x] = 1 - \Phi(z_I(x)),$$

where $z_I(x) \equiv [\kappa_I - \mu_0(x)]/\sigma_0\rho_I$ and $\rho_I \equiv Corr[\mu(x, \eta), y] = (1 - h_0/h_I)^{1/2}$. Note that we iterate expectations over η to obtain the *unconditional* hiring rate (i.e., not conditional on a specific value of η) for group x applicants based on interviews. Specifically, $E_\eta[Hire|x] = \int E[Hire|x, \eta] f(\eta|x) d\eta$.[16]

If both testing and interviews are used, the hiring decision is $Hire = I\{\mu(x, \eta, s) > \kappa_T\}$, where κ_T is the screening threshold that yields a total hiring rate of K using both interviews and test scores. The expected hiring rate of group x applicants who have received the interview and the test is:[17]

$$E_{\eta,s}[Hire|x] = 1 - \Phi(z_T(x))$$

where $z_T(x) \equiv [\kappa_T - \mu_0(x)]/\sigma_0\rho_T$ and $\rho_T \equiv Corr[\mu(x, \eta, s), y] = (1 - h_0/h_T)^{1/2}$.

[16] Since η is normally distributed and assessed productivity conditional on η is normally distributed, the unconditional distribution of perceived productivity is also normally distributed. It can be shown that the variance of the unconditional distribution is $V_{\eta,y}(\mu(x, \eta)) = \rho_I^2\sigma_0^2$.

[17] We iterate expectations over η and s to obtain the unconditional hiring rate for group x applicants based on interviews and tests. It can be shown that the variance of the unconditional distribution is $V_{s,\eta,y}(\mu(x, \eta)) = \rho_T^2\sigma_0^2$.

When hiring is based on interviews, the *hiring gap* between majority and minority workers is

$$\gamma_I = E_\eta[Hire|x_1] - E_\eta[Hire|x_2].$$

When hiring is based on testing and interviews, this gap is

$$\gamma_T = E_{\eta,s}[Hire|x_1] - E_{\eta,s}[Hire|x_2].$$

We denote the effect of testing on the hiring gap as $\Delta\gamma \equiv \gamma_T - \gamma_I$.

C Second outcome of interest: Productivity

A second outcome of interest is the effect of testing on productivity. If only interviews are used, the mean productivity for hired workers of group x is

$$E_\eta[y|Hire = 1, x] = \mu_0(x) + \sigma_0\rho_I\lambda(z_I(x)), \tag{5}$$

where $\lambda(z_I)$ is the inverse Mills ratio $\phi(z_I)/[1-\Phi(z_I)]$, equal to the density over the distribution function of the standard normal distribution evaluated at z_I.

If both tests and interviews are used, the mean productivity for hired workers of group x is

$$E_{\eta,s}[y|Hire = 1, x] = \mu_0(x) + \sigma_0\rho_T\lambda(z_T(x)). \tag{6}$$

A comparison of equations (5) and (6) shows that testing affects the productivity of hired applicants through two channels: selectivity (equal to one minus the hiring rate) and screening precision. All else equal, a rise in selectivity (i.e., a reduction in hiring) for group x raises the expected productivity of workers hired from group x by truncating the lower-tail of the group x productivity distribution. Screening precision refers to the accuracy of the firm's posterior, and its effect is seen in the terms ρ_I and ρ_T in equations (5) and (6), with $\rho_T > \rho_I$ (more precisely, both ρ_I and ρ_T are increasing functions of screening precision, so $h_T > h_I$ implies that $\rho_T > \rho_I$). All else equal, a rise in screening precision improves the accuracy of firms' assessments of worker productivity and so raises the quality of hires from each demographic group.

In addition to the impact of testing on overall productivity levels, we also study its effect on the *productivity gap*, defined as the mean productivity of majority workers minus the mean productivity of minority workers. This gap proves relevant to our empirical work because our model suggests that testing typically moves the hiring and productivity gaps in opposite directions.

When hiring is based on interviews, the majority/minority productivity gap is

$$\pi_I = E_\eta[y|Hire = 1, x_1] - E_\eta[y|Hire = 1, x_2].$$

When hiring is based on interviews and tests, this gap is

$$\pi_T = E_{\eta,s}\left[y|Hire = 1, x_1\right] - E_{\eta,s}[y|Hire = 1, x_2].$$

We denote the effect of testing on the productivity gap as $\Delta\pi \equiv \pi_T - \pi_I$.

D The effects of testing when both interviews and tests are unbiased

The potential for an equality-efficiency trade-off is relevant when one applicant group is less productive than the other. For concreteness, and without loss of generality, suppose that minorities are the less productive applicant group ($\mu_0(x_2) < \mu_0(x_1)$). These underlying population productivity differences imply observed differences in the hiring and productivity of minority and majority workers *prior to use of tests*. First, the hiring rate of minority applicants based on interviews will be lower than that of majority applicants ($\gamma_I > 0$). Second, minority workers hired using interviews will be on average less productive than majority workers ($\pi_I > 0$). Both inequalities follow from the firm's threshold hiring policy wherein applicants whose assessed productivity (equation (2)) exceeds a reservation value κ_I are hired.[18] This observation is significant for our empirical work because, as shown in Table I, minority workers hired using interviews are *less productive,* as measured by job tenure, than are majority workers hired using interviews.

To derive the effect of testing on the *hiring gap*, we note that the overall hiring rate in the model is constant at K. Hence, testing must either leave hiring of both groups unaffected or change the hiring rate of each group by equal but opposite amounts. It is straightforward to show by differentiation that: (1) it is not possible for hiring of both groups to be unaffected; and (2) testing *raises* minority hiring or, more generally, raises hiring of the applicant group with lower average productivity (see proof in Appendix):

$$\Delta\gamma < 0.$$

Intuitively, because the interview signal is error-ridden ($1/h_\eta > 0$) and expected majority applicant productivity exceeds expected minority applicant productivity, firms disproportionately hire applicants from the group favored by their prior—that is, majority applicants. Testing increases minority hiring because the posterior including the test score places more weight on observed signals and less weight on group means. However, simulations show that the effect of testing on the majority/minority hiring gap is typically small under the assumed normality of the productivity distributions. We therefore do not generally expect testing to induce a substantial change in minority hiring.

[18] The hiring rule ($Hire = I\{\mu(x,\eta) > \kappa_I\}$) equates the expected productivity of *marginal* hires from each applicant group. Because the *average* majority *applicant* is more productive than the average minority applicant, the average majority *hire* is also more productive than the average minority hire. As a referee pointed out, this result stems from the fact that the normal distribution is thin-tailed.

We obtain a similar, but stronger, result for the effect of testing on the majority/minority *productivity gap*: although minority workers hired using interviews are less productive than majority workers hired using interviews, testing leaves this majority/minority productivity gap essentially unaffected. More precisely, *testing raises productivity of both minority and majority hires approximately equally, with exact equality as selectivity approaches one* (see proof in Appendix). We write:

$$\Delta\pi \approx 0.$$

The intuition for this result stems from two sources: first, the threshold hiring rule equates the productivity of marginal minority and majority hires both before and after the introduction of testing; second, when selection is from the right-hand tail of the normal distribution, the truncated mean increases near-linearly with the point of truncation with a first derivative that is asymptotically equal to unity.[19] Consequently, a rise in screening precision raises the marginal and average productivity of hires almost identically for minority and majority workers.

Summarizing, *if both interviews and job tests are unbiased, testing does not pose an equality-efficiency trade-off.* Although job tests unambiguously raise productivity, the gains come exclusively from improved selection within each applicant group, not from hiring shifts against minorities.

These results are illustrated in Figure IIIa, which provides a numerical simulation of the impact of testing on hiring and productivity for a benchmark case where majority applicants are on average more productive than minority applicants and job interviews and job tests are both unbiased. The x-axis of the figure corresponds to the correlation between test scores and applicant ability ($\text{Corr}\langle s, y\rangle = 1/\left(1 + h_0/h_s\right)^{1/2}$), which is rising in test precision. The y-axis depicts the hiring rate of majority and minority applicants (left-hand scale) and the expected productivity (equivalently, ability) of majority and minority hires gap (right-hand scale).[20] Prior to the introduction of testing—equivalently, $\rho = 0$ in the figure—minority applicants are substantially less likely than majority applicants to be hired and are also less productive than majority workers conditional on hire. Job testing slightly reduces the minority/majority hiring gap. But this effect is small relative to the initial gap in hiring rates, even at maximal test precision. By contrast, testing leads to a substantial rise in the productivity of both

[19] Numerical simulations of the normal selection model show that this asymptotic equality is numerically indistinguishable from exact equality at selectivity levels at or above +0.1 standard deviation from the mean (i.e. $z_I, z_T \geq 0.1$). This result is also visible in the numerical simulation in Figure IIIa, where the productivity gap between minority and majority hires is invariant to testing. Recall from Table II that the overall hiring rate at this firm is 8.95 percent, implying that $z_I, z_T \approx 1.34$.

[20] In the simulation, the ability (equivalently productivity) of nonminority applicants is distributed $N\left(0, 0.29\right)$, the productivity of minority applicants is distributed $N\left(-0.19, 0.27\right)$, the precision of the informal ability signal is $1/0.45$, and 8.95 percent of applicants are hired. Thus, $h_0 = 1/0.27$, $h_\eta = 1/0.45$ $\mu_0\left(x_1\right) = 0$, $\mu_0\left(x_2\right) = -0.19$ and $K = 0.0895$. These values are chosen to match estimates from the parametric model simulation in Section VI of the paper. The precision of the job test ranges from $1/10,000$ to $1/0.0001$, corresponding to a correlation of $(0.0, 1)$ between test scores and applicant ability (plotted on the x-axis).

minority and majority hires, with the degree of improvement increasing in test precision. Consistent with the analytic results, testing has no detectable effect on the majority/minority productivity gap at any level of test precision.

E The effects of testing when interviews and tests are biased: The case of identical biases

Our main result so far is that use of an unbiased test introduced in an unbiased hiring environment raises productivity without posing an equality-efficiency trade-off. We now consider how test and interview biases affect this conclusion.

Suppose there is a *mean bias in interviews*. So, change equation (1) to

$$\eta^* \sim N\left(y + \nu_\eta(x), 1/h\right)$$

where $\nu_\eta(x_1) \neq \nu_\eta(x_2)$. We say that job interviews are *minority favoring* if $\nu_\eta(x_2) > \nu_\eta(x_1)$, and *majority favoring* if $\nu_\eta(x_1) > \nu_\eta(x_2)$). For example, managers may perceive majority applicants as more productive than equally capable minority applicants, or vice versa.[21]

Similarly, suppose there is a *mean bias in job tests*. So, change equation (3) to

$$s^* \sim N\left(y + \nu_s(x), 1/h_S\right)$$

where $\nu_s(x_1) \neq \nu_s(x_2)$, with the definition of minority favoring and majority favoring tests analogous to that for interviews. This might arise if tests are 'culturally biased' so that for given applicant ability, minority applicants score systematically below majority applicants.

Define the net bias of interviews as $\Delta\nu_\eta = \nu_\eta\left(x_1\right) - \nu_\eta\left(x_2\right)$ and, similarly, the net bias of tests as $\Delta\nu_s = \nu_s\left(x_1\right) - \nu_s\left(x_2\right)$. If $\Delta\nu_\eta > 0$, interviews favor majority applicants, and vice versa if $\Delta\nu_\eta < 0$ (and similarly for job tests). We refer to the difference in bias between tests and interviews $(\Delta\nu_s - \Delta\nu_\eta)$ as the 'relative bias' of tests.

Assume that firms' updated assessments of applicant productivity (based on interviews) and posteriors (based on interviews and tests) are still given by equations (2) and (4) except that we now substitute η^* and s^* for η and s. For consistency, suppose that firms' prior for each draw from the applicant distribution is mean-consistent with the information given by interviews, as in the unbiased case: $y|_x \sim N\left(\mu_0\left(x\right) + \nu_\eta\left(x\right), 1/h_0\right)$. Thus, firms do not compensate for biases in interviews or tests and we say that their *perceived* productivity of the applicant distribution is equal to true productivity plus interview bias.

[21] Equivalently, $\Delta\nu_\eta$ could be interpreted as taste-discrimination: firms' reservation productivity for minority and majority hires differs by $\Delta\nu_\eta$.

How do these biases affect our prior results for the impact of testing on equality and efficiency? Suppose initially that *interviews and tests are equally biased*—that is, both tests and interviews contain biases but these biases are identical ($\Delta\nu_s = \Delta\nu_\eta \neq 0$). In this no relative bias case, our prior results require only slight modification:

1. Use of tests that are unbiased *relative to* job interviews does not pose an equality-efficiency trade-off. In particular: (1) testing raises hiring of the applicant group with lower *perceived* productivity, $\Delta\mu + \Delta\nu_\eta$ (the minority group by assumption); and (2) testing raises productivity of both minority and majority hires approximately equally, with exact equality as selectivity approaches one. Thus, expanding on our earlier conclusion: unbiasedness of both interviews and tests ($\Delta\nu_s = \Delta\nu_\eta = 0$) is a sufficient but not a necessary condition for the no-trade-off result to hold. If both interviews and tests are equally biased ($\Delta\nu_s = \Delta\nu_\eta$)—thus, there is no *relative bias*—testing does not pose an equality-efficiency trade-off.

2. We showed above that if both interviews and tests are unbiased, the applicant group with lower average productivity will have a lower hiring rate and lower productivity conditional on hire than the group with higher average productivity ($\text{Sign}\langle\gamma_I\rangle = \text{Sign}\langle\pi_I\rangle$). Interview and testing biases can reverse this positive correlation. Because biases reduce selectivity of the favored group and raise selectivity of the non-favored group, it is possible for the group with a greater hiring rate to have lower productivity conditional on hire.[22] So, if minority hires are observed to be less productive than majority hires, this implies that either minority applicants have lower mean productivity than majority applicants (i.e., $\mu_0(x_2) < \mu_0(x_1)$) or that job interviews are minority-favoring ($\Delta\nu_\eta < 0$) or both.

F The effects of testing when interviews and tests have *non*-identical biases

We finally consider how job testing affects the productivity and hiring gaps when the test is biased *relative to* job interviews (i.e., $\Delta\nu_s \neq \Delta\nu_\eta$). For concreteness, we continue to assume that minority applicants are perceived as less productive than majority applicants: $\mu_0(x_1) + \nu_\eta(x_1) > \mu_0(x_2) + \nu_\eta(x_2)$. It is straightforward to establish the following three results:

1. Use of a job test that is biased relative to interviews: (1) raises the hiring rate of minorities if the test favors minorities (i.e., relative to interviews) but has ambiguous effects on minority hiring otherwise; and (2) reduces the productivity level of the group favored by the test relative to the group that is unfavored. For example, if minority applicants are perceived as less productive

[22] This result requires only that the absolute level of bias is sufficient to offset underlying mean majority/minority productivity differences, which can occur even if there is no relative bias in tests.

than majority applicants, use of a relatively minority-favoring test will raise minority hiring and reduce the productivity of minority relative to majority hires (thus, $\Delta\gamma < 0, \Delta\pi > 0$).

2. If the job test is *bias-reducing*—that is, if the test is less biased than are job interviews (formally, $\Delta\nu_\eta > \Delta\nu_s \geq 0$ or $0 \geq \Delta\nu_s > \Delta\nu_\eta$)—it unambiguously raises productivity.[23] Intuitively, a bias-reducing test improves hiring through two channels: (1) raising screening precision and (2) reducing excess hiring of the group favored by interviews (thus, increasing selectivity for this group). Both effects are productivity-enhancing.

3. By contrast, a *bias-increasing* test ($|\Delta\nu_s| > |\Delta\nu_\eta|$) has ambiguous effects on productivity. Although testing always raises screening precision—which is productivity-enhancing—a bias-increasing test causes excess hiring of the group that is favored by the bias, which is productivity-reducing. The net effect depends on the gains from increased screening precision relative to the losses from increased bias.

Figure IIIb illustrates result (2). Here, we simulate the impact of testing on hiring and productivity for a case where minority applicants are less productive than majority applicants and job interviews are minority-favoring.[24] Prior to testing (equivalently, $\rho = 0$ in the figure), the majority/minority hiring gap is small and the majority/minority productivity gap is large relative to a setting with no biases (Figure IIIa). This contrast with Figure IIIa reflects the fact that a minority-favoring interview raises minority hiring and reduces minority productivity. Job testing counteracts this bias, leading to a marked decline in the hiring of minority applicants and an equally marked decline in the productivity gap between majority and minority hires (with the magnitude depending upon test precision).[25] Thus, an unbiased job test increases efficiency at the expense of equality if job interviews are biased in favor of minorities.

Summarizing our three main conclusions: if job tests are relatively unbiased, they do not pose an equality-efficiency trade-off; if job tests are bias-reducing, they pose an equality-efficiency trade-off if and only if interviews are minority-favoring; if job tests are bias-enhancing, they may pose an equality-efficiency trade-off—or they may simply reduce equality and efficiency simultaneously.

[23] However, if the test and interview have biases of *opposite* sign ($\text{Sign}\langle\Delta\nu_s\rangle = \text{Sign}\langle\Delta\nu_\eta\rangle$), testing does not necessarily increase productivity even if job tests are *less* biased than interviews.

[24] We use the same parameter values as in Figure IIIa except that we now assume that $\Delta\nu_\eta = \mu_0(x_2) - \mu_0(x_1) = -0.19$.

[25] In the limiting case where job tests are fully informative, the unbiased and biased-interview cases converge to the same hiring rates and productivity levels.

G Empirical implications

Our illustrative model contains many specific—albeit, we believe reasonable—assumptions and so it is unwise to generalize too broadly based on this analysis. In fact, a key purpose of the conceptual framework is to demonstrate that, contrary to an influential line of reasoning, job testing does not pose an *intrinsic* equality-efficiency trade-off, even if minority applicants perform worse than majority applicants on job tests.

Beyond this observation, three general conclusions are warranted. First, the potential effects of job testing on minority hiring depend primarily on the biases of job tests *relative to* job interviews (and other existing screening methods). Job tests that are unbiased relative to job interviews are unlikely to reduce minority hiring because such tests do not adversely affect firms' average assessments of minority productivity.

Second, testing is likely to reduce minority hiring when tests are relatively biased against minorities (i.e., relative to interviews). In such cases, testing conveys 'bad news' about the productivity of minority relative to majority applicants and so is likely to adversely affect minority hiring. Nevertheless, if testing mitigates existing biases, it will still be efficiency-enhancing, and so an equality-efficiency trade-off will be present. If instead testing augments bias, it may be efficiency-reducing.

Finally, testing will generally have opposite effects on the hiring and productivity gaps between majority and minority workers; a test that reduces minority hiring will typically differentially raise minority productivity. This implication proves particularly useful for our empirical analysis.

Below, we use this model to interpret the empirical findings in light of their implications for the relative biases of the job test and the informal screen that preceded it. To make this interpretation rigorous, we parametrically simulate the model in section VI using observed applicant, hiring and productivity data to calculate a benchmark for the potential impacts of job testing on the majority/minority hiring and productivity gaps under alternative bias scenarios.

IV Estimating the productivity consequences of job testing

Our model is predicated on the assumption that job testing improves productivity. We verify this assumption here. The productivity measure we study is the length of completed job spells of workers hired with and without use of job testing. While job duration is clearly an incomplete measure of productivity, it is likely to provide a good proxy for worker reliability since unreliable workers are likely to quit unexpectedly or to be fired for poor performance. Notably, the firm whose data we analyze implemented job testing precisely because managers believed that turnover was too high. In the working paper version of this article [Autor and Scarborough, 2004], we also consider a second

productivity measure, involuntary terminations, and find results consistent with those below.

We begin with the following difference-in-difference model for job spell duration:

$$D_{ijt} = \alpha + \delta T_i + X_i\beta + \theta_t + \varphi_j + e_{ijt}, \tag{7}$$

where the dependent variable is the job spell duration (in days) of worker i hired at site j in year and month t. The X vector includes worker race and gender, and T is an indicator variable equal to one if the worker was screened using the job test, and zero otherwise. The θ vector contains a complete set of month $\times$ year-of-hire effects to control for seasonal and macroeconomic factors affecting job spell durations. Our main specifications also include a complete set of store effects, φ, which absorb time invariant 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.[26] For these and all subsequent models, we have also experimented with clustering the standard errors on stores' month-by-year of adoption to account for potential error correlations among adoption cohorts. These standard errors prove comparable to our main estimates.

Consistent with the bivariate comparisons in Table I, the estimate of equation (7) in column 1 of Table IV confirms that Black and Hispanic workers have substantially lower mean tenure than White employees. When $1,363$ site fixed effects are added to the model, these race differences are reduced by approximately 40 percent, indicating that minority workers are overrepresented at establishments where both minority and majority workers have high turnover. Nevertheless, race differences in tenure remain highly significant and economically large.

Columns 3 and 4 of Table IV show that job testing raises job spell durations significantly. In models excluding site effects and race dummies, we estimate that workers hired using the employment test worked 8.8 days longer than those hired without use of the employment test. When site fixed effects are added, this point estimate rises to 18.8 days.[27] Adding controls for worker race and gender does not change the magnitude or significance of these job-test effects. When state $\times$ time interactions are added in column 6 to account for differential employment trends by state, the main estimate rises slightly to 22.1 days. This represents about a 12 percent gain in average tenure relative to the pre-testing baseline. Models that include a full set of state $\times$ month-year-of-hire interactions (not tabulated) yield nearly identical and highly significant results.

The tenure gains accruing from job testing are also visible in Figure IV, which plots the density and cumulative distribution of completed job spells of tested and non-tested hires. The distribution of spells

[26] We exclude from these models the 2 percent of spells that are incomplete.

[27] The flow of hires in our sample intrinsically overrepresents workers hired at high-turnover stores (relative to the stock of hires). When testing is introduced, a disproportionate share of tested hires are therefore hired at high turnover establishments. Adding site effects to the model controls for this composition bias and hence raises the point estimate for the job testing indicator variable (compare columns 3 and 4).

for tested hires lies noticeably to the right of the distribution for non-tested hires and generally has greater mass at higher job durations and lower mass at shorter durations. As shown in the lower panel of the figure, the job spell distribution of tested hires almost entirely first order stochastically dominates that of non-tested hires. Quantile regression estimates for job spell durations (not tabulated) confirm that the effect of testing on job spell duration is statistically significant and monotonically increasing in magnitude from the 10^{th} to the 75^{th} percentiles. These models find that testing increased median tenure by 8 to 9 days, which is roughly a 10 percent gain (thus comparable to the estimated effect at the mean).

A Endogeneity of testing?

Our findings could be biased if the decision to test a worker is endogenous. We observe that in the one to two months following the rollout of testing at a site, 10 to 25 percent of new hires are not tested. This may be due to operational issues following system installation (i.e., the kiosk is offline) or due to record-keeping lags wherein workers offered a job shortly before the advent of testing do not appear on the payroll until after testing is already in use. A more pernicious concern, however, is that managers could circumvent testing to hire preferred candidates—a potential source of endogeneity bias.

To purge any potential endogeneity, we estimate a two-stage least squares (2SLS) version of equation (7) in which we use a dummy variable indicating that a store has adopted testing 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 testing kiosk at a store, we use the date of the first observed tested hire to proxy for the rollout date. The coefficient on the store-adoption dummy in the first stage equation of 0.89 indicates that once a store has adopted testing, the vast majority of subsequent hires are tested.[28] The 2SLS estimates of the effect of testing on job spell durations shown in columns 7 through 10 of Table IV are quite similar to the corresponding OLS models, suggesting that endogeneity of individual test status is not a substantial source of bias.

A second source of endogeneity is that the timing of stores' adoption of testing might be correlated with potential outcomes. Although all stores in our data adopt testing during the sample window, the timing of adoption is not necessarily entirely random. To our understanding, the rollout order of stores was determined by geography, technical infrastructure, and internal personnel decisions. If, however, 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.[29]

[28] A table of first-stage estimates for the 2SLS models is available from the authors.

[29] Managers who we interviewed were not aware of any consideration of store-level personnel needs in the choice of rollout order. They also pointed out that timely analysis of store-level personnel data was not feasible prior to the Unicru

As a check on this possibility, we augmented equation (7) for job spell duration with leads and lags of test adoption. These models (available from the authors) capture the trend in job spell durations for workers hired at each store in the nine months surrounding introduction of testing: four months prior to three months post adoption. While the lead estimates in these models are in no case significant and have inconsistent signs, the lag (post-rollout) dummies show that workers hired in the first month of testing have 12 days above average job spell duration, and workers hired in subsequent months have 17 to 25 days above average duration (in all cases significant). Thus, our main estimates do not appear confounded by pre-existing trends in job spell duration.[30]

B Do test scores predict productivity?

It would be valuable to corroborate these findings by showing that test scores predict worker productivity. We would ideally proceed by regressing gains in store level productivity on gains in test scores for cohorts of workers hired at the same stores before and after the advent of job testing. Our strong expectation is that stores that saw greater increases in worker 'quality' as measured by test scores would have experienced larger gains in productivity. Unfortunately, the firm that we study did not collect baseline test score data for cohorts of workers hired prior to the use of testing. As an alternative, we draw on the database of $189,067$ applications submitted to the $1,363$ stores in our sample during the year *after* the rollout of employment testing (Table II). Under the assumption that the characteristics of applicants by store were stable before and after the introduction of job testing, these data can be used to benchmark the relationship between test scores and productivity.[31]

We estimate the following variant of our main model for worker tenure:

$$D_{ijst} = \alpha + \zeta \bar{S}_j + X_i\beta + \theta_t + \chi_s + e_{ijt}. \tag{8}$$

Here, the dependent variable is the completed job spell duration of workers hired at each store j in state s, $\bar{S}_j$ is the average test score of store $j's$ applicants, which serves as a proxy for the average 'quality' of applicants at the store, and χ_s is a vector of state dummies. If test scores are predictive of worker productivity (as our analysis so far suggests) stores with lower average applicant quality should exhibit lower overall productivity prior to the advent of testing.

Table V presents estimates. Column 1 finds a sizable, positive relationship between store-level average applicant 'quality' and the productivity of workers hired prior to the use of testing. A one-

installation.

[30] We also estimated a version of equation (7) 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 $p < 0.05$. None of the 3 negative point estimates is significant.

[31] We cannot link test scores of workers in the primary sample to their employment outcomes, however.

standard deviation difference in mean applicant quality (cross store standard deviation of 0.16) predicts a 6.3 day difference in mean store job duration. Since we cannot include site effects in these cross-sectional models, subsequent columns add controls for state-specific trends and measures of minority resident share and median household income in the store's zip code (calculated from the Census STF files). These covariates have little impact on the coefficient of interest.

The next three columns provide estimates of the relationship between store-level productivity and mean test scores of *hired* workers. These point estimates are likely to be substantially attenuated by measurement error since the hired sample used to calculate the means is only 10 percent as large as the applicant sample. Despite this, we find an equally large coefficient for the test-score variable. Because the cross-store standard deviation of hired worker test scores (0.32) is about twice as large as the cross-store standard deviation of applicant test scores, the standardized effect size of 13 days of tenure is also twice as large. When we instrument the test scores of hires with those of applicants to reduce the influence of measurement error (columns 7 through 9), the point estimate on the test score variable increases in magnitude by about a third. Taken together, these results demonstrate that job test scores have significant predictive power for worker productivity.

C Testing for differential productivity impacts by race

A central implication of the model is that if job tests are unbiased relative to job interviews, they will raise the productivity of majority and minority hires equally. Conversely, if tests are relatively biased against minorities, they should raise the productivity of minority hires by more than majority hires (and vice versa if tests are majority-favoring). To assess the impact of testing on the productivity of majority and minority hires, we estimate an augmented version of equation (7) where we replace the 'tested' dummy variable with a full set of interactions between tested-status and the three race groups in our sample:

$$D_{ijt} = \alpha + \delta_w T_i \times \text{White}_i + \delta_b T_i \times \text{Black}_i + \delta_h T_i \times \text{Hispanic}_i + X_i\beta + \theta_t + \varphi_j + e_{ijt}. \tag{9}$$

The parameters of interest in this equation, δ_b, δ_h and δ_w, estimate the differential gains in job spell duration for tested Black, Hispanic and White hires relative to their non-tested counterparts.

Table VI presents OLS and 2SLS estimates of equation (9). In the baseline specification in column 1, which excludes site effects and state trends, we estimate that job testing raised spell durations by 14 days among White hires, 15 days among Black hires, and -1.2 days among Hispanic hires. When site effects and state trends are added, these point estimates rise to 23 days for both Black and White hires and 13 days for Hispanic hires. The tenure gains for Whites and Blacks are highly significant. Those for Hispanics are significant at the 10 percent level in the final specification but not otherwise.

A test of the joint equality of the tenure gains by race accepts the null at $p = 0.36$. In subsequent columns, we present 2SLS estimates using site adoption of testing as an instrument for whether or not an individual hire received the employment test. These models show comparable patterns.

In net, we find that testing had similar impacts on productivity for all worker groups. In the case of Black versus White productivity gains, the point estimates are extremely close in magnitude in all models. Although estimated tenure gains are smaller for Hispanic hires than for other groups, the data do not reject the hypothesis that tenure gains are identical for all three groups (Whites, Black and Hispanics).

D Robustness checks

A natural concern with job spell duration as a productivity measure is that it captures quantity but not quality of labor input. Consider, for example, that college students hired during their summer breaks may be more capable or reliable than average workers and yet may have shorter job stints. As one check on this possibility, we reestimated all models in Tables IV and VI while excluding all workers hired in May, June, November and December—that is, the cohorts most likely to include seasonal hires. These estimates, available from the authors, are closely comparable to our main results.

To supplement the evidence from the job duration analysis, it would be valuable to have a more direct measure of worker productivity. In Autor and Scarborough [2004], we explore one such measure: firing for cause. Using linked personnel records, we distinguished for-cause terminations (e.g., theft, job abandonment, insubordination) from neutral or positive terminations (e.g., return to school, relocation, new employment). Consistent with the results for job tenure above, these models find that job testing modestly reduced firing-for-cause and significantly reduced voluntary turnover without yielding differential impacts on minority and majority hires.

V Testing for disparate impacts of testing on minority hiring

Did the productivity gains from testing come at a cost of reduced minority hiring? Although the test score distributions of Black, White and Hispanic job applicants differ significantly (Figure II), the test score distributions of Black, White and Hispanic *hires* are quite comparable (Figure V)). The reason is that the hired population from each race group almost entirely excludes the lower tail of the applicant distribution, where a disproportionate share of Black and Hispanic applicants reside. The contrast between Figures II and V suggests that disparate hiring impacts were a real possibility.

Yet, initial evidence suggests that a disparate impact did not occur. Unconditional comparisons of hiring by demographic group in Table I show a slight *increase* in minority employment after the implementation of job testing. To provide a rigorous test, we contrast changes in minority versus

majority hiring at stores adopting testing relative to stores not adopting during the same time interval. The outcome variable of interest is the minority hiring *rate*, equal to the flow of minority hires over minority applicants. Unfortunately, our data measure the flow of hires but not the flow of applicants. To see the complication this creates, let $\Pr(x_2|Hire=1)$ equal the probability that a new hire is a minority worker and let $\Pr(x_1|Hire=1)$ be the corresponding probability for a majority worker. Applying Bayes rule, we obtain

$$\ln\left(\frac{\Pr(x_2|Hire=1)}{\Pr(x_1|Hire=1)}\right) = \ln\left(\frac{E(Hire|x_2)}{E(Hire|x_1)}\right) + \ln\left(\frac{\Pr\{x_2\}}{\Pr\{x_1\}}\right).$$

The odds that a newly hired worker is a minority is a function of both the minority/majority hiring rate and the minority/majority application rate ($\Pr\{x_2\}/\Pr\{x_1\}$). Since we lack data on application rates, we must assume that minority/majority application rates are roughly constant within stores over time to isolate the pure impact of testing on the minority/majority hiring rate.[32]

To perform this estimate, we fit the following conditional ('fixed-effects') logit model,

$$\Pr(x_2|Hire=1) = F\left(\xi T_i + X_i\beta + \theta_t + \varphi_j\right), \tag{10}$$

where x_2 indicates that a hired worker is a minority, the vectors φ and θ contain a complete set of store and month-by-year of hire dummies, and $F(\cdot)$ is the cumulative logistic function.[33] The coefficient, ξ, measures the total effect of job testing on the log odds that a newly hired worker is a minority. If our assumption is correct that minority/majority application rates are roughly constant within stores, these rates will be 'conditioned out' by the store fixed effects and $\hat{\xi}$ will capture the impact of job testing on the minority hiring rate. If this assumption fails, $\hat{\xi}$ still provides a valid estimate of the causal effect of job testing on minority hiring but in that case, we cannot separate the effect of testing on application versus hiring rates.

The top panel of Table VII reports estimates for the change in the log hiring odds of Black, White and Hispanic applicants. These models yield little 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 ($z < 1$).

As a supplemental test, we fit in panel B of Table VII a simple fixed-effects, linear probability model of the form:

[32] Unicru personnel believe that the application kiosks attract more job seekers overall but have no information on how kiosks affect application rates by race. One might speculate that the kiosks discourage minority applicants since minorities are disproportionately likely to have criminal records [Petit and Western, 2004] and completing the electronic application requires providing a social security number and authorizing a criminal background check. Such discouragement would bias our results towards the finding that job testing reduced the minority hiring rate.

[33] We use a conditional logit model to avoid the incidental parameters problem posed when estimating a conventional maximum likelihood model with a very large number of fixed effects (1,363).

$$E\left(x_2|Hire=1\right)=\alpha+\psi T_i+X_i\beta+\theta_t+\varphi_j. \tag{11}$$

This model measures the effect of testing on the *share* of hires who are minorities. So that coefficients may be read as percentage points, point estimates and standard errors are multiplied by 100. In all cases, the estimated impact of testing on hiring rates by race is under one half of one percentage point and insignificant. The 2SLS estimates of these models (panel C) are similar to the corresponding OLS models, implying an even smaller reduction in Black hiring and a slightly larger reduction in Hispanic hiring. These point estimates suggest that testing had negligible effects on the race distribution of workers.[34]

A Disparate hiring impacts: A second test

Since the hiring results are central to our conclusions, we test their robustness by analyzing a complementary source of variation. Prior to the advent of testing, we observe a tight link between the neighborhoods in which 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 use this link to explore whether testing changed the relationship between stores' neighborhood demographics and the race of hires. Specifically, we estimate a variant 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 STF-1 files.

Column 1 of Table VIII shows that, prior to the use of job testing, a store situated in a zip code with a 10 percentage point higher share of non-White residents would be expected to have an 8.7 percentage point higher share of non-White employees. The point estimate in column 2 shows that this relationship was essentially unchanged by testing. The next two columns 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, this interaction term is close to zero and insignificant. When site dummies are included (column 4)—thus absorbing the main effect of the zip code's non-White resident share—the interaction term is again small and insignificant.

Panel B provides analogous estimates for the relationship between the racial composition of employees and neighborhood household income. In the pre-testing period, stores in more affluent zip

[34] Lead-and-lag estimates for the effect of testing on race composition are generally insignificant and do not have consistent signs. The point estimates suggest a brief rise in black hiring and decline in white hiring in the first three months following the introduction of testing, followed by a slight reduction in Black hiring and rise in White hiring in months three forward. These latter effects are far from statistically significant. A table of estimates is available from the authors.

codes had a substantially higher share of White employees: 10 additional log points of neighborhood household income was associated with a 3.2 percentage point greater share of White employees. Employment testing did not alter this link. For all demographic groups, and for both measures of neighborhood demographics, the pre-post change in the relationship between neighborhood characteristics and the group membership of hires is insignificant and is close to zero in the model with site dummies.

VI What conclusions do the data support? A model-based test

In this final section, we apply the theoretical model from section III to synthesize and interpret the findings above. Drawing on the applicant, hiring and productivity databases summarized in Tables I and II, we parametrically simulate the model to assess what combinations of interview bias, test bias, and underlying majority/minority productivity differences are most consistent with the findings. One overriding conclusion emerges from this exercise: the data readily accept the hypothesis that both job tests and job interviews are unbiased and that the average productivity of White applicants exceeds that of Black applicants. By contrast, the plausible alternatives that we consider—most significantly, that the job test is relatively biased against minorities—are rejected.

A Simulation procedure

Let observed job spell durations, D, be a linear function of applicant ability y, with $D = \alpha + \vartheta y$, where $\vartheta > 0$ is a parameter to be estimated from the data. Suppose that the ability of an applicant drawn at random from the distribution of group x applicants is equal to $y = \mu_0(x) + \varepsilon_0$. Prior to the introduction of job testing, firms have access to an *interview signal*, η, for each applicant that is correlated with ability. When job testing is introduced, it provides a second signal, s, that is also correlated with ability. We assume initially that both interviews and tests are unbiased, with $\eta = y + \varepsilon_\eta$ and $s = y + \varepsilon_s$. In these expressions, ε_0, ε_s and ε_η are mean-zero error terms that are normally and independently distributed, with variances to be estimated from the data.

To estimate the variance parameters, we use the following empirical moments: the mean test score of applicants is normalized to zero and the mean test score of workers hired using the test is 0.707 (Table II); the variance of test scores is normalized at one (hence, $1 = \sigma_0^2 + \sigma_s^2$);[35] the observed hiring rate is equal to 8.95 percent; and the average gain in productivity from testing is 21.8 days (Table IV). We make a further adjustment for the fact that the observed hiring rate is only 22 percent at the 95th percentile of the score distribution (see Table II), implying either that stores are

[35] It is the ratio of variances (σ_η^2/σ_0^2, σ_s^2/σ_0^2), not their levels, that determines the informativeness of the signals. Thus, the normalization that $\sigma_0^2 + \sigma_s^2 = 1$ is innocuous.

extraordinarily selective or, more plausibly, that a portion of applicants is turned away because there are no vacancies.[36] Since ability is unobserved, we cannot directly estimate the structural relationship between ability and job spell duration, ϑ. Instead, we use the empirical relationship between test scores and productivity from Table V ($\hat{\zeta} = 53.9$ in equation (8)) to calculate the implied value of ϑ based on other moments of the model. Putting these pieces together, we calculate that $\hat{\sigma}_s^2 = 0.71$, $\hat{\sigma}_\eta^2 = 0.45$ and $\hat{\sigma}_0^2 = 0.29$. Hence, test scores have approximately 60 percent more measurement error than do interviews.[37]

Using these parameter estimates in combination with the database of 189,067 applications summarized in Table II, we implement the following simulation procedure:[38]

1. For each applicant, we draw a simulated ability level, y, as a function of the applicant's observed test score and the estimated error variance of the test. Although this simulated ability level is not observed by employers, it contributes to applicants' interview and test scores and completely determines their job spell durations conditional on hire.

2. Using the ability draws and the estimated variance parameters, we draw an 'interview signal' for each applicant. In contrast to applicant ability levels, these interview signals are observed by firms and are used for hiring.

3. Using applicants' interview signals, their race, and firms' priors, we calculate firms' 'interview-based' posterior productivity assessment for each applicant (see equation (2)).

4. We then simulate hiring under the interview-based regime by calculating a store-specific interview-based hiring threshold such that the count of applicants whose interview-based posterior assessment meets the threshold exactly equals the count of hires observed at the store. Applicants meeting the threshold are labeled 'interview-based hires.'

5. We next use the draws of ability, y, to calculate the job spell durations of interview-based hires (equal to $D = \hat{\alpha} + \hat{\vartheta} y$). In combination, steps (4) and (5) allow us to calculate the race composition and productivity of hires (both overall and by race) under the *interview-based* regime.

[36] To adjust for vacancies, we estimate the hiring rate conditional on a vacancy ('active hiring rate') by calculating what the model implies that the hiring rate *should be* at the 95th percentile of the test score distribution given other estimated parameters. If the observed rate is lower than the calculated rate, we attribute the difference to lack of vacancies and impute the active hiring rate as the ratio of the implied hiring rate to the observed hiring rate. In practice, the active hiring rate is solved simultaneously with the other parameters of the model since they are not independent. We estimate the active hiring rate at 40.4%; that is, 4 in 10 applicants are hired when a vacancy is present.

[37] It would be highly surprising to find that tests are *more informative* than interviews since the item response data gathered by the personality test appear (to us) crude relative to the nuances of attitude and behavior observable during interviews.

[38] We sketch the procedure here, with further detals available in an unpublished appendix.

6. To obtain analogous outcomes under the *test-based* regime, we repeat steps (3) through (5), making two modifications to the procedure. First, we replace firms' interview-based posterior productivity assessments with their test-based posterior productivity assessments (see equation (4)).[39] Second, when performing the simulated hiring process in step (4), we replace the interview-based hiring threshold with a test-based hiring threshold that generates an identical number of hires at each store.

7. In the final step, we compare the race composition and productivity of hires (overall and by race) under the interview-based and test-based regimes. Since the distribution of ability and the hiring rate are identical at each store under each regime, a comparison of (simulated) hiring and productivity outcomes under these two regimes provides an estimate of the pure screening effect of testing on equality and efficiency.

This baseline procedure simulates the case where both interviews and tests are unbiased. It must be slightly extended to explore cases where test or interview biases are present. Table II shows that applicants from the majority group score significantly higher on the job test than applicants from the minority group. We accordingly consider two cases for test bias: in the first case, tests are *unbiased* and, by implication, minority applicants are on average less productive than majority applicants; in the second case, we assume that job tests are *majority-favoring* while minority and majority applicants have the same average productivity.[40]

We allow for the possibility of interview bias in a parallel fashion. Because the data provide no guidance on the possible sign of interview bias, we consider three cases: no bias, minority-favoring bias, and majority-favoring bias. In the unbiased case, the interview signal is equal to $\eta = y + \varepsilon_\eta$, as above. In the minority-favoring case, the interview signal additionally includes an additive bias term that precisely offsets the mean test score differences between minority and majority applicants. Conversely, in the majority-favoring case, the interview signal contains a bias of equal magnitude and opposite sign to the minority-favoring case.

These assumptions give rise to six permutations of the simulation: two cases of testing bias (unbiased and majority-favoring) permuted with three cases of interview bias (unbiased, minority-favoring and majority-favoring). For each scenario, we perform $1,000$ trials of the simulation to obtain mean outcomes and bootstrapped standard errors, equal to the standard deviation of outcomes across trials.

[39] These test-based posteriors differ from the interview-based posteriors *only* in that they incorporate both interview and test signals.

[40] Since we do not know the *true* mean ability of each applicant group—only the group's mean test score—we make the following ancillary assumptions: if job tests are unbiased, mean ability for each applicant group is equal to the groups's mean test score. If job tests are majority favoring, mean ability for each applicant group is equal to the White mean.

Because our focus is on Black-White differences, we discuss and tabulate results for only these two groups. Hispanics are included in the simulation, however.

B Simulation results

Table IX summarizes the simulation results. Columns 1 through 6 present the simulated productivity and hiring effects of testing under each of the six scenarios considered. For comparison, column 7 lists the actual outcome statistics for each productivity and hiring measure (from Tables I, VI and VII). The bottom row of each panel provides chi-squared statistics for the goodness of fit of the simulated outcomes to their observed counterparts.[41]

As shown in column 7, prior to the use of job testing, the unconditional mean job spell duration gap between White and Black hires was 45 days. Our analysis found that testing raised mean White and Black job spell durations by 23 days each, leading to no change in the productivity gap. Testing also yielded no significant change in the racial composition of hires, though our point estimates suggest a increase in the White employment share of 0.24 percentage points. How do the simulation results compare to the actual outcomes?

Only one of the six simulation scenarios closely corresponds to the data. This is the case where interviews and job tests are *unbiased* and average White productivity exceeds average Black productivity (column 1). Under this scenario, the simulation predicts a gain of 18.6 and 19.9 days respectively for White and Black job spells, as compared to an observed rise of 23.2 days. The simulation further implies a 52 day gap in mean job spell duration between White and Black applicants hired using the informal screen, as compared to the observed difference of 44.9 days. A chi-squared test of goodness of fit of these estimates (bottom row of panel A) readily accepts the null of equality between the observed and simulated statistics ($p = 0.50$). Alongside these productivity impacts, the simulation predicts a small rise in Black employment (panel B). This predicted value does not differ significantly from the small observed decline in Black employment.[42] An omnibus test of goodness of fit for both productivity and employment outcomes (panel C) accepts the null at $p = 0.33$.

It is also instructive to consider the cases that do not fit the observed outcomes. The alternative scenario that comes closest to matching the data is one in which job interviews are *biased towards*

[41] To compare each simulated statistic with its observed value, we calculate the following chi-square square statistic with one degree of freedom:

$$\chi^2(1) = \left(\frac{[\text{sim} - \text{obs}]}{(SE^2_{\text{sim}} + SE^2_{\text{obs}})^{1/2}} \right)^2 .$$

To calculate pooled summary tests for each simulation scenario, we sum the χ^2 statistics and the degrees of freedom (thus, treating each statistic as independent). When performing pooled tests, we exclude redundant statistics. For example, we include the changes in White and Black productivity but exclude the change in the Black-White productivity gap.

[42] We have also performed goodness of fit tests for changes in log-odds of hiring rather than changes in employment shares. Results are similar to those tabulated but statistical power is lower.

whites, the job test is unbiased and the expected productivity of White applicants exceeds that of Black applicants (column 2). As in the prior case, the simulation suggests comparable tenure gains for Whites and Blacks of 20.4 and 19.7 days.[43] Here, however, the predicted initial productivity gap of 30.1 days falls far short of the observed difference of 44.9 days. Where this scenario departs most substantially from the data, however, is in its predictions for minority hiring. Because the job test is assumed to be relatively minority-favoring, the simulation predicts a substantial gain in Black employment, leading to a closing of the White-Black employment gap of 4.1 percentage points. This prediction is rejected by the data since the actual change in White-Black employment is negligible.

A second scenario of particular interest is shown in Column 3. Here, the informal screen is minority-favoring, the job test is unbiased (thus, the job test is *relatively* biased against minorities) and average White productivity exceeds average Black productivity. As per the Introduction, this is the focal case where job testing could produce a disparate impact—reducing Black hiring while raising productivity differentially for Blacks hires (as well as overall). Consistent with expectations, the simulation predicts a significant differential gain in job duration of 6.3 days for Black relative to White hires and a small decline in Black employment. In practice this scenario is rejected by the data ($p = 0.00$) since there was neither a differential gain in Black productivity nor a fall in Black hiring. We therefore reject the presence of an equality-efficiency trade-off in this setting.

Consider finally the cases in columns 4 through 6, where Black and White applicants are assumed to have identical mean productivity. These scenarios are at odds with the data in one key respect: all predict substantially smaller minority-majority gaps in initial productivity than are observed in the data; in two of three cases, these gaps are of the wrong sign. This pattern underscores that it is difficult to square the hypothesis that minority and majority applicants have the same underlying productivity with the fact that job spell durations of minority hires are substantially shorter than those of majority hires. To reconcile these two observations, one must assume, as in column 6, that job interviews are heavily minority-favoring. Under this assumption, however, job testing is predicted to substantially raise White employment, which does not occur.

In net, the simulation results reject each of the scenarios considered except one in which both job interviews and job tests are unbiased and White applicants are on average more productive than Black applicants. This leads us to conclude that job testing increased the precision of worker selection without yielding disparate impacts because both job interviews and job tests were unbiased.

[43] Given that the job test in this scenario is relatively biased towards minorities, one may wonder why the model does not predict a greater rise in the productivity of White relative to Black hires. We do not have a precise answer to this question, but believe it stems from the fact that the distribution of Black productivity is relatively flat near the hiring threshold. Thus, a marginal increase in the selectivity of Black hires does not yield a substantial change in the productivity of black hires.

VII Conclusion

An influential body of research concludes that the use of standardized job tests for employment screening poses an intrinsic equality-efficiency trade-off: testing improves selection but reduces minority hiring. We develop a simple conceptual model that demonstrates that this equality-efficiency trade-off is only unambiguously present when job tests counter pre-existing screening biases favoring minority applicants. By contrast, if job tests are unbiased relative to the screens they supplement, there is no equality-efficiency trade-off—gains in productivity do not come at a cost of lower minority hiring. Since we see little reason to suspect that existing informal screens are minority-favoring or that job tests are *more* biased against minorities than are informal screens, we believe that the presumed equality-efficiency trade-off likely to be largely irrelevant in practice.

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 one year from informal, paper-based hiring to a computer-supported screening process that relies heavily on a standardized personality test. The advent of employment testing increased productivity, raising mean and median employee tenure by 10 to 12 percent and slightly lowering the frequency of terminations for cause. Consistent with expectations, minority applicants performed significantly worse on the employment test. Had the informal screen that predated testing been comparatively minority-favoring such that it masked underlying majority/minority differences in average productivity, our model suggests that employment testing would have slightly diminished Black employment and raised the productivity of Black hires by approximately 40 percent more than it raised the productivity of White hires. In point of fact, we detect no change in the racial composition of hires and, moreover, productivity gains were equally large among minority and majority hires. These findings, paired with evidence from a parametric simulation of our theoretical model, lead us to conclude that both the job test and the informal screen were unbiased. Thus, job testing did not pose an equality-efficiency trade-off because the job test raised screening precision without introducing bias.[44]

In considering the external validity of these findings, several caveats are in order. First, our data come 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. Nevertheless, analysis of other cases is clearly warranted. A second caveat is that the between-group differences in test scores found

[44] A final alternative interpretation of the findings is that minority and majority workers *are* equally productive but that the productivity measures are themselves contaminated by race bias. In this case, our findings would indicate only that the job test was unbiased *relative* to the informal screen but not necessarily unbiased in a cardinal sense. While we cannot dismiss this alternative hypothesis using available data, previous studies that benchmark standardized tests scores against objective productivity measures in civilian and military settings find no evidence to suggest that such tests are race-biased [Hartigan and Wigdor, 1989; Wigdor and Green, 1991; and Jencks and Philips, 1998, chapter 2].

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. This fact limits the power of our analysis to distinguish competing scenarios, and one might posit that an alternative employment test that revealed larger group productivity differences might potentially generate disparate impacts. Although we do not discount this possibility, 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 pose an equality-efficiency trade-off. Moreover, employment testing guidelines issued by the Equal Employment Opportunity Commission make it legally precarious for firms to use employment tests that 'pass' minority applicants at less than 80 percent of the pass-rate of majority applicants. Hence, employment tests will not generally show greater group differences than those found here.

An important final point of interpretation is that our results speak only to firms' private gains from improved worker selection. The extent to which these private gains translate into social benefits depends 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 [Costrell and Loury 2004]. 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; Masters, 2006]. Moreover, since testing itself is costly, the net social benefits in the pure screening case could potentially be negative. Though our results provide little guidance as to which of these scenarios is most relevant, it appears unlikely that the social benefits from testing exceed the private benefits. Quantifying these social benefits is an important area for research.

VIII Appendix: Proofs of selected propositions

A Preliminaries

Suppose there is no formal testing. Decompose η^* as $\eta^* = \eta + \nu_\eta(x)$; an applicant from group x is hired iff $\mu(x,\eta^*) > \kappa_I$, where $\mu(x,\eta^*) = \frac{h_\eta}{h_I}\eta^* + \frac{h_0}{h_I}[\mu_0(x)+\nu_\eta(x)] = \mu(x,\eta)+\nu_\eta(x)$. Thus the applicant is hired iff

$$\begin{aligned} \mu(x,\eta) &= \frac{h_\eta\eta + h_0\mu_0(x)}{h_I} > \kappa_I - \nu_\eta(x) \\ \eta &> \frac{h_I(\kappa_I - \nu_\eta(x)) - h_0\mu_0(x)}{h_\eta}. \end{aligned}$$

Since we can decompose η as $\eta = y + \varepsilon_\eta$, with $\varepsilon_\eta \sim N(0, 1/h_\eta)$ and independent of y, the distribution of η within group x is $N\left(\mu_0(x), \frac{1}{h_0} + \frac{1}{h_\eta}\right)$. So we can write the condition for an applicant to be hired

as

$$\begin{aligned}
\frac{\eta-\mu_0(x)}{\frac{1}{h_0}+\frac{1}{h_\eta}} &> \frac{h_I[\kappa_I-\nu_\eta(x)-\mu_0(x)]}{h_\eta(\frac{1}{h_0}+\frac{1}{h_\eta})^{1/2}} \\
&= \frac{h_I[\kappa_I-\nu_\eta(x)-\mu_0(x)]}{(h_\eta h_I/h_0)^{1/2}} \\
&= \left(\frac{h_0 h_I}{h_\eta}\right)^{1/2}[\kappa_I-\nu_\eta(x)-\mu_0(x)] \\
&= \frac{\kappa_I-\nu_\eta(x)-\mu_0(x)}{\sigma_0\rho_I} \\
&\equiv z_I^*(x),
\end{aligned}$$

where $\rho_I \equiv Corr\left[\mu\left(x,\eta\right),y\right] = \left(1-h_0/h_I\right)^{1/2}$ and the probability that an applicant from group x is hired is $1-\Phi(z_I^*(x))$.

Decomposing y as $y=\mu_0(x)+\varepsilon_y$, expected productivity conditional on hire is

$$\begin{aligned}
E_\eta\left[y|Hire=1,x\right] &= \mu_0(x)+E_\eta\left[\varepsilon_y|\frac{\eta-\mu_0(x)}{\frac{1}{h_0}+\frac{1}{h_\eta}}>z_I^*(x),x\right] \\
&= \mu_0(x)+\sigma_0\rho_I E\left[\frac{\eta-\mu_0(x)}{\frac{1}{h_0}+\frac{1}{h_\eta}}|\frac{\eta-\mu_0(x)}{\frac{1}{h_0}+\frac{1}{h_\eta}}>z_I^*(x),x\right] \\
&= \mu_0(x)+\sigma_0\rho_I\lambda(z_I^*(x)).
\end{aligned}$$

Consider now the case with formal testing. Decomposing s^* as $s^*=s+\nu_s(x)$, an applicant is hired iff $\mu(x,\eta^*,s^*)>\kappa_T$, where $\mu(x,\eta^*,s^*)=\frac{h_s}{h_T}s^*+\frac{h_I}{h_T}\mu(x,\eta^*)=\mu(x,\eta,s)+\frac{h_s}{h_T}\nu_s(x)+\frac{h_I}{h_T}\nu_\eta(x)$. Thus the applicant is hired iff

$$\begin{aligned}
\mu(x,\eta,s) &= \frac{h_s s+h_\eta\eta+h_0\mu_0(x)}{h_T} \\
&= \frac{(h_s+h_\eta)y+h_s\varepsilon_s+h_\eta\varepsilon_\eta+h_0\mu_0(x)}{h_T} \\
&> \kappa_T-\frac{h_s}{h_T}\nu_s(x)-\frac{h_I}{h_T}\nu_\eta(x),
\end{aligned}$$

where we write $s = y + \varepsilon_s$ and $\eta = y + \varepsilon_\eta$. Since y, ε_s, and ε_η are independent and $h_s + h_\eta + h_0 = h_T$,

$$\begin{aligned}
\frac{h_s s + h_\eta \eta + h_0 \mu_0(x)}{h_T} &\sim N\left(\mu_0(x), \frac{(h_s + h_\eta)^2}{h_0 h_T^2} + \frac{h_s}{h_T^2} + \frac{h_\eta}{h_T^2}\right) \\
&\sim N\left(\mu_0(x), \frac{1}{h_0 h_T^2}(h_T^2 - h_0^2 - h_0 h_s - h_0 h_\eta)\right) \\
&\sim N\left(\mu_0(x), \frac{h_T - h_0}{h_0 h_T}\right) \\
&\sim N\left(\mu_0(x), \sigma_0^2 \rho_T^2\right).
\end{aligned}$$

It follows that an applicant is hired iff

$$\begin{aligned}
\frac{\mu(x,\eta,s) - \mu_0(x)}{\sigma_0 \rho_T} &> \frac{\kappa_T - \mu_0(x) - \frac{h_s}{h_T}\nu_s(x) - \frac{h_I}{h_T}\nu_\eta(x)}{\sigma_0 \rho_T} \\
&\equiv z_T^*(x),
\end{aligned}$$

and the probability of a hire is $1 - \Phi(z_T^*(x))$.

By the same reasoning as above, expected productivity conditional on being hired under formal testing is

$$\begin{aligned}
E_{\eta,s}\left[y | Hire = 1, x\right] &= \mu_0(x) + E_{\eta,s}\left[\varepsilon_y | \frac{\mu(x,\eta,s) - \mu_0(x)}{\sigma_0 \rho_T} > z_T^*(x)\right] \\
&= \mu_0(x) + \sigma_0 \rho_T \lambda(z_T^*(x)),
\end{aligned}$$

where $\rho_T \equiv Corr\left[\mu\left(x,\eta,s\right), y\right] = \left(1 - h_0/h_T\right)^{1/2}$.

We will also need to make use of the following lemma:

Lemma Let $\lambda(z) = \frac{\phi(z)}{1-\Phi(z)}$. Then $\lim_{z\to\infty} \lambda(z) - z = 0$.

Proof We have

$$\lim_{t\to\infty} \lambda(t) - t = \lim_{t\to\infty} \frac{\frac{1}{\sqrt{2\pi}} e^{-t^2/2} - t[1 - \Phi(t)]}{[1 - \Phi(t)]}$$

Applying l'Hopital's rule,

$$\begin{aligned}
\lim_{t\to\infty} \lambda(t) - t &= \lim_{t\to\infty} \frac{-\frac{1}{\sqrt{2\pi}} t e^{-t^2/2} - [1-\Phi(t)] + \frac{1}{\sqrt{2\pi}} t e^{-t^2/2}}{-\frac{1}{\sqrt{2\pi}} e^{-t^2/2}} \\
&= \lim_{t\to\infty} \frac{[1-\Phi(t)]}{\frac{1}{\sqrt{2\pi}} e^{-t^2/2}} \\
&= \lim_{t\to\infty} \frac{-\frac{1}{\sqrt{2\pi}} e^{-t^2/2}}{-\frac{1}{\sqrt{2\pi}} t e^{-t^2/2}} \\
&= \lim_{t\to\infty} \frac{1}{t} = 0.
\end{aligned}$$

B Claim III.D.1

> Testing reduces the majority/minority hiring gap (unbiased case with minority applicants less productive than majority applicants).

Assume without loss of generality that x_1 is the more productive group, and let $\gamma_1 = E_{\eta,s}[Hire|x_1] - E_\eta[Hire|x_1]$ and $\gamma_2 = E_{\eta,s}[Hire|x_2] - E_\eta[Hire|x_2]$. A constant hiring rate implies that testing either leaves hiring of both groups unaffected or moves the hiring rate of each group by equal but opposite amounts (either $\gamma_1, \gamma_2 = 0$ or $\gamma_1 = -\gamma_2 \neq 0$). We note that $\Delta\gamma$ can be expressed as $\Delta\gamma = \gamma_1 - \gamma_2$.

The introduction of testing without a change in bias is identically equal to a rise in screening precision. We can therefore sign $\Delta\gamma$ by differentiating γ_1 and γ_2 with respect to ρ_T, bearing in mind that the screening threshold, κ_T, also depends upon screening precision. This derivative is:

$$\begin{aligned}
\gamma_x &= \frac{\partial}{\partial \rho_T} \left[1 - \Phi\left(z_T(x)\right)\right] \\
&= \phi\left(z_T(x)\right) \left[\frac{z_T(x)}{\rho_T} - \frac{\partial \kappa_T / \partial \rho_T}{\rho_T \sigma_0}\right],
\end{aligned}$$

where $z_T(x) \equiv [\kappa_T - \mu_0(x)]/\sigma_0\rho_T$. Since $\phi(\cdot) > 0$ and $z_T(x_1) < z_T(x_2)$, it cannot be the case that γ_1 and γ_2 are simultaneously equal to zero. Therefore $\text{Sign}\langle z_T(x_1) - [\partial\kappa_T/\partial\rho_T]/\sigma_0\rangle = -\text{ Sign } \langle z_T(x_2) - [\partial\kappa_T/\partial\rho_T]/\sigma_0\rangle$. Given that $z_T(x_1) < z_T(x_2)$, we conclude that $\gamma_1 < 0$, $\gamma_2 > 0$ and $\Delta\gamma = \gamma_1 - \gamma_2 < 0$. Testing therefore raises minority hiring or, more generally, raises hiring of the group with lower average productivity.

C Claim III.D.2

The effect of testing on the productivity gap approaches zero as the proportion of applicants hired approaches zero (unbiased case).

We can write

$$\Delta\pi = \sigma_0\rho_T[\lambda(z_T(x_1)) - \lambda(z_T(x_2))] - \sigma_0\rho_I[\lambda(z_I(x_1)) - \lambda(z_I(x_2))].$$

Define

$$\begin{aligned} \alpha_T &= z_T(x_1) - z_T(x_2) = -\frac{\Delta\mu_0}{\sigma_0\rho_T}, \\ \alpha_I &= z_I(x_1) - z_I(x_2) = -\frac{\Delta\mu_0}{\sigma_0\rho_I}. \end{aligned}$$

Now consider taking limits as $K \to 0$; as $K \to 0$, $z_T(x_2) \to \infty$ and $z_I(x_2) \to \infty$ while α_T and α_I remain fixed constants. Recall that $\lim_{z\to\infty} \lambda(z) - z = 0$, so that

$$\begin{aligned} \lim_{z_j(x_2)\to\infty} [\lambda(z_j(x_2) + \alpha_j) - (z_j(x_2) + \alpha_j)] - [\lambda(z_j(x_2)) - z_j(x_2)] &= 0 \\ \lim_{z_j(x_2)\to\infty} [\lambda(z_j(x_2) + \alpha_j) - \lambda(z_j(x_2))] &= \alpha_j, \end{aligned}$$

for $j \in \{T, I\}$. It follows that

$$\begin{aligned} \lim_{K\to 0} \Delta\pi &= \sigma_0\rho_T\alpha_T - \sigma_0\rho_I\alpha_I \\ &= \Delta\mu_0 - \Delta\mu_0 \\ &= 0. \end{aligned}$$

IX References

Aberdeen Group, *Hourly Hiring Management Systems: Improving the Bottom Line for Hourly Worker-Centric Enterprises*, (Boston, MA: Aberdeen Group 2001).

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

Altonji, Joseph G. and Rebecca M Blank, "Race and Gender in the Labor Market," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3C. (Amsterdam: North-Holland, 1999).

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

Angrist, Joshua D, "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 Discussion Paper 1017-93, 1993.

Autor, David H. and David Scarborough, "Will Job Testing Harm Minority Workers?" MIT Department of Economics Working Paper No. 04-29, 2004.

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

Bertrand, Marianne and Sendhil Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, XCIV (2004), 991-1013.

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

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

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

Costrell, Robert M. and Glenn C. Loury, "Distribution of Ability and Earnings in a Hierarchical Job Assignment Model., *Journal of Political Economy*, CXII (2004), 1322-1363.

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

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

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

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

Goldin, Claudia and Cecilia Rouse, "Orchestrating Impartiality: The Impact of Blind Auditions on the Sex Composition of Orchestras," *American Economic Review*, XC (2000), 715-41.

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

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

Holzer, Harry J., Steven Raphael and Michael A. Stoll, "Perceived Criminality, Criminal Background Checks, and the Racial Hiring Practices of Employers," *Journal of Law and Economics*, XLIX (2006), 451-480.

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

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

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

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

Lundberg, Shelly J. and Richard Startz, "Private Discrimination and Social Intervention in Competitive Labor Markets," *American Economic Review*, LXXIII (1983), 340-347.

Masters, Adrian, "Matching with Interviews," Mimeograph, State University of New York at Albany, 2006.

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

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

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

Stiglitz, Joseph, "The Theory of "Screening," Education, and the Distribution of Income," *American Economic Review*, LXV (1976), 283-300.

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

U.S. Census Bureau, *Census 2000 Summary File 1: Census of Population and Housing*, (Wash-

ington, DC: U.S. Census Bureau, 2001).

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

U.S. Department of Labor, Equal Employment Opportunity Commission, "Uniform Guidelines on Employee Selection Procedures," Title 41 Code of Federal Regulations, Pt. 60-3, 1978.

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

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

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

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

A. Frequencies

	Full Sample		Non-Tested Hires		Tested Hires	
	Frequency	% of Total	Frequency	% of Total	Frequency	% of Total
All	33,924	100%	25,561	75%	8,363	25%
White	23,560	69.5	18,057	70.6	5,503	65.8
Black	6,262	18.5	4,591	18.0	1,671	20.0
Hispanic	4,102	12.1	2,913	11.4	1,189	14.2
Male	17,444	51.4	13,008	50.9	4,436	53.0
Female	16,480	48.6	12,553	49.1	3,927	47.0

B. Employment Spell Duration (days)

	Full Sample		Non-Tested Hires		Tested Hires	
	Median	Mean	Median	Mean	Median	Mean
All	99	173.7	96	173.3	107	174.8
	[97, 100]	(1.9)	[94, 98]	(2.1)	[104, 111]	(2.9)
White	106	184.0	102	183.0	115	187.1
	[103, 108]	(2.1)	[100, 105]	(2.3)	[112, 119]	(3.6)
Black	77	140.1	74	138.1	87	145.7
	[75, 80]	(3.0)	[71, 77.4]	(3.5)	[81.9, 92]	(4.8)
Hispanic	98	166.4	98	169.3	99	159.5
	[93, 103]	(4.6)	[92, 104]	(5.4)	[90, 106]	(6.4)

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.

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

A. Test Scores of Applicants (n = 189,067)

			Percent in each category		
	Mean	SD	Quartile 1: 'Red'	Quartile 2: 'Yellow'	Quartile 3 & 4: 'Green'
All	0.000	1.000	23.2	24.8	52.0
White	0.064	0.996	20.9	24.5	54.6
Black	-0.125	1.009	27.8	25.2	47.1
Hispanic	-0.056	0.982	24.9	25.6	49.6
Male	0.019	0.955	24.4	24.3	51.3
Female	-0.014	1.033	21.6	25.5	52.9

B. Test Scores of Hires (n = 16,925)

			Percent in each category		
	Mean	SD	Quartile 1: 'Red'	Quartile 2: 'Yellow'	Quartiles 3 & 4: 'Green'
All	0.707	0.772	0.18	16.1	83.8
White	0.720	0.772	0.14	15.7	84.2
Black	0.667	0.777	0.39	16.4	83.2
Hispanic	0.695	0.768	0.13	17.3	82.6
Male	0.749	0.750	0.23	14.9	84.8
Female	0.657	0.788	0.13	17.4	82.5

C. Hire Rates by Applicant Group

By Race and Gender

Race/Sex	% Hired	Obs
All	8.95	189,067
White	10.16	113,354
Black	7.17	43,314
Hispanic	7.12	32,399
Male	8.59	106,948
Female	9.42	82,119

By Test Score Decile

Decile	% Hired	Obs
1	0.07	19,473
2	0.06	20,038
3	3.96	18,803
4	5.65	18,774
5	7.97	19,126
6	10.99	18,264
7	11.71	18,814
8	13.76	18,029
9	16.14	19,491
10	20.43	18,255

Table Notes:
- N=189,067 applicants and 16,925 hires at 1,363 sites.
- Sample includes all applicants and hires between Aug 2000 and May 2001 at sites used in treatment sample.
- Test score sample is standardized with mean zero and unit variance.

Table III. The Relationship Between Applicant Characteristics and Test Scores

Dependent Variable: Standardized Test Score

	(1)	(2)	(3)	(4)	(5)
Black	-0.192	-0.183	-0.125	-0.113	-0.113
	(0.008)	(0.007)	(0.008)	(0.008)	(0.008)
Hispanic	-0.121	-0.148	-0.100	-0.093	-0.093
	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)
Male	-0.044	-0.045	-0.052	-0.053	-0.053
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Median income in applicant's zip code				0.066	0.062
				(0.015)	(0.016)
Percent non-white in applicant's zip code				-0.071	-0.071
				(0.023)	(0.023)
State effects	No	Yes	No	No	No
1,363 Site effects	No	No	Yes	Yes	Yes
State trends	No	No	No	No	Yes
R-squared	0.0070	0.0113	0.0265	0.0269	0.0277
Obs	189,067				

Table Notes:

-Robust standard errors in parentheses account for correlation between observations from the same site (1,363 sites).

-Sample includes all applications from August 2000 through May 2001 at sites in treatment sample.

-All models include controls for the year-month of application and an 'other' race dummy variable to account for 25,621 applicants with other or unidentified race.

-Income and fraction non-white for stores and applicants are calculated using store zip codes merged to 2000 Census SF1 and SF3 files.

Table IV. 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)
			OLS Estimates				2SLS Estimates			
Employment test			8.9	18.4	18.4	21.8	6.3	14.9	14.8	18.1
			(4.5)	(4.0)	(4.0)	(4.3)	(5.1)	(4.6)	(4.6)	(5.0)
Black	-43.5	-25.9			-25.9	-25.8			-25.9	-25.8
	(3.2)	(3.5)			(3.5)	(3.5)			(3.5)	(3.5)
Hispanic	-17.5	-11.8			-11.8	-11.7			-11.8	-11.7
	(4.4)	(4.1)			(4.1)	(4.1)			(4.1)	(4.1)
Male	-4.2	-2.0			-2.0	-1.9			-2.0	-1.9
	(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	0.0112	0.1089	0.0049	0.1079	0.1094	0.1116				

Table Notes:
-N=33,266
-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 V. The Relationship between Site-Level Mean Test Scores and Job Spell Duration of Hires
Dependent Variable: Length of Employment Spell (days)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Non-Tested Hires (OLS)			Tested Hires (OLS)			Tested Hires (2SLS)		
Mean test score of applicants	39.2	40.3	36.5	36.4	36.2	40.9			
	(11.9)	(12.1)	(13.6)	(18.3)	(18.3)	(19.7)			
Mean test score of hires							57.5	57.0	53.9
							(25.8)	(25.5)	(23.7)
Log median income in store zip code			-12.3			-23.7			-18.4
			(7.1)			(11.2)			(11.5)
Share non-white in store zip code			-19.4			-12.8			-21.5
			(11.2)			(16.2)			(15.3)
Black	-37.2	-36.6	-34.8	-34.2	-33.2	-33.8	-35.8	-34.9	-33.3
	(4.0)	(4.0)	(4.1)	(6.1)	(6.0)	(6.3)	(6.0)	(5.9)	(6.2)
Hispanic	-9.9	-9.7	-8.2	-23.2	-22.9	-24.1	-25.7	-25.5	-24.7
	(5.5)	(5.5)	(5.3)	(7.0)	(7.0)	(7.1)	(7.1)	(7.1)	(7.2)
Male	-5.4	-5.5	-5.3	0.0	-0.7	-0.2	0.3	-0.5	-0.3
	(2.8)	(2.8)	(2.8)	(4.8)	(4.8)	(4.9)	(4.8)	(4.8)	(4.8)
State effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State trends	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
R-squared	0.0227	0.0257	0.0260	0.0257	0.0335	0.0343			
N	25,039			8,177			8,169		

Table Notes:
-Robust standard errors in parentheses account for correlation between observations from the same site (1,363 clusters)
-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 (189,067 applicants total).
-Cross-store standard deviation of mean applicant scores and mean hire scores are 0.159 and 0.315 respectively.
In Panel C, test scores of hired are instrumented using test scores of applicants.

Table VI. OLS and IV Estimates of the Effect of Job Testing on the Job Spell Duration of Hires: Testing for Differential Impacts by Race
Dependent Variable: Length of Completed Employment Spell (days)

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS Estimates			2SLS Estimates		
White x tested	13.8	19.7	23.2	12.3	17.0	20.4
	(5.0)	(4.6)	(4.8)	(5.7)	(5.2)	(5.6)
Black x tested	15.4	22.2	23.2	12.4	18.1	18.8
	(6.4)	(5.9)	(6.0)	(7.0)	(6.7)	(6.9)
Hispanic x tested	-1.2	7.0	12.8	-5.6	0.5	6.4
	(8.8)	(7.3)	(7.6)	(9.2)	(7.7)	(8.1)
Black	-44.5	-26.5	-25.8	-44.0	-26.2	-25.4
	(3.8)	(3.9)	(3.9)	(3.9)	(3.9)	(3.9)
Hispanic	-14.0	-8.2	-8.8	-13.1	-7.2	-7.8
	(5.5)	(4.8)	(4.9)	(5.6)	(4.9)	(4.9)
Male	-4.2	-2.0	-1.9	-4.2	-2.0	-1.9
	(2.4)	(2.4)	(2.4)	(2.4)	(2.4)	(2.4)
Site effects	No	Yes	Yes	No	Yes	Yes
State trends	No	No	Yes	No	No	Yes
H_0: Race interactions jointly equal	0.19	0.15	0.36	0.14	0.08	0.21
R-squared	0.012	0.109	0.112			

Table Notes:
-N=33,266
-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 VII. Estimates of The Effect of Job Testing on Hiring Odds by Race (Panel A) and the the Share of Hires by Race (Panels B and C)
Dependent Variable:Equal to One (Zero) if Hired Worker is (not) of Specified Race

	(1)	(2)	(3)	(4)	(5)	(6)
	White		Black		Hispanic	
	Panel A. Hiring odds: 100 x Fixed Effects Logit Estimates					
Employment test (logit coefficient)	2.90	2.06	-2.35	-0.13	-2.48	-5.78
	(5.63)	(5.89)	(6.77)	(7.14)	(7.33)	(7.62)
State trends	No	Yes	No	Yes	No	Yes
N	30,921	23,957	26,982	26,982	22,453	22,453
	Panel B. Hiring Shares: 100 x OLS Estimates					
Employment test (OLS coefficient)	0.41	0.24	-0.27	-0.04	-0.14	-0.21
	(0.84)	(0.89)	(0.69)	(0.72)	(0.62)	(0.67)
State trends	No	Yes	No	Yes	No	Yes
N	33,924	33,924	33,924	33,924	33,924	33,924
	Panel C. Hiring Shares: 100 x 2SLS Estimates					
Employment test (2SLS coefficient)	0.78	0.69	-0.15	0.09	-0.63	-0.78
	(0.95)	(1.02)	(0.78)	(0.81)	(0.70)	(0.77)
State trends	No	Yes	No	Yes	No	Yes
N	33,924	33,924	33,924	33,924	33,924	33,924

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 VIII. The Relationship Between Store Zip Code Demographics and Race of Hires
Before and After Use of Applicant Testing
Dependent Variable: An Indicator Variable Equal to 100 if Worker is of Given Race

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	White				Black				Hispanic			
	Not Tested	Tested	All	All	Not Tested	Tested	All	All	Not Tested	Tested	All	All
Panel A: Race of Hires and Minority Share in Store Zip-Code												
Share non-white in zip code	-87.4 (2.3)	-86.1 (3.4)	-87.6 (2.2)		56.5 (3.5)	56.7 (4.9)	56.5 (3.3)		30.9 (3.0)	29.4 (4.4)	31.2 (2.9)	
Share non-white x tested			1.3 (3.3)	-0.3 (1.8)			1.1 (4.9)	1.3 (1.7)			-2.4 (4.5)	-1.1 (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	0.232	0.253	0.236	0.353	0.169	0.197	0.174	0.356	0.130	0.110	0.124	0.296
N	25,561	8,363	33,924	33,924	25,561	8,363	33,924	33,924	25,561	8,363	33,924	33,924
Panel B: Race of Hires and Log Median Income in Store Zip-Code												
Log median income in zip	32.0 (2.5)	39.5 (3.1)	32.2 (2.4)		-20.0 (2.5)	-23.0 (3.2)	-20.0 (2.4)		-12.1 (1.6)	-16.5 (2.5)	-12.3 (1.6)	
Log median income x tested			5.9 (3.8)	0.6 (1.6)			-3.0 (3.7)	-0.4 (1.4)			-2.8 (2.8)	-0.3 (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	0.117	0.155	0.125	0.353	0.099	0.129	0.104	0.356	0.102	0.095	0.099	0.296
N	25,561	8,363	33,924	33,924	25,561	8,363	33,924	33,924	25,561	8,363	33,924	33,924

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.

Table IX. The Impact of Job Testing on Hiring and Job Spell Durations of White and Black Applicants under Six Bias Scenarios: Comparing Simulation Results with Observed Outcomes.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Simulation Results						
Avg. ability	W > B	W > B	W > B	W = B	W = B	W = B	
Interview bias	Neutral	Favors W	Favors B	Neutral	Favors W	Favors B	
Test bias	Neutral	Neutral	Neutral	Favors W	Favors W	Favors W	Observed
			A. Productivity: Job Spell Durations in Days				
1. Initial tenure gap: W - B	52.0 (5.1)	30.1 (5.9)	80.7 (5.0)	-13.2 (4.9)	-41.9 (5.1)	15.6 (4.5)	44.9 (3.9)
2. Δ W tenure	18.6 (1.2)	20.4 (1.1)	16.8 (1.3)	16.8 (1.3)	18.6 (1.2)	16.0 (1.3)	23.2 (4.8)
3. Δ B tenure	19.9 (2.7)	19.7 (3.2)	23.1 (2.3)	23.2 (2.3)	20.0 (2.7)	27.3 (2.1)	23.2 (6.0)
4. Δ W - Δ B tenure	-1.4 (3.0)	0.7 (3.4)	-6.3 (2.7)	-6.4 (2.7)	-1.4 (3.0)	-11.3 (2.6)	0.0 (6.2)
5. χ^2(3) rows 1, 2, 3 P-value	2.4 0.50	5.1 0.17	34.0 0.00	88.1 0.00	185.5 0.00	26.6 0.00	
			B. Employment Shares and Log Odds of Hiring				
6. ΔW emp share x 100	-0.97 (0.18)	-2.38 (0.18)	0.86 (0.18)	0.86 (0.18)	-0.98 (0.19)	2.69 (0.19)	0.24 (0.89)
7. ΔB emp share x 100	0.82 (0.15)	1.72 (0.15)	-0.53 (0.16)	-0.53 (0.15)	0.82 (0.15)	-1.88 (0.16)	-0.04 (0.72)
8. ΔW - ΔB emp share x 100	-1.79 (0.31)	-4.10 (0.30)	1.39 (0.31)	1.39 (0.30)	-1.79 (0.31)	4.57 (0.32)	0.28 (1.42)
9. χ^2(2) rows 6, 7 P-value	3.4 0.33	14.9 0.00	1.0 0.79	1.0 0.79	3.4 0.33	15.0 0.00	
			C. Omnibus Goodness of Fit Statistics for Productivity and Employment				
10. χ^2(5) rows 5, 9 P-value	5.8 0.33	20.0 0.00	35.0 0.00	89.2 0.00	188.9 0.00	41.6 0.00	

Notes:
- 1,000 replications of each of six scenarios (corresponding to columns 1 through 6) using 189,067 applicant files.
- In columns 1 through 6, standard deviations of estimates from 1,000 simulations are in parentheses.
- In column 7, standard errors from empirical estimates are in parentheses.
- Point estimates and standard errors for results in column 7 are obtained from Tables I, V and VII.
- Chi-squared goodness of fit statistics are obtained by comparing simulation estimates in columns 1 through 6 to observed outcomes in column 7.

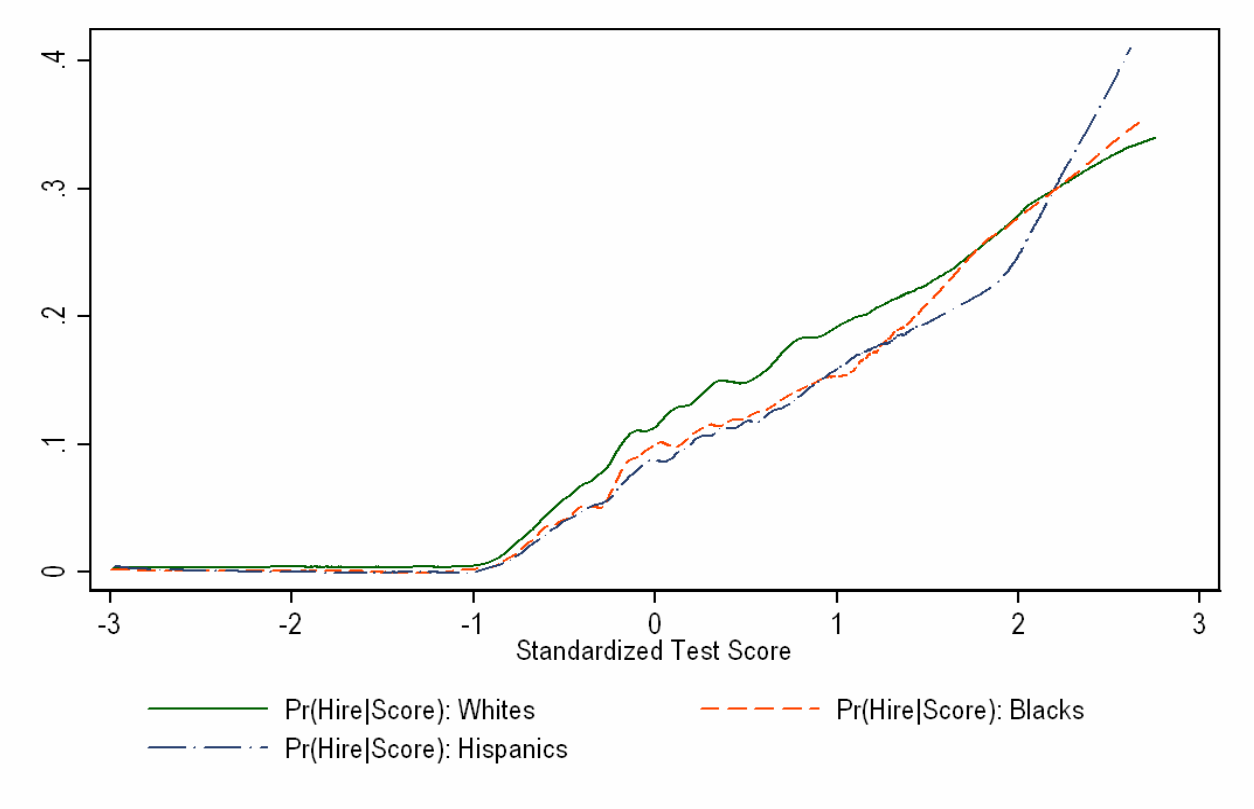
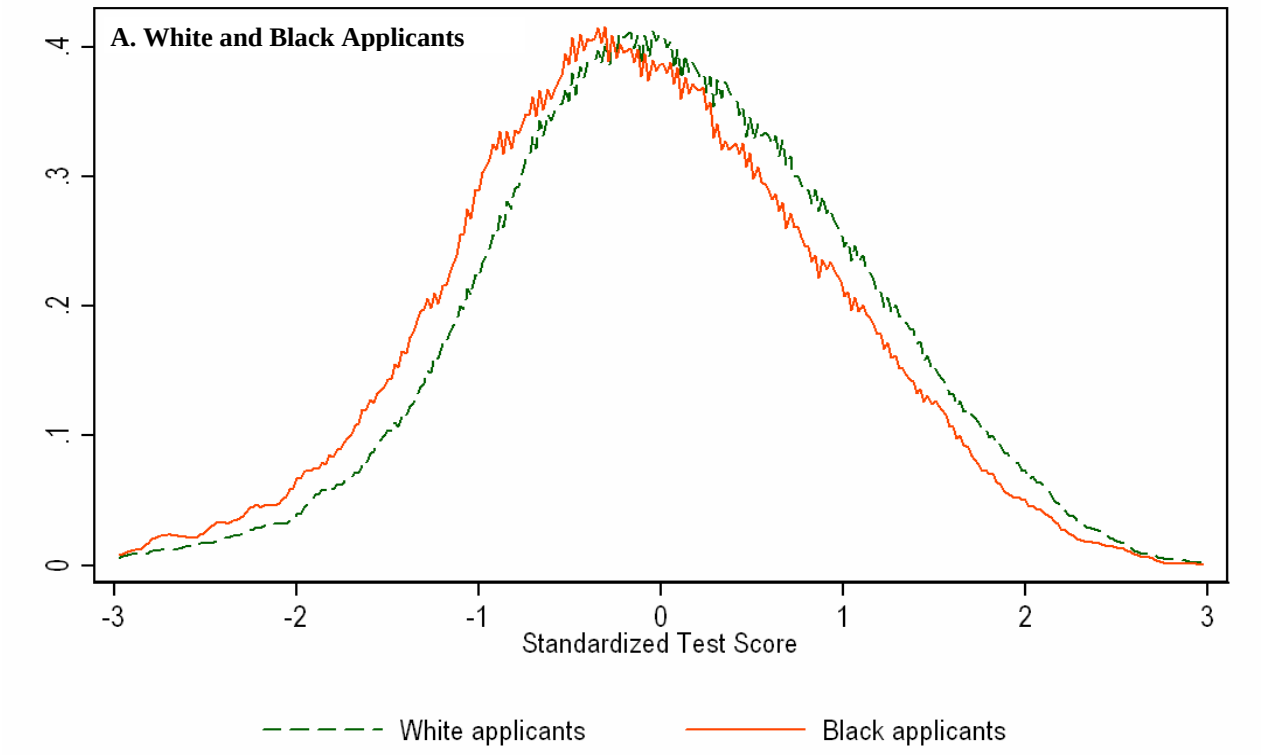


Figure I. Conditional Probability of Hire as a Function of Test Score by Race: Locally Weighted Regressions. Sample: All White, Black and Hispanic applicants, June 2000 - May 2001 (n=189,067).



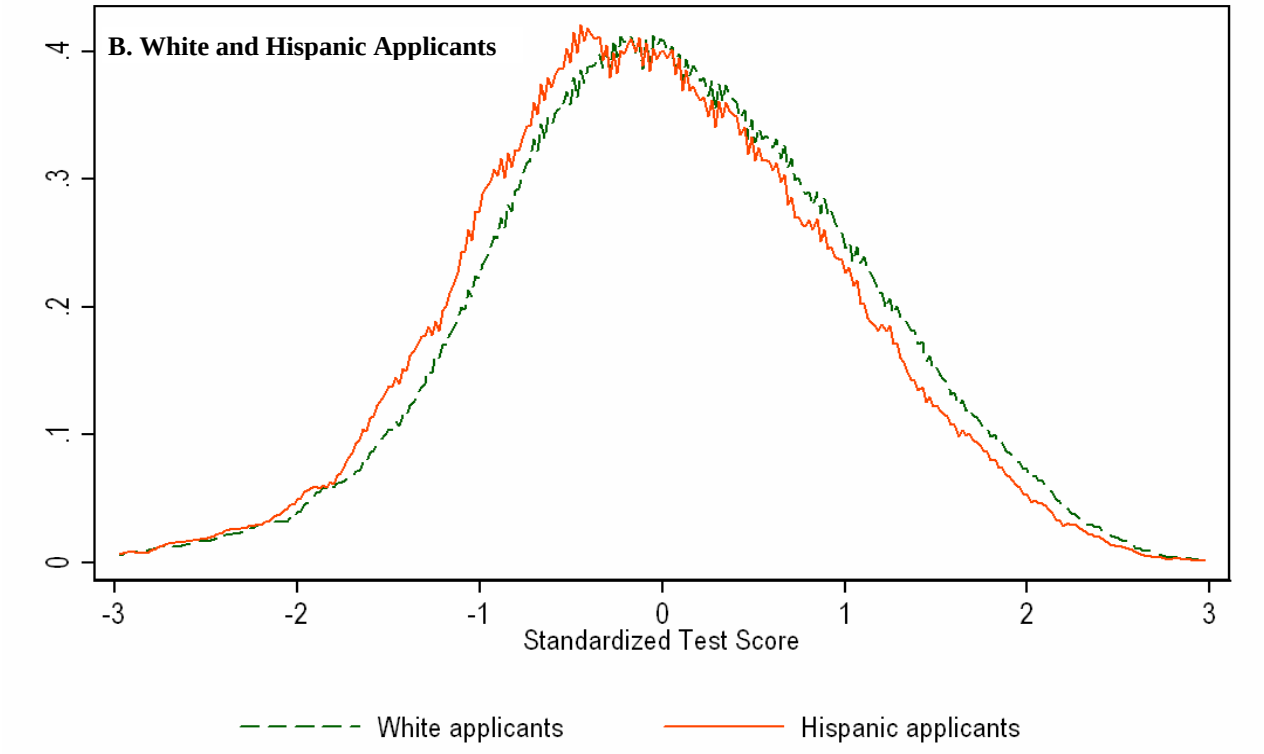


Figure II. Density of Applicant Test Scores
Sample: All White, Black and Hispanic applicants, June 2000 - May 2001 (*n* =189,067)

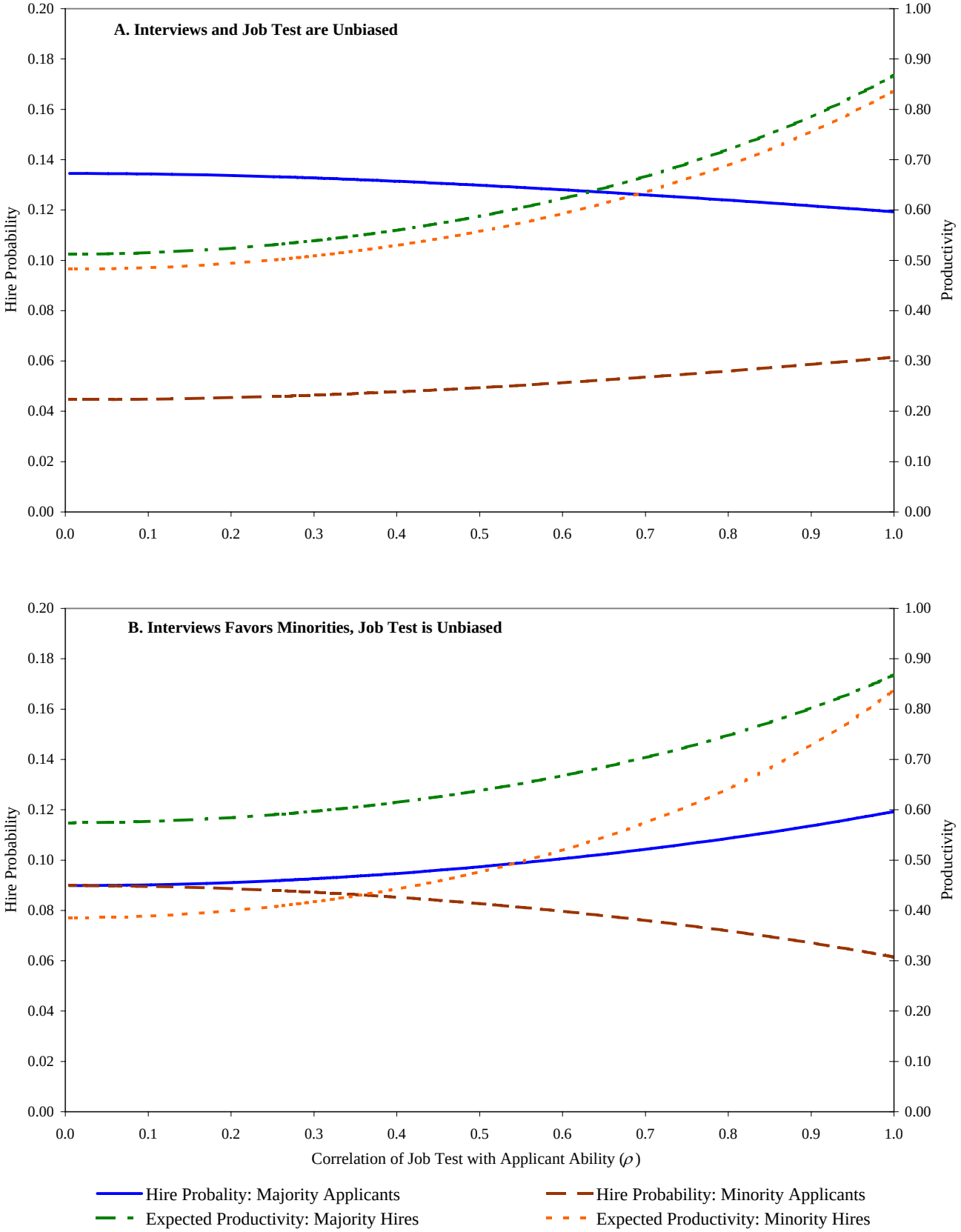


Figure III.
Simulation of the Impact of Job Testing on the Hiring and Productivity Gaps between Minority and Non-Minority Workers under Two Screening Scenarios: (A) Interviews and Job Tests are Unbiased; (B) Interviews Favor Minorities, Job Test is Unbiased.

Figure III note.

In the simulation, nine percent of applicants are hired, the productivity (equivalently ability) of majority applicants is distributed N(0,0.27), the productivity of minority applicants is distributed N(-0.19,0.27), the precision of the interview signal is 1/0.45 and the precision of the job test ranges from 1/10,000 to 1/0.0001 corresponding to a correlation between test scores and applicant ability of (0,1). These values are chosen to match estimates from the parametric simulation of the model in Section 6 of the text. In panel (A), both interviews and tests are mean-consistent with true applicant ability. In panel (B), the job test is mean-consistent with the true applicant ability and interviews are mean-biased in favor of minority applicants by +0.19.

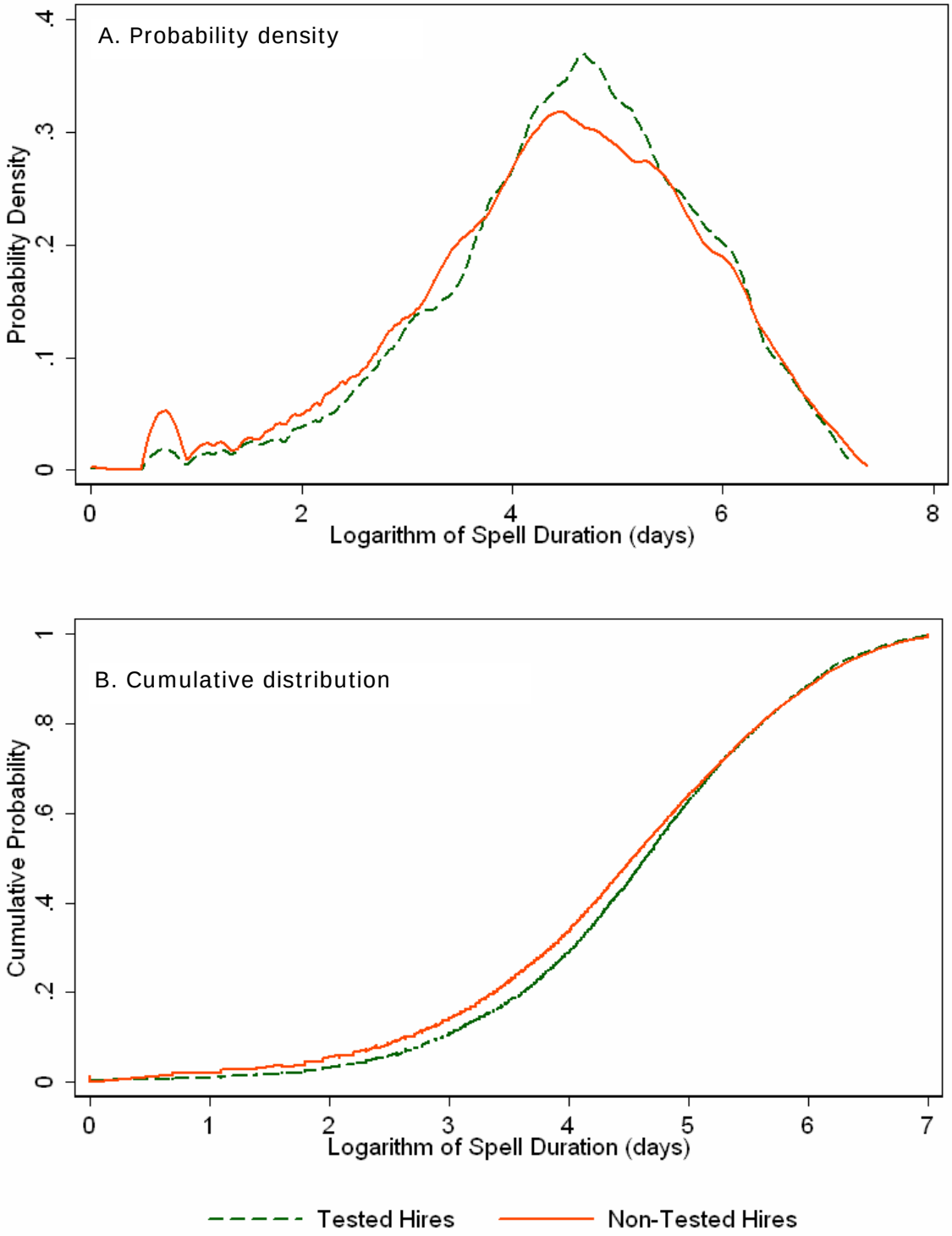


Figure IV. Completed Job Spell Durations of Tested and Non-Tested Hires.
Sample: Hires June 2000 - May 2001 with Valid Outcome Data (n = 33,266)

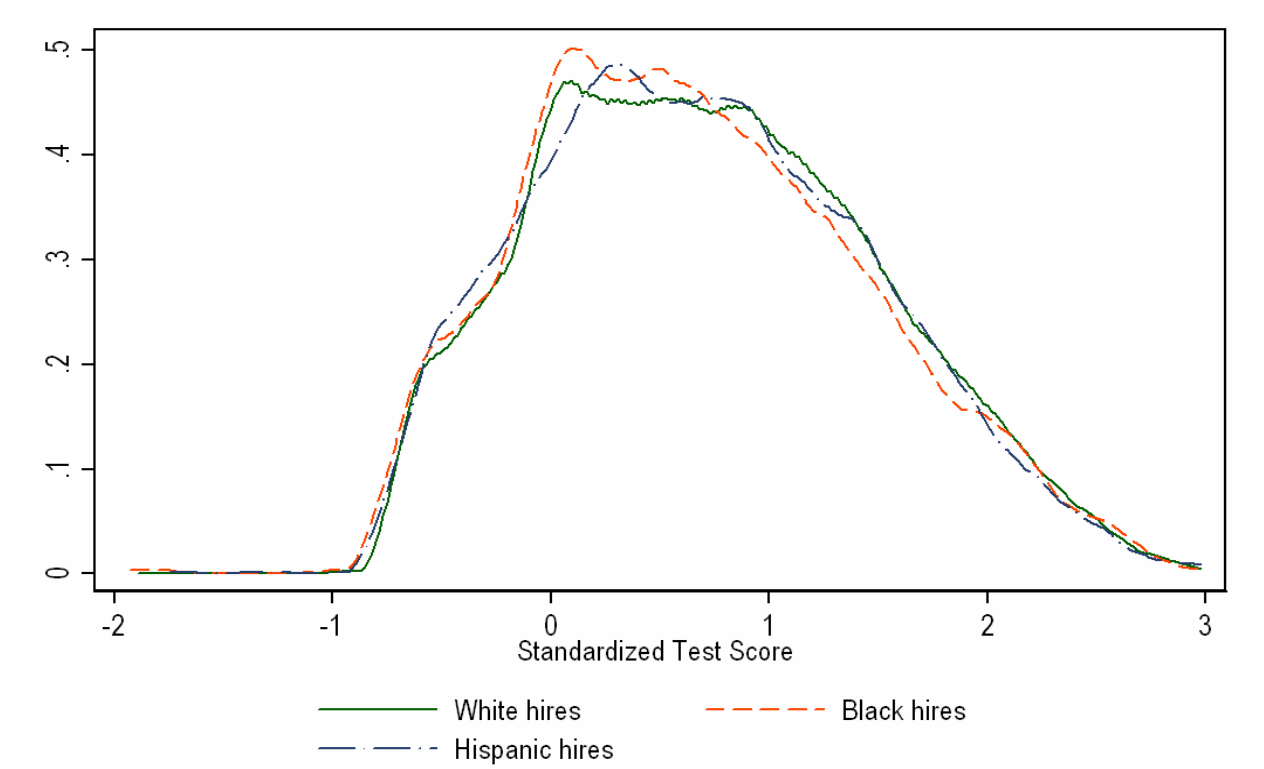


Figure V. Test Score Densities of Hired Workers by Race