Will Job Testing Harm Minority Workers? Evidence from the Retail Sector*

David H. Autor
MIT and NBER

David Scarborough
Unicru, Inc.

July 2005
Revised from December 2003

Abstract

Because minorities typically fare poorly on standardized tests, job testing is thought to pose an equality-efficiency trade-off: testing improves selection but reduces minority hiring. We evaluate this trade-off using data from a national retail firm whose 1,363 stores switched from informal to test-based worker screening. We find that testing yielded more productive hires – raising mean and median tenure by 10 percent and slighlty reducing the frequency of firing for cause. Consistent with prior research, minorities performed significantly 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 employers already accounted for expected productivity differences between minority and non-minority applicants prior to the introduction of testing – that is, they statistically discriminated. Consequently, testing raised productivity without disparate impacts on minority workers.

JEL: D63, D81, J15, J71, K31, M51
Keywords: Job testing, Discrimination, Economics of minorities and races, Worker screening, Productivity, Personnel economics

*We thank Daron Acemoglu, Joshua Angrist, David Card, John Donohue, Roland Fryer, Caroline Hoxby, Lawrence Katz, Edward Lazear, Michael Greenstone, Sendhil Mullainathan, Roberto Fernandez, and numerous seminar participants for insightful suggestions. We are indebted to Tal Gross for superb research assistance and Alan Baumbusch of Unicru, Inc. for generous assistance with all data matters. Autor gratefully acknowledges financial support from the National Science Foundation (CAREER SES-0239538) and the Alfred P. Sloan foundation.

1 Introduction

In the early 20th century, the majority of unskilled, industrial employees in the United States were hired with no systematic efforts at selection (Wilk and Cappelli, 2003). Sanford Jacoby's well-known industrial relations text describes an early 20th century Philadelphia factory at which foremen tossed apples into crowds of job-seekers, and hired the men who caught them (Jacoby, 1985, p. 17). These hiring practices are no longer commonplace. During the 1980s, as much as one-third of large employers adopted systematic skills testing for job applicants (Bureau of National Affairs, 1980 and 1988). But skills testing has remained rare in hiring for hourly wage jobs, where training investments are typically modest and employment spells brief (Aberdeen, 2001). Due to advances in information technology, these practices are now poised for change. With increasing prevalence, employers use computerized job applications and assessments to administer and score personality tests, perform online background checks and guide hiring decisions. Over time, these tools are likely to become increasingly sophisticated, as for example has occurred in the consumer credit industry.

Widespread use of job testing has the potential to raise aggregate productivity by improving the quality of matches between workers and firms. But there is a pervasive concern, reflected in public policy, that job testing may have adverse distributional consequences, commonly called 'disparate impacts.' Because of the near universal finding that minorities, less-educated and low socioeconomic-status 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?"[1]

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 governing the use of employment tests,[2] several Equal Employment Opportunity Commission guidelines regu-

[1] Nor is this expression of concern merely rhetorical. In the conclusion of their volume, 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.

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

lating employee selection procedures (U.S. Department of Labor, 1978),[3] and two National Academy of Sciences studies evaluating the efficacy and fairness of job testing (Hartigan and Wigdor, 1989; Wigdor and Green, 1991).[4]

Yet, despite a 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. As our illustrative model below demonstrates, there are two assumptions underlying the presumed trade-off, and these assumptions do not appear equally palatable. The first assumption is that employment tests hold the potential to improve employee selection. This assumption is supported by a large body of research and we view it as non-controversial.[5] The second assumption, implicit in the 'disparate impact' view, is that absent job testing, firms are essentially uninformed about the distribution of worker productivity and hence do not implicitly account for minority/non-minority differences when hiring. Under these assumptions, the introduction of an unbiased test is likely to reduce the hiring rate of minority workers – a disparate impact – because an improvement in screening precision implicitly 'reveals' minority group members to be less productive than non-minority workers on average.

Because competitive employers face a strong incentive to select and remunerate workers according to their productive value, however, a setting where employers are essentially unaware of the distribution of worker productivity may appear somewhat artificial. Indeed, economists have long recognized that employers face an incentive to use prior knowledge of the distribution of worker productivity, both overall and by race, to assess the expected productivity of individual job applicants, i.e., to statistically discriminate (Phelps, 1972; Aigner and Cain, 1977). Consider instead a case where firms screen informally for a tested attribute and testing improves the accuracy of screening. Will the resulting gain in screening precision reduce hiring from low scoring groups? We show below that if firms hold

[3] For example, 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."

[4] Donohue (1994) discusses the evolving notions of equality in U.S. employment law and contends that its objectives have moved beyond a market view of equality, in which all workers are paid their intrinsic productive value (equality of treatment), to a notion of "constructed equality" in which employers are compelled to make workers equal. The perceived trade-off between equality and efficiency is only a trade-off when viewed through the "constructed equality" lens; in the pure market context, testing is likely to increase equality of treatment.

[5] 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).

accurate ('informed') beliefs about the distribution of productivity by applicant group and use this information for screening, testing's main effect will be to raise the precision of screening *within* each applicant group rather than to shift hiring against minority applicants. In this case, job testing has the potential to raise productivity with no disparate impact on minority workers.

This discussion, and our formal model, suggest that the presumed trade-off between efficiency and equality in hiring is an empirical possibility rather than a theoretical certainty. To evaluate 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. To our knowledge, there is no prior research that performs this comparison.[6] 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 computer-supported method also places substantial weight on an electronically administered and scored personality test. We use the rollout of this technology over a twelve month period to contrast contemporaneous changes in productivity and minority hiring at establishments differing only in the date that they rolled out the employment test at their sites.

We find strong evidence that testing yielded more productive hires – increasing mean and median employee tenure by approximately 10 percent, and slightly lowering the frequency at which workers were fired for cause. Consistent with a large body of work, our analysis of applicant data reveals that minorities performed significantly worse on the employment test than non-minorities. Had managers initially been 'uninformed' about these group differences, simple calculations suggest that testing would have lowered minority hiring by approximately 10 to 15 percent. This did not occur. We find no evidence that employment testing changed the racial composition of hiring at this firm's 1,363 sites. Moreover, productivity gains were uniformly large among both minority and non-minority hires.

The combination of uniform productivity gains and a lack of adverse hiring impacts suggests that employers were aware of minority/non-minority differences even prior to the introduction of testing and that they used this information for hiring. Our findings are consistent with the hypothesis that firms effectively statistically discriminated prior to the introduction of employment testing; hence testing raised productivity with no disparate impacts on minority workers. Available precision does not allow us to reject a weaker interpretation of these facts, however, which is that firms imperfectly screened for the tested attribute prior to the use of testing and testing raised the precision of this

[6] Although a large literature evaluates the likely impacts of testing on private sector hiring, all studies that we are aware of compare anticipated or actual hiring outcomes using an employment test to a *hypothetical* case in which, absent testing, firms are uninformed about the distribution of worker productivity.

screening. In either case, our analysis clearly rejects the strong 'disparate impact' view; despite the measurable improvement in worker selection, evidenced by increased productivity, these gains came at no cost in reduced minority hiring.

Our paper is related to a broad theoretical and empirical literature on worker selection under uncertainty. Key contributions include Spence (1973), Stiglitz (1975) and Salop and Salop (1976), who analyze models of screening, signaling, and self-selection, and Phelps (1972) and Aigner and Cain (1977), who provide classic theoretical treatments of statistical discrimination. A number of recent studies assess the role of race in employer hiring decisions. Altonji and Pierret (2001) apply a dynamic learning model to test for employer statistical discrimination in a longitudinal panel of worker earnings, and find little evidence of race-based statistical discrimination.[7] Holzer, Raphael, and Stoll (2002) analyze whether employer-initiated criminal background checks affect the likelihood that employers hire Black workers and conclude that, in the absence of criminal background checks, employers statistically discriminate against Black applicants. A resume audit study conducted by Bertrand and Mullainathan (2004) finds that job applicants with 'Black-sounding' names receive significantly fewer job application call-backs than applicants with 'White-sounding' names, a result that is consistent with either taste-based or statistical discrimination. In the product market context, List (2004) presents a panoply of evidence from field experiments that dealers of sportscards engage in substantial statistical discrimination in trade with minority versus non-minority buyers.

Our analysis is also closely related to studies of ability testing used for military selection. Eitelberg et. al. (1984) provide a comprehensive history of ability testing in the U.S. military and discuss its implications for racial composition. Wigdor and Green (1991) provide the definitive validation study of the Armed Forces Qualification Test (AFQT) as a predictor of soldiers' in-field performance. Closest in spirit to our study – though answering a conceptually distinct question – is Angrist (1993), who demonstrates that successive increases in the military's AFQT qualification standard differentially reduced minority enlistment.

Our research contributes to the prior literature on testing, race and employment in three respects. First, despite substantial regulatory concern about the possible adverse consequences of job testing, 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 interesting. A final unusual feature of our research is that we look beyond the

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

hiring impacts of job testing to evaluate its consequences for the productivity of hires, both overall and by race, as measured by turnover and firing for cause. 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.

The next section describes our data and details the hiring procedures at the firm under study before and after the introduction of testing. Section 3 offers a simple conceptual model to illustrate how the possible disparate impacts of job testing depend critically on the employer's *prior* knowledge of population parameters. Sections 4 and 5 provide our empirical analysis of the consequences of testing for productivity and hiring. Section 6 concludes.

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

We analyze the application, hiring, and employment outcome data of a large, geographically dispersed service sector firm with outlets in 47 continental U.S. states.[8] Our data includes all 1,363 outlets of this firm operating during our sample period. All sites are company-owned, each employing approximately 10 to 20 workers in line positions and offering near-identical products and services. Line positions account for approximately 75 percent of total (non-headquarters) employment, and a much larger share of hiring. Line job responsibilities include checkout, inventory, stocking, and general customer assistance. These tasks are comparable at each store, and most line workers perform all of them. Line workers are primarily young, ages 18 - 30, and many hold their jobs for short durations. As is shown in the first panel of Table 1, 70 percent of hires are White, 19 percent are Black, and 12 percent are Hispanic.[9] Median duration of completed job spells of line workers is 99 days, and the corresponding mean is 174 days (panel B).

Worker screening

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

[8] This firm was selected for study by the first author among all Unicru clients because its phased rollout of job testing across company sites provided a clean quasi-experimental design. The data analyzed were provided by Unicru, Inc. under a non-disclosure agreement with MIT. Consent was not required (or requested) from the firm studied. Unicru personnel had not previously analyzed the firm's data to evaluate the effect of job testing on the racial distribution of hiring or productivity. 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.

[9] These figures pertain to the flow of hires rather than the stock. Since Whites at this firm typically have longer job spells than non-Whites, they will be relatively over-represented among the stock of workers as compared to the flow data.

some occasions, applicants were interviewed and hired at the time of application.

Commencing in June of 1999, the firm began rolling out electronic application kiosks provided by Unicru, Inc. By June of 2000, all 1,363 stores in our sample were equipped with the technology, which supplanted the paper application process. At the kiosk, applicants complete a questionnaire administered by a screen-phone or computer terminal, or in a minority of cases, by a web-based application. Like the paper application form, the electronic questionnaire gathers basic demographic information 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).[10]

Once the electronic application is completed, the data are sent to Unicru for automated processing. The results are transmitted to the store's manager, typically within a few minutes, 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 'red' applicants, and, as is shown in Table 2, fewer than 1 percent of all 'red' applicants are hired. Beyond this near-prohibition, managers retain considerable discretion. There are many more applicants than jobs, and only 9 percent of applicants

[10] An identical paper and pencil personality test could readily have been used in the pre-electronic application hiring regime. The cost of administering and scoring this paper and pencil test may have made it unattractive.

is hired. Even for those who score well above the 'red' threshold, the customer service test score has substantial predictive power for hiring. As shown in panel C of Table 2 and Figure 1, hiring rates are strongly increasing in the test score.[11] Only 1 in 18 of those scoring in the fourth decile (in the 'yellow' range) is hired relative to 1 in 5 applicants scoring in the highest decile.

Hiring and termination data

Our analysis draws on company personnel records that contain worker demographics (gender, race), date of hire, and termination date and termination reason for each worker hired during the sample frame. These data allow us to calculate length of service for employment spells in our sample, 98 percent of which are completed by the close of the sample. We code worker terminations into two groups: voluntary terminations and terminations for cause. Voluntary terminations include return to school, geographic relocation, or any non-adverse separation that is initiated by the worker. Firings for cause include incidents of theft, insubordination, unreliability, unacceptable performance or job abandonment. In addition, we utilize data on applicant's self-reported gender, race (White, Black, Hispanic), and the zip code of the store to which they applied for employment.[12] We merge these zip codes to data from the 2000 U.S. Census of Populations Summary Files 1 and 3 (U.S. Census Bureau, 2001 and 2003) to obtain information on the racial composition and median household income in each store's location.

An important feature of our analysis is that personnel (but not application) records are available for workers hired prior to implementation of the Unicru system at each store. Hence, we build a sample that includes all line workers hired from January 1999, five months prior to the first Unicru rollout, through May 2000, when all stores had gone online. After dropping observations in which applicants had incompletely reported gender or race, we were left with $33,924$ 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.[13]

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

[11] Figure 3 plots the results of locally weighted regressions of hiring rates on test scores by race, conditioning on store effects and application year × month. We also estimated linear probability models for hiring odds as a function of test score, store effects, time effects, and race and gender. We estimate that a one standard deviation (20 point) increase in the test score raises an applicant's hiring probability by 4.6 percentage points ($t = 67$). Given a baseline hiring rate of 9 percent, this is a large effect.

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

[13] We closed the sample at the point when all hires at this firm were made through the Unicru system. Because the rollout accelerated very rapidly in the final three of twelve months, the majority of hires during the rollout period 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 the majority of these positions pay the minimum wage. We therefore suspect that controlling for year and month of hire, as is done in all models, should purge much of the wage variation in the data.

Applicant test scores

To analyze test score differences in our sample, we draw on a database containing all 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). Although we would ideally analyze applications submitted during the rollout, these paper records were not retained. In section 4, we demonstrate that applicant test scores from this database are highly correlated with the productivity of workers hired at each store *before* the introduction of employment testing. This suggests that the applicant sample provides a reasonable characterization of workers applying for work during the rollout period.

As shown in Table 2, there are marked differences in the distribution of test scores among White, Black and Hispanic applicants. Kernel density comparisons of standardized raw test scores, shown in Figure 2, 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 (5.4 points and 3.5 'centiles'), accord closely with the representative test data reported by Goldberg et al. (1998).[14] As we show below, these test score gaps are also economically significant.[15]

Prior to undertaking the empirical analysis, we provide a simple conceptual model to explore the conditions under which testing is likely to generate disparate impacts on minority hiring.

3 When does job testing have disparate impacts?

How does the introduction of job testing affect the employment opportunities of minority job seekers in a competitive labor market? As discussed in the Introduction, the presumed answer to this question

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

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

is that testing reduces the labor market opportunities of members of low scoring groups. Here, we present a brief, illustrative model to explore when this presumption is likely to hold. Our conceptual framework is closely related to well known models of statistical discrimination by Phelps (1972), Aigner and Cain (1977), Lundberg and Startz (1984), Coate and Loury (1993) and Altonji and Pierret (2001). The primary contribution of our model is to show how the impact of testing on the employment opportunities and productivity (conditional on hire) of minority and non-minority workers depends critically on the employer's *prior* knowledge of population parameters. As we show below, if firms hold 'uninformed' prior beliefs about the distribution of worker productivity, the introduction of an unbiased test is likely to reduce the hiring rate of minority workers – a disparate impact – and also raise the productivity of minority relative to non-minority hires. These impacts occur because if firms are unaware of population parameters, an improvement in screening precision implicitly 'reveals' minority group members to be relatively less productive, thereby reducing their hiring. By contrast, if firms hold accurate ('informed') beliefs about the distribution of productivity by applicant group, testing's main effect is to raise the precision of screening within each applicant group rather than to shift hiring against minority applicants. In fact, in the case considered by our model, testing raises hiring of minority workers yet leaves the productivity differential between minorities and non-minorities essentially unaffected.

To develop these results, we consider three economic priors in order of increasing structure, ranging from a completely uninformed prior to the canonical case of statistical discrimination. We show how these priors give rise to distinct changes in hiring and productivity when an informative job test is introduced. When we bring the model to the data, we use these predictions to interpret the findings in light of what they imply for firms' screening practices prior to the introduction of job testing.

3.1 The environment

Consider a large set of firms facing job applications from two identifiable demographic groups $x \in \{a, b\}$, each comprising half of the applicant population. Firms employ one worker at a time and search for a replacement when a vacancy opens. Firms have a linear, constant returns to scale production technology, a positive discount rate, and are risk neutral. Workers produce output, $f(\eta_i) = \eta_i$, in flow terms, where η_i is the measure of worker quality. Job spell durations are independent of η and wages (also expressed in flow terms) are fixed at $\omega < \bar{\eta}_a, \bar{\eta}_b$.[16] While holding a vacancy, firms receive applications drawn at random from the pooled distribution of a and b workers who are of heterogeneous productivity. Firms choose either to hire the current applicant or to wait a non-zero interval for a new applicant. In this case, the prior applicant becomes unavailable.

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

Given fixed wages, firms strictly prefer to employ workers with higher η. Since holding a vacancy forfeits potential profits, firms apply a selection policy that trades off the costs and benefits of waiting for a superior applicant. As is well understood, this trade-off leads to a threshold rule where firms hire applicants whose expected productivity exceeds an optimally chosen value, and a constant fraction of worker-firm matches leads to hire. We analyze a reduced form version of this setup. Firms in our model select applicants using a hiring threshold, and this produces a constant hire rate of K. We further assume that $K < \frac{1}{2}$ to reduce the number of cases considered.[17] In a complete model, this hiring threshold would depend on technology and labor market conditions. In our reduced form model, the unconditional hiring probability is held constant at $\Pr(H) = K$. This simplification focuses our analysis on the first-order impacts of job testing on the distribution of hiring across applicant types, leaving total employment fixed.

A key simplifying assumption of our model is that the two demographic groups, a and b, differ only in mean productivity. Specifically, we assume that $\eta \sim N\left(\bar{\eta}_x, \sigma_\eta^2\right)$ with $x \in \{a, b\}$, $\bar{\eta}_a > \bar{\eta}_b$, and $\sigma_\eta^2 > 0$. Our assumption stands in contrast to several other well known models of statistical discrimination in which testing is differentially informative (or uninformative) for minority groups due to their higher underlying productivity variance, e.g., Aigner and Cain (1977), Lundberg and Startz (1984), and Masters (2004). 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 2 and Table 2, mean test scores in our sample also differ significantly among White, Black and Hispanic applicant groups but the variance of test scores is near identical for all three groups. While this does not prove that the underlying variance of ability is equivalent for all race groups, it is consistent with this possibility.

3.2 Screening with no prior information

Consider initially an extreme case where firms hold essentially no information about applicant productivity, either at the individual or aggregate level. Specifically, we assume that firms hold the prior that applicant productivity is distributed normally with $u_0 \sim N(\bar{\eta}, 1/h_0)$ where $h_0 = 1/\sigma_0^2$ is the 'precision' of these beliefs and $\bar{\eta}$ is the pooled mean of the applicant distribution. We consider the limiting case of $h_0 \to 0$, implying that u_0 is almost uninformative about applicant productivity. Accordingly, prior to the introduction of testing, firms effectively 'screen' workers by hiring at random; specifically, each applicant is hired with probability K. Consequently, a and b applicants face equal probability of hire,

[17] As above, fewer than 1 in 10 applicants at the stores in our sample are hired.

a representative subset of all applicants is hired, and the expected productivity gap between a and b hires is equal to the difference in population mean productivity: $\bar{\eta}_a - \bar{\eta}_b$.

Now, consider the introduction of an informative job test administered to applicants in this hiring environment. We model the job test as providing an unbiased productivity signal for each applicant, η_1 that is centered on the applicant's true productivity. Specifically, $\eta_1 \sim N\left(\eta, \sigma_1^2\right)$ with $\sigma_1^2 > 0$. Upon observing this test score, a firm's posterior distribution of an applicant's productivity is (suppressing individual subscripts i)

$$u_1 = \frac{h_0 u_0 + \eta_1/\sigma_1^2}{h_0 + 1/\sigma_1^2} \simeq \eta_1, \text{ with precision } h_1 = h_0 + \sigma_1^2 \simeq 1/\sigma_1^2, \tag{1}$$

(see DeGroot, 1986, Chapter 7.6). Because the firm's prior is almost uninformative, its best estimate of applicant productivity *is* the test score.

Introduction of job testing in this setting has three effects. First, because b's are overrepresented among low-scoring applicants and a's overrepresented among high-scoring applicants, it is immediately apparent that job testing reduces the hiring of b relative to a applicants (a 'disparate impact') and raises the productivity of b relative to a hires.

Formally, holding the overall hiring rate at K, the firm selects a hiring threshold κ_1 that solves the following condition:

$$\Pr\left[h_1 > \kappa_1\right] = K.$$

For any bounded value of κ_1, $\Pr\left[h_1 > \kappa_1 | x = a\right] > \Pr\left[h_1 > \kappa_1 | x = b\right]$.[18] Consequently, testing causes b hiring to fall and a hiring to rise. Aggregate productivity of the pool of hired workers also rises since $E\left[\eta | h_1 > \kappa_1\right] > E\left[\eta\right]$. More subtly, the change in overall selectivity leads to a differential rise in the productivity of b relative to a hires. This is shown formally in the appendix but intuition is immediate. By differentially truncating the lower end of the productivity distribution of b applicants, testing compresses – but does not eliminate – the productivity gap between a and b hires.

The first panel of Figure 3 illustrates these results by simulating the impact of job testing on the hiring and productivity gaps between a vs. b applicants under this 'uninformed' hiring scenario.[19] The introduction of testing generates a discontinuous rise in the relative hiring of a's and the relative productivity of b's. Logically, the more precise is the test (h_1 larger), the larger is the disparate impact on hiring.

[18] Although the firm's prior is *almost* uninformative, we assume that it is feasible for the firm to select κ given its (imprecise) knowledge of the distribution.

[19] Parameter values used are: $\sigma_\eta = 1$, $\bar{\eta}_a = 0.50$, $\bar{\eta}_b = 0.25$ and $K = 0.20$. The figure plots the change in $a - b$ hiring and productive gaps for $\sigma_1^2 \in [0, 3]$.

3.3 A more realistic case: Multiple screening measures

While the above case rationalizes the concern that job testing may have disparate impact on low-scoring applicant groups, one may question its relevance. Specifically, it appears implausible that firms would have *no* independent assessment of applicants' productivity in the absence of testing. Consider instead a case where prior to the introduction of testing, firms assess applicant productivity via a job interview that, like the job test, provides an unbiased productivity measure for each applicant, $\eta_2 \sim N\left(\eta, \sigma_2^2\right)$. For simplicity we model the interview and the job test as having independent measurement error, though this is not essential to the results. We continue to assume that firms have a diffuse prior about the productivity of each applicant before administering the interview or job test.

Upon administering the interview and job test, the firm's posterior assessment of applicant productivity becomes:

$$u_2 = \frac{h_0 u_0 + \eta_1/\sigma_1^2 + \eta_2/\sigma_2^2}{h_0 + 1/\sigma_1^2 + 1/\sigma_2^2} \simeq \frac{\eta_1/\sigma_1^2 + \eta_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}, \tag{2}$$

with precision

$$h_2 = h_0 + 1/\sigma_1^2 + 1/\sigma_2^2 \simeq 1/\sigma_1^2 + 1/\sigma_2^2.$$

A comparison of u_1 and u_2 indicates that the combination of interviews and job tests is more precise than either tool used alone. In fact, the introduction of testing in the interview-only environment is identically equivalent to a rise in the precision of the interview or job test (i.e., a reduction in σ_2^2 or σ_1^2).

How does the rise in precision affect the hiring and productivity of b relative to a applicants? Interestingly, the same results apply as in the previous, and arguably less plausible, case above. A proof of these results is given in the appendix. We provide intuition here. Consider again the firm's screening threshold. Let κ_2 be the selection thresholds that solve:

$$\Pr\left[u_2 > \kappa_2\right] = K$$

Recalling our assumption that less than half of all applicants from each applicant group is hired, it is easy to show that $\kappa_2 < \kappa_1$, that is the hiring threshold must fall as precision rises (reducing the dispersion of the firm's posterior) to keep overall hiring constant.

Three results follow. First, an increase in the precision of screening raises the hiring odds for high ability applicants and reduces the hiring odds for low ability applicants. The average productivity of hires, both overall and from each applicant group, rises. Second, the hiring rate of b applicants falls. To see why this occurs, observe that a and b applicants with the same posterior value (u_2) do not have identical expected productivity; rather, since screening is noisy and a and b applications are drawn from different distributions (with $\bar{\eta}_a > \bar{\eta}_b$), the expected productivity of a applicants exceeds

that of b applicants at any given posterior u_2.[20] Implicitly, noisy screening generates a bias towards b applicants. By mitigating this screening bias, testing reduces b hiring.

More formally, we can write the correlation between the firm's posterior and revealed applicant productivity ($\text{Corr}(u_2, \eta)$) as:

$$\rho = \left(\frac{\sigma_1^2 \sigma_\eta^2 + \sigma_2^2 \sigma_\eta^2}{\sigma_1^2 \sigma_2^2 + \sigma_2^2 \sigma_\eta^2 + \sigma_1^2 \sigma_\eta^2} \right)^{1/2} = \frac{\sigma_\eta}{\left(\sigma_\eta^2 + 1/h_2\right)^{1/2}}$$

with $\rho \in [0, 1]$. As precision increases (h_2 grows), $\rho \to 1$. The odds that an individual applicant is from the b versus a group at any given posterior productivity assessment, u, is:

$$L = \frac{\phi\left(\rho\left(u_2 - \bar{\eta}_b\right)/\sigma_\eta\right)}{\phi\left(\rho\left(u_2 - \bar{\eta}_a\right)/\sigma_\eta\right)}, \tag{3}$$

with $L \leq 1$ for $u_2 > \bar{\eta}_a, \bar{\eta}_b$. Holding κ constant, a rise in screening precision (a rise in h_2, implying a rise in ρ), reduces L at every point u_2 for $u_2 > \bar{\eta}_a, \bar{\eta}_b$. Consequently, the likelihood that an 'acceptable' applicant ($u_2 \geq \kappa_2$) is from the b group falls. Intuitively, if precision is low, the posterior u is nearly uninformative, and so a given value of u is almost equally likely to be drawn for either an a or b applicant. As precision rises, high values of u are more likely to be drawn for a applicants and low values for b applicants. Hence, holding κ constant, a rise in precision lowers the hiring rate of b applicants.[21]

By a similar argument, a rise in screening precision raises the productivity of b hires relative to a hires. As noted above, the expected productivity gap between a and b hires at any given u decreases as precision rises; in fact, with perfect screening, this difference falls to zero. Holding κ constant, it is immediate that the productivity gap between b and a hires must therefore fall. Again, however, κ drops and so this conclusion could be overturned if the hiring rate of b's rose relative to a's. But we have just established that the hiring rate of b's falls while that of a's rises. Hence, increased screening precision raises relative b productivity.

In short, the introduction of testing again generates a disparate negative impact on b hiring – comparable to the no-prior-information case above – and raises the productivity of b relative to a hires. These effects, illustrated in panel B of Figure 3, occur despite the fact that the firm's pre-existing screening mechanism, the job interview, is informative and mean unbiased.

[20] This 'bias' is a direct consequence of the fact that firms do not have an 'informed' prior about the distribution of worker productivity by applicant group. Given the information available to them under the 'uninformed' prior, firms are forming their posteriors (u_2) optimally.

[21] As noted above, the screening threshold must drop to maintain a constant hiring rate, which offsets this result slightly. To see that this factor cannot be fully offsetting, note that because L falls everywhere with a rise in precision (for κ exceeding $\bar{\eta}_a, \bar{\eta}_b$), it is not possible for the hiring rate of b's to remain constant without an *increase* in the hiring rate of a's. This would in turn imply an overall rise in the hiring rate, which violates the constant hiring rate assumption.

3.4 Screening with an 'informed' prior

What the previous cases have in common is that firms do not use any prior information about the *distribution* of applicant ability, either overall or by applicant group, to assess the productivity of individual applicants – this, despite the fact that firms presumably could have learned these parameters through repeated screening and hiring. Consider instead a case where firms have learned the population parameters of the productivity distribution and use this information for screening. We refer to this knowledge as an 'informed prior.' Under this prior, firms assess applicant ability as $u' \sim N(\bar{\eta}_x, 1/h')$ with $h' = 1/\sigma_\eta^2$ before administering interviews or job tests. In contrast to previous cases, this prior differs by applicant group and hence constitutes statistical discrimination.

Upon interviewing and testing, the firm's posterior of applicant productivity becomes:

$$u_3 = \frac{u'h' + \eta_1/\sigma_1^2 + \eta_2/\sigma_2^2}{1/\sigma_\eta^2 + 1/\sigma_1^2 + 1/\sigma_2^2}, \tag{4}$$

with precision

$$h_3 = h' + 1/\sigma_1^2 + 1/\sigma_2^2.$$

This posterior is nearly identical to the cases above, with the important difference that the firm's prior knowledge of applicant group means influences its posterior. To see this more clearly, it is useful to rewrite the firm's posterior as follows:

$$u_3 = \rho^2 u_2 + \left(1 - \rho^2\right) \bar{\eta}_x, \tag{5}$$

where u_2 is the firm's posterior from the non-informed prior case (equation (2)) and ρ is the correlation coefficient defined above. Equation (5) underscores that the 'informed' posterior is simply a linear combination of the *uniformed* posterior above and the group productivity mean, where the weight placed on the mean is proportional to the measurement error in the screening process $(1 - \rho^2)$.[22] Notably, because the informed prior correctly accounts for group differences in expected productivity, it eliminates the 'group bias' in screening found in the cases above; a and b applicants with the same posterior value u_3 have the same expected productivity.

How does job testing affect hiring and productivity in this setting? Logically, as in the examples above, job testing raises the average productivity of hires overall and from each applicant group. Opposite to prior cases, however, the rise in screening precision *raises* the hiring rate of b applicants. Intuitively, this occurs because as precision rises, firms put more weight on observed applicant scores and less weight on the prior. Since the firm's prior for b applicants is less favorable than for a applicants, added precision differentially benefits b applicants.

[22] Notice that as $\rho \to 1$, $u_3 \to u_2$.

Slightly more formally, we can again write the likelihood ratio that an applicant is from the b versus a group at a given posterior u_3 (and using the fact that $u_3 = \rho^2 u_2 + \left(1-\rho^2\right)\bar{\eta}_x$) as:

$$L' = \frac{\phi\left(\left(u_2 - \bar{\eta}_b\right)/\rho\sigma_\eta\right)}{\phi\left(\left(u_2 - \bar{\eta}_a\right)/\rho\sigma_\eta\right)}.$$

Opposite to the uninformed prior case above, a rise in ρ *increases* this odds ratio at given u_2 (for $u_2 > \bar{\eta}_a, \bar{\eta}_b$) as firms place more weight on the observed data and less on the prior. Consequently, with κ held constant, job testing raises the hiring rate of b applicants. This effect is in fact partly offset by a fall in κ, which is required to maintain constant hiring with a falling posterior variance ($V\left(u_3|x\right) = \rho^2\sigma_\eta^2$). The net effect of rising precision, however, is that testing *increases* b hiring.

Finally, consider the effect of testing on the productivity differential between a and b hires. Under the informed prior – and again in contrast to previous cases – the productivity of marginal hires from each applicant group is equated at all levels of test precision (i.e., $E\left(\eta|u_3 = \kappa, x\right) = u_3$ for $x = \{a, b\}$). Consequently, any impact of testing on the gap between a and b productivity can *only* arise through changes in the conditional mean gap among inframarginal a and b hires. Under the assumed normality of a and b productivity distributions, these inframarginal differences are of second order importance, as shown in the Appendix. The net effect of testing on relative productivity is therefore zero to a first approximation.[23]

These results are illustrated in panel C of Figure 3, which plots the change in the $a - b$ hiring and productivity gaps induced by testing. In this simulation, b hiring rises with testing while the impact on the productivity gap is essentially undetectable. The simple explanation for *the lack* of adverse impacts of job testing on the hiring (or productivity) of minority workers is that firms holding an informed prior accurately account for group differences in expected productivity independently of screening precision. Consequently, job testing increases the overall efficacy of job screening from both a and b applicant pools without shifting hiring against members of lower-scoring groups.[24]

3.5 Implications

Our illustrative model contains many specific (albeit, we believe reasonable) assumptions and it would be unwise to generalize very broadly based on this analysis. In fact, a key purpose of our model is to demonstrate that, contrary to an influential strand of reasoning, job testing has no *intrinsic*

[23] To be clear, one can readily construct cases where disparate impacts occur (in either direction) by assuming large cross-group dissimilarities between the productivity distributions of inframarginal hires.

[24] The NBER working paper version of this manuscript (Autor and Scarborough, 2004) shows how these conclusions are affected if firms apply a hiring 'quota' whereby a constant share of a and b workers is hired. Briefly, a hiring quota inhibits firms from equating the productivity of marginal hires from each group. Consequently, as precision rises with testing, the relative productivity of b hires – rather than their hiring rate – rises. Our results below do not suggest that this case is relevant.

implications for the relative well-being of different worker groups, even in cases where the job test indicates large between-group differences in expected productivity.

Beyond this observation, there are three general conclusions that we believe are warranted. First, in the extreme case where firms are fully unaware of group differences in productivity – i.e, their priors are 'uninformed' – testing has the potential to generate real, disparate impacts on low scoring groups. Interestingly, these disparate hiring impacts will generally be accompanied by differential rises in productivity among hired members of low-scoring groups – an implication we test below. Second, if employers are accurately informed about group productivity characteristics, job testing appears unlikely to have disparate (negative) impacts on the hiring of minority group members since firms will already account for these expected differences when screening. Lastly, in the 'informed' case, job testing should have little or no effect on the relative productivity of minority and non-minority hires since 'informed' firms will optimally equate the marginal productivity of hires across different worker groups in both the pre- and post-testing regimes.[25]

4 Estimating the productivity consequences of job testing

We now present an empirical examination of the effects of testing on the employment and productivity of minority and non-minority workers at the approximately $1,400$ sites of the firm described above. While our model above takes it as a given that job testing improves screening, this assumption requires verification. As an initial productivity measure, we compare 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 be fired for poor performance. Notably, the firm whose data we analyze implemented job testing precisely because managers believed that turnover was too high.[26] In section (4.2), we also consider a second productivity measure: involuntary terminations.

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

$$D_{ijt} = \alpha + \delta T_i + X_i\beta + \theta_t + \varphi_j + e_{ijt}, \tag{6}$$

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 contains 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 turnover.

[25] A fourth implication, almost too obvious to mention, is that job testing raises productivity by increasing the accuracy of selection.

[26] Company managers whom we interviewed report that turnover is costly because unplanned absences disrupt staffing and customer service. By contrast, they report that direct training costs are virtually nil in these line positions.

Most specifications also include a complete set of store site effects, φ, which absorb fixed factors affecting job duration at each store. Since outcomes may be correlated among workers at a given site, we use Huber-White robust standard errors clustered on store and application method.[27]

Consistent with the bivariate comparisons in Table 1, the estimate of equation (6) in column 1 of Table 3 confirms that Black and Hispanic workers have substantially lower conditional 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 minorities and non-minorities have high turnover. Nevertheless, race differences in tenure remain significant and large. While it is possible to interpret this pattern as evidence that stores engage in 'reverse discrimination' favoring Black and Hispanic applicants, these results are also fully consistent with the use of threshold based hiring, as posited by our model. As shown in Figure 2 and Table 2, the White, Black and Hispanic test score distributions have almost identical shapes but differ significantly in means, with the White mean exceeding the Black mean by 0.19 standard deviations and the Hispanic mean by 0.12 standard deviations. If firms engage in threshold-based hiring – where applicants are hired if they exceed an expected productivity cutoff – screening should reduce but not eliminate these gaps, leading to significant minority/non-minority differences in the productivity of hires. This is consistent with what we find in Table 3.[28]

Columns 3 and 4 of Table 3 demonstrate that job testing raised job spell durations. In models excluding site effects and race dummies, we estimate that workers selected using the employment test worked 8.8 days longer than those selected without use of the employment test ($t = 1.97$). When site fixed effects are added, this point estimate rises to 18.8 days ($t = 4.6$).[29] Adding controls for worker race and gender does not change the magnitude or significance of these job-test effects. When we augment the model with state $\times$ time interactions in column 6 to account for differential employment trends by state, the main estimate rises slightly to 22.1 days, which is approximately a 13 percent gain on the pre-testing baseline. Models that include a full set of state $\times$ month-year-of-hire interactions

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

[28] Several additional pieces of evidences support the hypothesis that the threshold model roughly describes hiring at this firm. As is shown in panels A and B of Table 2, test score differentials among hired Black, White and Hispanic are substantially smaller than they are among *applicants*, yet these differences are not zero. For example, the average White applicant ranks 5.4 centiles above the average Black applicant, while the average White hire ranks 1.5 centiles above the average Black hire. These race differences among applicants and hires are highly significant. Plots of the test score distributions of hired workers in Figure 5 provide further evidence for the threshold model. Relative to the test score distribution of applicants depicted in Figure 2, the pool of hired workers almost entirely excludes the lower tail of the applicant pool. Finally, the lowest scoring White, Black and Hispanic *hires* have nearly identical test scores, yet the distribution of test scores among White hires generally lies to the right of Black and Hispanics hires. Each of these results is a natural consequence of the use of a threshold-based hiring rule.

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

(not tabulated) yield nearly identical (and highly significant) results.

These tenure gains accruing from job testing are also visible in Figure 4, which plots the density and cumulative distribution of completed job spells of tested and non-tested hires. The distribution of spells for tested hires lies noticeably to the right of that of 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 Figure 4, 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 in Table 4 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.[30] We estimate that testing increased median tenure by 8 to 9 days, which is roughly a 10 percent increase and comparable in effect size to the OLS models.

Endogeneity of testing?

Though it appears unambiguous that job testing raised job spell durations, it is conceivable that our findings could be biased if job-test status were endogenous. This endogeneity could take one of two forms. A first concern is that we observe in the data that in the 1 to 2 months following the rollout of testing at a site, 10 to 25 percent of new hires were not tested. This may have occurred for several reasons: due to the lag time between hiring and the start of employment, individuals hired shortly before the advent of testing may appear as non-tested, post-testing hires in our data; operational and training issues in the weeks following the Unicru installation in some cases caused the online system to be unavailable or unused; and managers might deliberately have circumvented testing to hire preferred candidates.[31]

To purge the possible endogeneity of tested status among hires at a store using the test, we re-estimate equation (6) using a dummy variable indicating store-test-adoption as an instrumental variable for the tested status of all applicants at the store. Since we do not know the exact installation date of the electronic application kiosk at a store, we use the date of the first observed tested hire to proxy for the rollout date. The coefficient on the store-adoption dummy in the first stage equation of 0.89 ($t = 111$) indicates that, once a store has adopted testing, the vast majority of subsequent hires are tested (see Appendix Table 1).

Instrumental variables estimates of the effect of testing on job spell durations, shown in panel B of Table 3, are quite similar to OLS estimates. In all cases, they are approximately 80 percent as large and nearly as precisely estimated. In general, we cannot reject the hypothesis that IV and OLS estimates are identical. This suggests that the potential endogeneity of tested status within stores is

[30] Median regression models retain the 2 percent of observations for which the job spell had yet to be completed by the end of the sample. Since it is not feasible to estimate a large number of store fixed effects in quantile regression models, we exclude store effects and instead include 46 state dummies.

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

not a substantial source of bias.

A second source of concern 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 the best of our understanding, the rollout order of stores was determined by geography, technical infrastructure, and internal personnel decisions. It is this last factor that is of concern. If, for example, stores adopted testing when they experienced a rise in turnover, mean reversion in the length of employment spells could cause us to overestimate the causal effect of testing on workers' job spell durations.[32]

As a check on this possibility, we augmented equation (6) for job spell duration with leads and lags of test adoption. These models, summarized in Appendix Table 2, depict the trend in job spell durations for workers hired at each store in the 9 months surrounding introduction of testing: 5 months prior to 4 months post adoption. If job spell durations rose or fell significantly prior to test adoption, the lead and lag models would make this evident. In point of fact, the lead estimates are in no case significant and, moreover, do not have consistent signs. By contrast, the lag (post-rollout) dummies show a discontinuous rise in job duration for workers hired immediately after the adoption of testing. Workers hired in the first month of testing have 14 days above average job spell duration; workers hired in subsequent months have 19 to 28 days above average duration (in all cases significant). These results confirm that our main estimates are not confounded by pre-existing trends in job spell duration.[33]

4.1 Testing for differential productivity impacts by race

Prior to the use of job testing, Hispanic and especially Black workers had substantially shorter mean job durations than Whites. As our model underscores, if employers did not already account for expected productivity differences among demographic groups, testing would be expected to differentially improve the job durations of workers hired from low-scoring demographic groups. To assess the evidence for a positive 'disparate impact' of testing on productivity, we estimate an augmented version of equation (6) where we replace the 'tested' dummy variable with a full set of interactions between 'tested' and the three race groups in our sample:

$$D_{ijt} = \alpha + \delta_w T_i \cdot \text{White}_i + \delta_b T_i \cdot \text{Black}_i + \delta_h T_i \cdot \text{Hisp}_i + X_i\beta + \theta_t + \varphi_j + e_{ijt}. \quad (7)$$

[32] Managers 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 installation.

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

The parameters of interest in this equation, δ_b, δ_h and δ_w estimate the differential gains in job spell duration for cohorts of tested Black, Hispanic and White hires relative to their non-tested counterparts. Given the test score data summarized in Figure 2, the 'disparate impact' hypothesis predicts that $\hat{\delta}_b > \hat{\delta}_h > \hat{\delta}_w > 0$, that is, tenure gains should be largest among groups with the lowest scores. By contrast, if firms fully accounted for expected between-group differences prior to the use of testing, our model suggests that productivity gains should be roughly equal among demographic groups.

Table 5 presents OLS and IV estimates of equation (7). In the baseline specification, which excludes site effects and state trends (panel A, column 1), 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 panel B, we present instrumental variables 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 the final IV specification, the estimated tenure gain is 20 days for Whites, 19 days for Blacks and 6 days for Hispanics.

Taken together, these findings do not suggest that testing had a differential impact on productivity by worker group. Nor is this primarily a matter of precision. In the case of Black versus White productivity gains, the effect sizes are numerically comparable for both groups in all models.[34] One puzzling result, however, is that tenure gains are considerably smaller for Hispanic hires than other groups, and in some specifications are close to zero – a result that is not predicted by the disparate impact hypothesis or its alternatives. Hispanics are the smallest race group in our data, comprising 12 percent of hires, and the Hispanic-White test score gap is about half the size of the Black-White test score gap. We do not have a clear understanding of why the tenure gains are smaller for Hispanics. One possible explanation is that the test was initially offered only in English, which may have reduced its predictive validity for non-native Hispanic applicants.

A limitation of job duration as a measure of productivity is that it is conceivable that the most productive employees may not have the longest job spells. For example, college students who return to school after summer employment may be more capable or reliable than average workers and yet have shorter stints of employment. As one check on this possibility, we reestimated all models in Tables 3 and 5 excluding workers hired in May, June, November and December, i.e., the cohorts most likely to

[34] Though not the subject of this paper, tenure gains by gender are also similarly large: 22 days for males and 25 for females. For details, see Table 6 of Autor and Scarborough (2004).

include seasonal hires. Estimates that exclude these potential seasonal hires are closely comparable to our main estimates (results available from the authors).

4.2 A second productivity measure: Involuntary terminations

To buttress the conclusions from job duration analysis, it would be valuable to have a more direct measure of worker productivity. We explore one such measure here: firing for cause. Using linked personnel records, we distinguish for-cause terminations (e.g., theft, job abandonment, insubordination) from neutral or positive terminations (e.g., return to school, relocation, new employment). For brevity, we refer to these categories as 'involuntary' and 'voluntary' terminations, but we stress that all terminations for adverse reasons are coded as involuntary, even if initiated by the employee.

As shown in Table 1, close to half (47 percent) of all job spells have ended within 90 days of hire. To evaluate the effect of job testing on these terminations, we estimate the following linear probability model:

$$E\left[1\left\{\text{Term}_{ijt}^{90}=k\right\}\right]=\alpha+\gamma^{k}T_{i}+X_{i}\beta^{k}+\theta_{t}^{k}+\varphi_{j}^{k}, \tag{8}$$

where $1\{\cdot\}$ is the indicator function and k corresponds to each of the three potential termination statuses. The coefficient of interest, γ^k, estimates the conditional mean difference in the probability of each outcome (k) for tested relative to non-tested hires. We estimate this equation using ordinary least squares, applying robust standard-errors that are clustered on site and testing method.[35] So that coefficients may be read as percentage points, the dependent variable is multiplied by 100.

Within 90 days of hire, 30 percent of workers have ended their employment spells voluntarily and 16 percent have been terminated for cause. As shown in Table 6, job testing reduces the frequency of both types of terminations. At 90 days following hire, tested workers are 4.6 percentage points (14 percent) less likely than non-tested hires to have been terminated voluntarily and 1.6 percentage points (10 percent) less likely to have been terminated involuntarily. Both effects are statistically significant, the first highly significant and the second significant at $p = 0.06$. Adding state specific trends to these models (column 3) raises both point estimates and increases their statistical significance. Instrumental variables estimates (panel B) show slightly smaller effects, and these are less precisely estimated.[36]

[35] It would also be desirable to fit this model using a multinomial choice estimator but the very large number of included fixed effects makes this impractical. We have, however, fit multinomial logit versions of these models that exclude site effects. Like the OLS models, these estimates show that testing reduced voluntary and involuntary terminations among White and Black but not Hispanic hires. These point estimates are quantitatively similar to the OLS estimates. However, the effects of testing on involuntary terminations are not significant. Results are available from the authors.

[36] We have also estimated the effect of job testing on terminations at 60 to 360 days following hire (results available from the authors). The substantive results are closely comparable in all models, while the point estimates differ according to the time interval – since, in the long run, all workers are terminated, whether tested or not. After approximately 270 days following hire, we find no significant effect of job testing on termination rates.

To test for 'disparate impacts' on this second measure of productivity, we add to equation (8) a full set of interactions between the race category variables and the job testing dummy. Table 7 shows that job testing reduced voluntary turnover by 4 to 6 percentage points among all demographic groups (significant in all cases) and reduced involuntary turnover among both Whites and Blacks (but not Hispanics) by 1.7 to 2.8 percentage points (both significant at $p \leq 0.10$). The bottom of each column in Table 7 provides the p-value for the null hypothesis that the impact of job testing on voluntary and involuntary terminations is equivalent for all three demographic groups. In all cases, this null is accepted. In net, consistent with the results for job tenure, these models show that job testing improved a second dimension of productivity – firing for cause – without yielding differential productivity impacts for minority relative to non-minority hires.

One unexpected finding in these models is that the primary effect of job testing was to reduce voluntary rather than involuntary terminations, an impression also suggested by the raw means in Table 1.[37] One speculative explanation for this finding is that voluntary and involuntary terminations may themselves be competing risks – workers who do not quit on their own may eventually be fired for cause. If so, an increase in job spell durations coupled with a reduction in voluntary turnover may ultimately increase the odds that workers are terminated involuntarily. As is well known, it is not feasible to empirically distinguish competing risk models from alternatives, such as unobserved heterogeneity, without imposing strong *a priori* restrictions on the joint distribution of underlying hazards (cf. Honoré and Lleras-Muney, 2004). We therefore do not pursue this angle further.

4.3 The link between test scores and job performance

In light of our theoretical framework, the finding that job testing generated relatively uniform productivity gains among minority and non-minority hires is consistent with the hypothesis that firms held accurate ('informed') beliefs about the distribution of productivity by applicant group prior to the advent of testing and that they used this information for screening. If this hypothesis is correct, we would not expect testing to adversely effect minority hiring, an implication we take up in the next section. Before doing so, we briefly consider two questions that bear on the above results: 1) is there direct evidence that test scores predict worker productivity?; and if so, 2) how much of a differential productivity gain by race group could we plausibly have expected if firms were initially 'uninformed' about group differences?

To answer both questions, 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

[37] The raw comparisons also suggest that involuntary turnover rose after the introduction of testing, though this pattern is not confirmed once we condition on site effects (Tables 6 and 7).

of job testing. Our strong expectation is that stores that saw greater increases in worker 'quality' would have experienced greater gains in productivity. It would be straightforward to combine these estimates with data on test scores by demographic group among applicants and hires to calculate the expected impact of testing on productivity by race.

Unfortunately, the firm that we study did not collect any baseline test score data for cohorts of workers hired prior to the use of test-based screening. Hence, this simple benchmarking exercise is infeasible. 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 (summarized in Table 2). Under the working assumption that the average characteristics of applicants by store were roughly stable before and after the introduction of job testing, these data can be used to benchmark the relationship between test scores and productivity.

Consider the following variant of our main estimating equation for worker tenure:

$$D_{ijt} = \alpha + \delta T_i + \phi_1 \bar{S}_j + \phi_2 \left(\bar{S}_j \times T_{jt}\right) + X_i\beta + \theta_t + \varphi_j + e_{ijt}. \tag{9}$$

Here, the dependent variable is the completed job spell duration of workers hired at each store j, and $\bar{S}_j$ is the average test score of store $j's$ applicants. In this model, we use $\bar{S}_j$ as a proxy for the average 'quality' of applicants at each store. If test scores are predictive of worker productivity – as our analysis so far suggests – the following two relationships should hold in our data: first, *prior to* the advent of testing, stores with lower average applicant quality should exhibit lower overall productivity, as measured by job durations and involuntary termination ($\phi_1 > 0$); second, following the adoption of testing, productivity gains from employment testing should be larger at stores with lower quality applicants ($\phi_2 < 0$) since, prior to the use of testing, a greater fraction of hires at these stores would have been low scoring applicants.

Table 8 presents estimates of equation (9) for job tenure. Column 1 shows a sizable, positive relationship between average applicant quality and the productivity of hires in the pre-testing regime. The highly significant coefficient on the test score variable indicates that, conditional on race, gender, time and state effects, stores with a 1 point higher mean applicant test score obtained an average of 2.7 additional days of job service per non-tested hire. This effect is economically large. A one-standard deviation (equal to 3.7 points across stores) difference in store-level applicant test scores predicts a 10 day difference in store-level job duration. In subsequent columns we add measures of minority resident share and median household income in the store's zip code (calculated from the Census STF files). These measures provide further control for store-level heterogeneity since it is not possible to include store effects in these cross-sectional models. Interestingly, stores in low-income neighborhoods have higher average job spell duration than those in higher income neighborhoods. Inclusion of these

neighborhood-level covariates has little effect on the coefficient of interest, however.

We next estimate this equation for the sample of workers hired under the testing regime. Since these workers were selected based partly upon their test scores, we expect little residual relationship between the average quality of applicants and the productivity of tested hires. Estimates in columns 3 and 4 confirm this expectation. The coefficient on the average applicant test score is only half as large for tested relative to non-tested hires, and it is not significant.

We finally pool workers hired with and without use of the testing to evaluate whether productivity *gains* were larger at stores with lower quality applicant pools, as is implied by the estimates in earlier columns. Column 7, which includes site effects, shows that the mean store in our sample gained 18.3 days of tenure after job testing, whereas a store whose applicants were 5 percentage points below the average store gained 24.3 days of tenure, and a store whose applicants were 5 percentage points above the average store gained 12.2 days of tenure. These results appear to strongly affirm that test scores are predictive of job performance.

Benchmarking the productivity impacts of testing under an 'uninformed' prior

Although we have so far interpreted the *lack of* an effect of testing on productivity differentials by demographic group as evidence that firms were already 'informed' about group differences, an alternative – and far less interesting – explanation for this finding is simply that group differences in test scores were not large enough to render a meaningful change in relative productivity. To assess the relevance of this hypothesis, we benchmark the effects that testing might have had on measured group productivity differentials under the 'uninformed' null.

Consider a hypothetical case where, prior to the advent of testing, firms' hiring practices were uncorrelated with the test measure. This corresponds to the setting first considered by our model in which firms have no prior information about expected worker productivity.[38] Panel A of Table 2 shows that among tested applicants, the Black-White test score gap is 5.4 points (47.7 versus 53.1 points). Under the assumption that pre-test screening was uncorrelated with test scores, this score gap would carry over into the hired sample in its entirety. By contrast, panel B of Table 2 shows that among tested hires, the Black-White test score gap is only 1.5 points. By implication, if screening were initially 'uninformed,' the advent of testing would have reduced the Black-White test score gap among hires by 3.9 points. The analogous figure for Hispanic hires is 2.9 points.

To convert these hypothetical test score gains into potential productivity gains, we multiply the potential change in the score gap by the point estimates in Table 8 (column 2) for the effect of test scores on job spell durations. This calculation indicates that testing could have reduced the Black-

[38] The finding that testing resulted in more productive hires indicates that the opposite extreme assumption – screening was perfectly correlated with the test score – cannot be true.

White job duration gap by a sizable 13 days (3.9×3.2) and the Hispanic-White gap by 9 days (2.9×3.2). These are large effects. In Table 5, we found that testing raised the job spell duration of White hires by 20 days. Under the 'uninformed hiring' null, these calculations imply that Black job spell duration could have risen by 33 days and Hispanic job spell durations by 29 days (i.e., 65 and 45 percent more than the gains for Whites!). Clearly, the evidence in Table 5 does not support these predictions. In fact, the single-tailed 95 confidence intervals for the impact of testing on the job spell durations of Black and Hispanic hires exclude both of these upper bounds.[39]

These calculations indicate that testing had the potential to differentially raise productivity among low-scoring applicant groups. The fact that this did *not* occur confirms that firms were implicitly screening for worker attributes that were correlated with the test even prior to the advent of testing. This does not imply, however, that testing was redundant. Since overall productivity rose significantly, it is clear that testing improved the accuracy of the screening process.[40]

5 Did testing have a disparate impact on minority hiring?

We finally turn to an empirical test of the question posed in the Introduction: did the productivity gains accruing from testing come at a cost of reduced minority hiring? As shown in Figure 2, the test score distributions of Black, White and Hispanic job applicants differ significantly. Yet the test score distributions of Black, White and Hispanic *hires* are far more comparable. The reason, visible in Figure 5, is that the hired population almost entirely excludes the lower tail of the applicant distribution. Since a disproportionate share of Black and Hispanic job applicants are drawn from the lower tail, Figure 5 underscores that job testing had the potential to reduce minority hiring. To benchmark how large this reduction might be, we calculate the expected disparate impact of testing on minority relative to non-minority hiring, again using the null that prior to testing, screening was uncorrelated with test scores (i.e., the 'uninformed' prior).

To form this benchmark, we estimate the following linear probability model for hiring:

$$\Pr\left(H_i = 1\right) = \sum_{n=1}^{100} \pi_n \times 1\left\{S_i = n\right\},$$

[39] A similar set of calculations applied to the termination data show that testing could potentially have closed the Black-White gap in involuntary terminations by 1.7 percentage points (0.44×3.9) and the Hispanic-White gap by 1.3 percentage points (0.40×2.9). Unfortunately, our estimates for the impact of job testing on terminations do not provide adequate precision to distinguish among effect sizes of this magnitude. (Estimates are available from the authors.)

[40] It is interesting to calculate a similar productivity benchmark for overall productivity (that is, not by race group). The overall rise in job spell duration of 18 to 22 days (Table 3, columns 5 and 6) following job testing is equivalent to a rise of 7 to 8 points in the average test scores of hires. By contrast, the data in Table 2 indicate that average scores of hires could have risen by fully 21 points (i.e., by the difference between average scores of applicants and average scores of tested hires) had screening initially been uncorrelated with the test. This calculation underscores that although testing significantly improved screening, pre-test screening was implicitly correlated with the attributes measured by the job test.

where the outcome measure is an indicator variable equal to one if applicant i was hired, S_i is the applicant's test score percentile and $1\{\cdot\}$ is the indicator function. The coefficient vector, π, estimates the hiring rates for White applicants at each test score percentile (hence, the model is saturated).[41] To calculate expected hiring rates by group under the 'uninformed' null, we apply $\hat{\pi}$ to the test score distribution of Whites, Blacks and Hispanics. The hiring rates that emerge from this calculation are 10.2 percent for White applicants (equal to the White mean by construction), 8.8 percent for Black applicants and 9.3 for Hispanic applicants. Under the 'uninformed' prior, these benchmarks imply that job testing would have reduced Black relative to White hiring by 1.4 percentage points and Hispanic relative to White hiring by 0.9 percentage points, that is 14 and 9 percent respectively.[42] As we show below, disparate impacts in the range of 1.4 percentage points (though not 0.9 percentage points) are detectable with available precision in our sample. Did these disparate impacts occur?

As shown in Table 1, simple mean comparisons of hiring by demographic group before and after the use of testing do not suggest that job testing reduced minority employment. In fact, the White share of new hires fell by roughly 5 percentage points in the year following the introduction of testing. This uncontrolled comparison could potentially mask within-store shifts against minority hiring, however. We purge these confounds by again contrasting changes in minority versus non-minority hiring at stores adopting testing relative to stores not adopting testing during the same time interval.

Relative to our prior difference-in-difference estimates, these models present one important complication. The outcome variable of interest in these models 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 $\Pr(M|H,A)$ equal the probability that a new worker is a minority (M) given that he applied (A) and was hired (H). Applying Bayes rule, the log odds ratio that a new hire is a minority (M) versus non-minority (N) is,

$$\begin{aligned} \ln\left(\frac{\Pr(M|H,A)}{\Pr(N|H,A)}\right) &= \ln\left(\frac{\Pr(H|M,A)\cdot\Pr(M|A)}{\Pr(H|A)}\right) - \ln\left(\frac{\Pr(H|N,A)\cdot\Pr(N|A)}{\Pr(H|A)}\right) \\ &= \ln\left(\frac{\Pr(H|M,A)}{\Pr(H|N,A)}\right) + \ln\left(\frac{\Pr(M|A)}{\Pr(N|A)}\right) \end{aligned} \tag{10}$$

This equation shows that the odds that a newly hired worker is a minority is an increasing function

[41] We control for site effects to purge unobserved store-level heterogeneity affecting the hiring rates of all applicants. This has little effect on the results.

[42] As is visible in Table 2 panel C, observed hiring rates for tested Black and Hispanic applicants are in fact *lower* than the predicted rates. This discrepancy is also suggested by Figure 1 where, conditional on test scores, minority applicants are generally less likely to be hired than non-minorities (except as the figure suggests at very high values of the test score – but there is essentially no data in these high ranges). Our model also predicts this pattern. During job interviews, firms will observe applicant characteristics that are not visible in our data, such as dress, comportment, and maturity. Our model immediately implies that at any given test score $\bar{\eta}_1$, minority applicants will have weaker observables (represented by η_2 in our model) than non-minority applicants: $E(\eta_2|\eta_1=\bar{\eta}_1, x=a) > E(\eta_2|\eta_1=\bar{\eta}_1, x=b)$. This implies that minority applicants will have a lower hiring rate than non-minorities conditional on their test scores, which is what we observe in Figure 1.

of both the minority/non-minority hiring rate and the minority/non-minority application rate.

Since we lack data on the minority/non-minority applicant rate prior to the advent of job testing, the second term in equation (10) is a confounding variable that we would like to eliminate.[43] One might speculate, for example, that because the computerized application kiosk requires applicants to submit a social security number and authorize a criminal background check, this could differentially discourage minority applicants.[44] If so, this would bias our results towards the finding that job testing reduced minority hiring – which is not in fact what we find.

To bring equation (10) to the data, we fit the following conditional ('fixed-effects') logit model,

$$E\left(M_i|H_i, A_i\right) = F\left(\xi T_i + X_i\beta + \theta_t + \varphi_j\right), \tag{11}$$

where M 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.[45] The coefficient, ξ, measures the total effect of job testing on the log odds that a newly hired worker is a minority. As underscored by equation (10), this effect potentially operates through both the hiring odds and application rate channels. To isolate the pure effect of testing on the minority/non-minority hiring rate, we would need to make an additional assumption, which is that minority/non-minority application rates are roughly constant within stores over time. If so, the store-level application rates will be 'conditioned out' by the inclusion of store fixed effects and $\hat{\xi}$ will capture the pure impact of job testing on the minority hiring rate. Even if this assumption fails, however, $\hat{\xi}$ still provides a clean 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 9 reports estimates of equation (11) for the hiring rates of Blacks and Hispanics relative to Whites and for the hiring rate of Whites relative to non-Whites. These estimates yield no evidence that employment testing affected relative hiring odds by race. In all specifications, the logit coefficient on the job testing dummy variable is small relative to its standard error ($z < 1$).

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

$$E\left(M_i|H_i, A_i\right) = \zeta T_i + X_i\beta + \theta_t + \varphi_j. \tag{12}$$

[43] Unicru personnel interviewed for this research believe that application kiosks attract more applicants overall but have no information on how kiosks effect applications by race.

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

[45] We use a conditional logit model to estimate equation (11) to avoid the incidental paramaters problem posed when estimating a standard maximum likelihood model with a very large number of fixed effects (1,363).

Although the linear model is technically misspecified for this problem, it may provide more power to detect small changes in the racial composition of hires. Panel B of Table 9 summarizes these estimates. The dependent variable is multiplied by 100 so that coefficients may be read as percentage points.

In all cases, the impact of testing on hiring rates by race is precisely estimated and close to zero. The point estimates imply that testing reduced Black relative to White hiring rates by less than 0.3 percentage points and reduced Hispanic relative to White hiring rates by less than 0.15 percentage points. None of these estimates is close to significant. The third panel of Table 9 performs instrumental variable estimates of these models, using stores' adoption of testing as an instrument for applicants' tested status. The IV estimates are similar to the corresponding OLS models, implying an even smaller reduction in Black hiring and a slightly larger reduction in Hispanic hiring.

Earlier, we calculated that testing could potentially lower the hiring rate of Black and Hispanic applicants relative to White applicants by 1.4 and 0.9 percentage points respectively. The point estimates in Table 9 suggest that this did not occur. In fact, for Black hiring we can reject the null of a disparate impact of 1.4 percentage points at the 7 percent and 5 percent levels, respectively, for the estimates in columns 1 and 2.[46] We do not have adequate power, however, to reject either the null of 0.9 percentage points or the null of zero for Hispanic hiring.

5.1 Disparate hiring impacts: A second test

Since these results are central to our conclusions, we test their robustness by analyzing a complementary source of variation. As we show below, there is a tight link between the neighborhoods in which stores operate and the race of workers that they hire: stores in minority and low-income zip codes hire a disproportionate share of minority workers. We can use this link to explore whether the introduction of testing systematically changed the relationship between stores' neighborhood demographics and the race of hires. Specifically, we estimate a version of equation (12) augmented with measures of the minority share or median income of residents in the store's zip code, calculated from the 2000 U.S. Census. We first estimate this model separately for tested and non-tested hires at each store (excluding site effects) to assess the cross-sectional relationship between zip code characteristics and the race of hires. We next test formally if job testing changed this relationship.

Table 10 summarizes these estimates. Column 1 of the first panel documents a close correspondence between the race of neighborhood residents and the race of hires. The coefficient of 77 ($t = 25$) on the non-White resident variable indicates that, prior to the use of testing, a store situated in an entirely non-White zip code would be expected to have 77 percent more Black than White hires. Column 2 shows the analogous estimate for tested hires. The point estimate of 78 indicates that the relationship

[46] We use a one-tailed test given our one-sided null.

between store location and worker race was little changed by employment testing.

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

Panel B performs analogous estimates for the racial composition of hires using neighborhood household income in place of zip code minority share. In the pre-testing period, stores in more affluent zip codes had a substantially higher share of White than Black hires; 10 additional log points of neighborhood household income is associated with a 2.6 percentage point higher fraction of White relative to Black hires. Employment testing does not appear to have altered 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.

In net, despite sizable racial differences in test scores, we find no evidence that job testing had disparate racial impacts on hiring. This finding comports with our earlier evidence that testing did not differentially raise the productivity of minority hires. As underscored by our model, if prior to testing, firms were 'uninformed' about group differences in expected productivity measured by the job test, the introduction of job testing would have been expected to differentially reduce minority hiring (a disparate impact) and differentially raise the productivity of minority hires. The fact that neither occurred strongly suggests that, prior to the advent of testing, firms in our sample were 'informed' about group differences in expected productivity and used this information for screening purposes.[47]

6 Conclusion

An influential body of research concludes that the use of standardized tests for employment screening poses an intrinsic equality-efficiency trade-off; raising productivity through better worker selection comes at a cost of reduced opportunity for groups with lower average test scores. This inference rests on the presumption that in the absence of formal job tests, employers do not already informally account for expected productivity differences among applicants from different demographic groups. Accordingly, a test that reveals these differences will disproportionately reduce hiring (and improve

[47] An additional notable pattern visible in the Table 10 estimates is that in most cases, the explanatory power of the regressions linking neighborhood characteristics to the race of hires increases from the pre- to the post-testing regime. This finding suggests that testing reduced the heterogeneity of hiring behavior across stores situated in similar neighborhoods, which is consistent with a rise in screening precision leading to a reduction in the random component of hiring outcomes.

productivity) of workers from low-scoring groups. In a competitive hiring environment, however, this may not be the most relevant case. As our model underscores, if absent testing, employers already roughly account for expected productivity differences among applicant groups, it is possible for employment testing to improve selection without adversely affecting equality.

We studied the evidence for an equality-efficiency trade-off in employment testing at a large, geographically dispersed retail firm whose 1,363 stores switched over the course of 12 months from informal, paper-based hiring to a computer-supported screening process that relies heavily on applicant's scores on a standardized 'customer service' test. We found that the advent of employment testing increased productivity at treated stores, raising mean and median employee tenure by 10 percent, and slightly lowering the frequency of terminations for cause. Consistent with expectations, minority applicants performed significantly worse on the employment test. Had the pre-testing hiring screen been 'uninformed' about the expected productivity differences revealed by the test, we estimated that employment testing would have reduced minority hiring by approximately 10 to 15 percent and raised the productivity of minority hires by 45 to 65 percent more than it raised the productivity of non-minority hires. In fact, neither outcome occurred. We found no evidence that employment testing changed the racial composition of hiring. Moreover, productivity gains were equally large among minority and non-minority hires.

The strongest interpretation of these findings is that firms were optimally statistically discriminating prior to the use of testing, and so the gain in screening precision came at no cost in reduced opportunities for minority applicants. However, our benchmarking exercises makes clear that we lack the statistical power to distinguish this 'optimal' interpretation from a weaker reading of the same facts. It remains plausible, for example, that firms were neither perfectly informed nor perfectly uninformed about group differences; rather they imperfectly screened for the productive attributes measured by the test prior to testing, and the advent of testing raised the precision of this process. What our analysis clearly rejects is the strong disparate impact view embodied in public policy. Had firms been initially 'uninformed' about expected group differences in productivity, testing would have generated disparate impacts of a magnitude ruled out by our analysis.

Several caveats apply to our results. First, our data are from only one large retailer. Since retail firms in the U.S. operate in a competitive environment, we might anticipate that other firms would respond similarly. Nevertheless, analysis of other cases is clearly warranted before firm conclusions may be drawn. A second caveat is that the between-group differences found by the employment test used at this firm are not as large as differences found on other standard ability tests such as the Armed Forces Qualification Test. An alternative employment test that revealed larger group productivity differences might potentially generate disparate impacts. Although we do not discount

this possibility, we generally expect that employers *will* account for expected group productivity differences; hence, a test that reveals large disparities on some measure should not necessarily generate large disparate impacts. Moreover, employment testing guidelines issued by the Equal Employment Opportunity Commission make it difficult and potentially risky for firms to use employment tests that 'pass' minority applicants at less than 80 percent of the pass-rate of non-minority applicants.[48] We therefore do not expect typical employment tests to show greater group differences than those found here.

A final caveat for 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. 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, 2004). Moreover, since testing is itself costly, the net social benefits in the pure screening case could well be negative. Though our results provide little guidance as to which of these scenarios is more relevant, it appears unlikely that the social benefits from testing exceed the private benefits. Quantifying these social benefits is an important topic for future work.

7 Appendix: Details of the theoretical model

In this section, we prove the main theoretical results in section 3.

7.1 Case 1: Screening with no prior information (section 3.2)

Prior to the use of testing, a and b applicants face equal probability of hire, a representative subset of all applicants is hired, and the expected productivity gap between a and b hires is equal to the difference in population mean productivity: $\bar{\eta}_a - \bar{\eta}_b$. After the introduction of testing firms assess applicant productivity as u_1, given by equation (1) in the text. The selection threshold, κ_1 solves $\Pr[u_1 > \kappa_1] = K$, which may be calculated as:

$$\begin{aligned} K &= \frac{1}{2}\left[\Pr\left(u_1 > \kappa_1 | x = a\right) + \Pr\left(u_1 > \kappa_1 | x = b\right)\right] \\ &= \frac{1}{2}\left[\Phi\left(\frac{\rho_0\left(\bar{\eta}_a - \kappa_1\right)}{\sigma_\eta}\right) + \Phi\left(\frac{\rho_0\left(\bar{\eta}_b - \kappa_1\right)}{\sigma_\eta}\right)\right] \end{aligned}$$

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

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $\rho_0 = \left[\sigma_\eta^2/\left(\sigma_1^2+\sigma_\eta^2\right)\right]^{1/2}$ is the correlation between u_1 and η (that is, actual applicant productivity) for each applicant group. Since $\Phi(\cdot)$ is continuous, bounded between 0 and 1, and declining in κ_1, this equation will have a unique solution for κ_1.

1. The effect of testing on a versus b hiring. For any bounded value of κ_1, $\Pr[u_1 > \kappa_1 | x = a] > \Pr[u_1 > \kappa_1 | x = b]$. Consequently, b hiring must fall and a hiring must rise. Aggregate productivity of the pool of hired workers also rises since $E[\eta | u_1 > \kappa_1] > E[\eta]$.

2. The effect of testing on aggregate productivity. Testing raises productivity of both a and b hires. The expected productivity of hired workers from each group x is:

$$E(\eta | x, u_1 > \kappa_1) = \bar{\eta}_x + E(\varepsilon_\eta | x, u_1 > \kappa_1) = \bar{\eta}_x + \rho_0 \sigma_\eta \lambda\left(\frac{\rho_0(\kappa_1 - \bar{\eta}_x)}{\sigma_\eta}\right) > \bar{\eta}_x,$$

   where $\lambda(.)$ is the Inverse Mills Ratio, equal to the density over distribution function of the standard normal: $\phi(\cdot)/(1-\Phi(\cdot)) \geq 0$. By truncating the lower tail of test-takers (those with $u_1 < \kappa_1$), testing increases the expected productivity of hires. Since $\lambda(z) > 0$ for $z \in (-\infty, \infty]$, testing raises the expected productivity of hires from each group.

3. The effect of testing on a versus b productivity. $\partial E(\eta | x, u_1 > \kappa_1)/\partial \bar{\eta}_x = 1 - \lambda'(\rho_0(\kappa_u - \bar{\eta}_x)/\sigma_\eta) < 0$. Since $\bar{\eta}_b < \bar{\eta}_a$, $\lambda'(\cdot), \lambda''(.) \geq 0$, and $\rho_0(\kappa_u - \bar{\eta}_x)/\sigma_\eta$ is increasing in $\bar{\eta}_x$, the expected productivity gains from testing are larger for b than a hires.

7.2 Case 2: Screening with multiple signals (section 3.3)

Firms initially screen using an unbiased job interview and secondarily add the job test. Use of two unbiased screening instruments with uncorrelated measurement error is equivalent to a rise in the accuracy of screening with a single instrument. The correlation between assessed and true applicant productivity using both instruments is:

$$\text{corr}(u_2, \eta) \equiv \rho = \left(\frac{\sigma_1^2\sigma_\eta^2 + \sigma_2^2\sigma_\eta^2}{\sigma_1^2\sigma_2^2 + \sigma_2^2\sigma_\eta^2 + \sigma_1^2\sigma_\eta^2}\right)^{1/2}.$$

This exceeds the correlation between assessed and true productivity using the interview alone: $\text{corr}(u, \eta) = \left[\sigma_\eta^2/\left(\sigma_2^2+\sigma_\eta^2\right)\right]^{1/2} < \rho_0$ for bounded σ_2. Thus, the introduction of testing is equivalent to an increase in ρ.

1. The effect of testing on the hiring threshold. Holding κ at some initial level $\bar{\kappa}$, a rise in ρ reduces hiring from both a and b groups: $\partial \Pr(u_1 > \bar{\kappa} | x)\, \partial\rho = \phi(\cdot)((\bar{\eta}_x - \bar{\kappa})/\sigma_\eta) < 0$ (since

$\bar{\eta}_x < \bar{\kappa}$ for $x \in \{a, b\}$). To maintain overall hiring at K, κ must decline with a rise in ρ. Hence, $\partial\kappa(\rho)/\partial\rho < 0$.

2. The effect of testing on the a versus b hiring. Since total hiring is fixed at K, we have:

$$\frac{\partial K}{\partial \rho} = z_a \cdot \phi\left(\frac{\rho\left(\bar{\eta}_a - \kappa(\rho)\right)}{\sigma_\eta}\right) + z_b \cdot \phi\left(\frac{\rho\left(\bar{\eta}_b - \kappa(\rho)\right)}{\sigma_\eta}\right) = 0,$$

where $z_x = \partial\rho\left[\left(\bar{\eta}_x - \kappa(\rho)\right)/\sigma_\eta\right]/\partial\gamma = \left[\bar{\eta}_x - \kappa(\rho) + \rho\kappa'(\rho)\right]/\sigma_\eta$. Noting that $\phi(\cdot) > 0$, the above expression implies that either Sign$\langle z_a\rangle \neq$Sign$\langle z_b\rangle$. Since we have established that $\kappa'(\rho) < 0$ (and noting that $\bar{\eta}_a > \bar{\eta}_b$), it follows that $z_a > 0$ and $z_b < 0$. Hence testing increases a hiring and reduces b hiring.

3. The effect of testing on aggregate productivity. The expected productivity of hires is

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

and

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

As shown above, $\kappa'(\rho) < 0$, $\kappa(\rho) - \bar{\eta}_b + \rho\kappa'(\rho) > 0$ and $\kappa(\rho) - \bar{\eta}_a + \rho\kappa'(\rho) < 0$. Noting that $\lambda(\cdot), \lambda'(\cdot) > 0$, equation (13) is positive for b hires. Hence, b productivity rises. (This case is made simple by the fact that ρ rises and b hiring falls). To show that equation (13) is also positive for a hires, we use the fact that $0 > \rho\kappa'(\rho) > \bar{\eta}_b - \kappa_n(\rho)$, and substitute into equation (13):

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

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

4. The effect of testing on a versus b productivity. The derivative of expected productivity with respect to testing accuracy is given by equation (13). The first parenthetical expression in this equation is larger (more positive) for b hires since the numerator is declining in $\bar{\eta}_x$ and $\lambda'(\cdot) > 0$. As established above, the second parenthetical expression in this equation is negative for a hires and positive for b hires. Hence, testing differentially increases the productivity of b hires.

7.3 Case 3: Screening with an informed prior (section 3.4)

As in the previous case, the introduction of testing is identically equivalent to a rise in screening precision. But the use of an informed prior means that the expected productivity of a given applicant is adjusted by the group mean in inverse proportion to the accuracy of screening (represented by ρ). Firms assess applicant productivity as u_3 given by equation (4).

1. The effect of testing on the hiring threshold. The selection threshold, κ_3 solves $\Pr[u_3 > \kappa_3] = K$, which may be calculated as:

$$\begin{aligned} K &= \frac{1}{2}\left[\Pr\left(u_3 > \kappa_3 | x = a\right) + \Pr\left(u_3 > \kappa_3 | x = b\right)\right] \\ &= \frac{1}{2}\left[\Phi\left(\frac{\bar{\eta}_a - \kappa_3}{\rho\sigma_\eta}\right) + \Phi\left(\frac{\bar{\eta}_b - \kappa_3}{\rho\sigma_\eta}\right)\right]. \end{aligned}$$

   Holding κ at some initial level $\bar{\kappa}$, a rise in ρ increases hiring from both a and b groups:

$$\frac{\partial \Pr\left(H|x\right)}{\partial\rho} = \frac{\bar{\kappa} - \bar{\eta}_b}{\rho^2\sigma_\eta} \cdot \phi\left(\frac{\bar{\kappa} - \kappa\left(\rho\right)}{\rho\sigma_\eta}\right) > 0.$$

   To maintain overall hiring at K, κ must rise with an increase in ρ. Therefore, $\partial\kappa\left(\rho\right)/\partial\rho > 0$.

2. The effect of testing on the a versus b hiring. Since total hiring is fixed at K, we have:

$$\frac{\partial K}{\partial\rho} = t_b \cdot \phi\left(\frac{\bar{\eta}_b - \kappa\left(\rho\right)}{\rho\sigma_\eta}\right) + t_a \cdot \phi\left(\frac{\bar{\eta}_a - \kappa\left(\rho\right)}{\rho\sigma_\eta}\right) = 0,$$

   where $\tilde{z}_x = \partial\left[\left(\bar{\eta}_x - \kappa\left(\rho\right)\right)/\rho\sigma_\eta\right]/\partial\rho = \left[\kappa\left(\rho\right) - \bar{\eta}_x - \rho\kappa'\right]/\rho^2\sigma_\eta$. Since $\text{Sign}\langle\tilde{z}_a\rangle \neq \text{Sign}\langle\tilde{z}_b\rangle$ and $\kappa'\left(\rho\right) > 0$, the above equation implies that $t_a < 0$ and $t_b > 0$. Hence testing reduces a hiring and increases b hiring.

3. The effect of testing on aggregate productivity. The expected productivity of hires is

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

   The effect of testing on this expectation is:

$$\frac{\partial E(\eta|H, x)}{\partial\rho} = \sigma_\eta\lambda\left(\frac{\kappa\left(\rho\right) - \bar{\eta}_x}{\rho\sigma_\eta}\right) + \sigma_\eta\lambda'\left(\frac{\kappa\left(\rho\right) - \bar{\eta}_x}{\rho\sigma_\eta}\right)\left(\frac{\bar{\eta}_x - \kappa\left(\rho\right) + \rho\kappa'\left(\rho\right)}{\rho\sigma_\eta}\right). \tag{14}$$

   As shown above, $\kappa'\left(\rho\right) > 0$,$\kappa\left(\rho\right) - \bar{\eta}_b - \rho'\kappa\left(\rho\right) > 0$ and $\kappa\left(\rho\right) - \bar{\eta}_a - \rho'\kappa\left(\rho\right) < 0$. This immediately implies that equation (14) is positive for a hires. Hence, a productivity rises with testing. To show that equation (14) is also positive for b hires, we substitute the second of these inequalities ($\rho\kappa'_s\left(\rho\right) > \kappa_s\left(\rho\right) - \bar{\eta}_a$) for $\rho\kappa_s\left(\rho\right)$ into equation (14):

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

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

4. The effect of testing on a versus b productivity. Testing raises the productivity of a relative to b hires, but this effect is small.

$$\begin{aligned}\frac{\partial^2 E(\eta|H,x)}{\partial\rho\partial\bar{\eta}_x} &= \frac{1}{\rho}\left[-\lambda'\left(\frac{\kappa(\rho)-\bar{\eta}_x}{\rho\sigma_\eta}\right) + \lambda''\left(\frac{\kappa(\rho)-\bar{\eta}_x}{\rho\sigma_\eta}\right)\left(\frac{\bar{\eta}_x - \kappa(\rho) + \rho\kappa'}{\rho\sigma_\eta}\right) + \lambda'\left(\frac{\kappa(\rho)-\bar{\eta}_x}{\rho\sigma_\eta}\right)\right] \\ &= \lambda''\left(\frac{\kappa(\rho)-\bar{\eta}_x}{\rho\sigma_\eta}\right)\left(\frac{\bar{\eta}_x - \kappa(\rho) + \rho\kappa'}{\rho^2\sigma_\eta}\right).\end{aligned}$$

This expression is weakly negative for b hires and weakly positive for a hires, implying that testing raises the productivity of a relative to b hires. However, the second derivative of the Inverse Mills Ratio is extremely shallow at all points and asymptotes to zero as the argument of the IMR becomes large. Hence, we conclude that $\partial\left[E(\eta|H,x=a) - E(\eta|H,x=a)\right]/\partial\rho \approx 0$.

8 References

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

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

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

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

Autor, David H. and David Scarborough. 2004. "Will Job Testing Harm Minority Workers?" NBER Working Paper No. 10763, September.

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

Bertrand, Marriane and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4), September, 991 - 1013.

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

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

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

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

Donohue, John J. III. 1994. "Employment Discrimination Law in Perspective: Three Concepts of Equality" *Michigan Law Review*, 92(2), 2583-2612.

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

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

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

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

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

Honoré, Bo and Adriana Lleras-Muney. 2004. "Competing Risks and the War on Cancer," NBER Working Paper #10963, December.

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

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

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

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

List, John A. 2004. "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field" *Quarterly Journal of Economics*, 119(1), February, 49-90.

Lundberg, Shelly J. and Richard Startz. 1983. "Private Discrimination and Social Intervention in Competitive Labor Markets." *American Economic Review*, 73(3), June, 340-347.

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

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

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

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

Salop, Joanne and Steven Salop. 1976. "Self-Selection and Turnover in the Labor Market," *Quarterly Journal of Economics*, 60, 619-627.

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

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

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

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

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

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

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

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

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

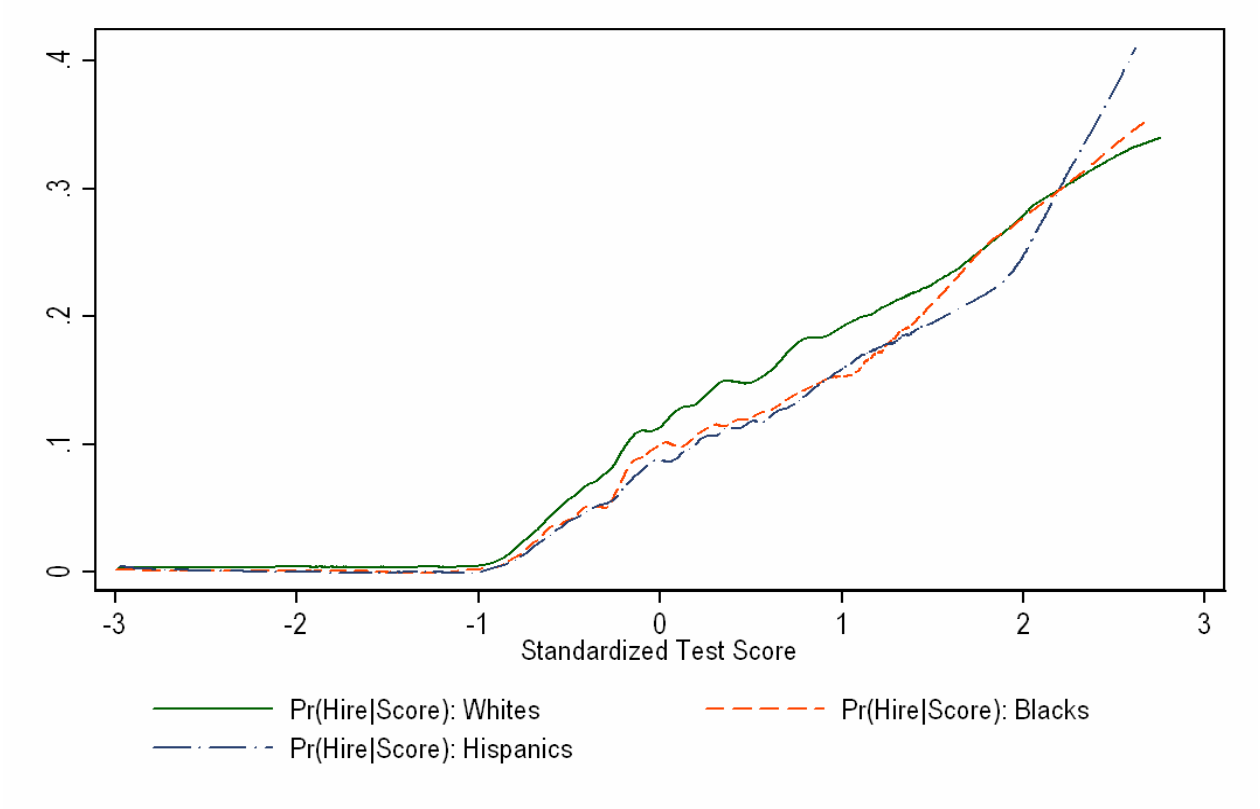

Figure 1. Conditional Probability of Hire as a Function of Test Score by Race: Locally Weighted Regressions

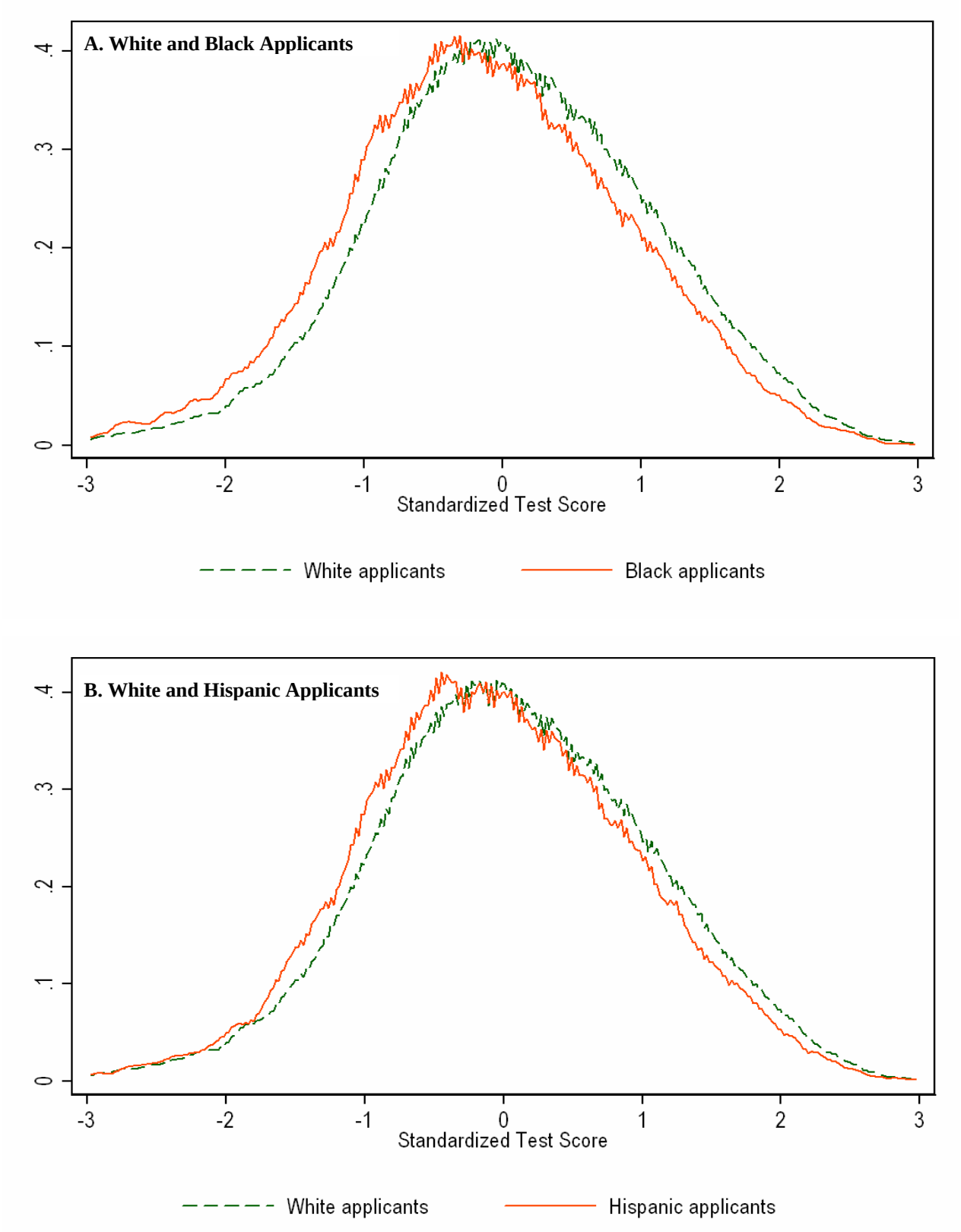

Figure 2. Density of Applicant Test Scores
Sample: All White, Black and Hispanic applicants, June 2000 - May 2001 (*n* =189,067)

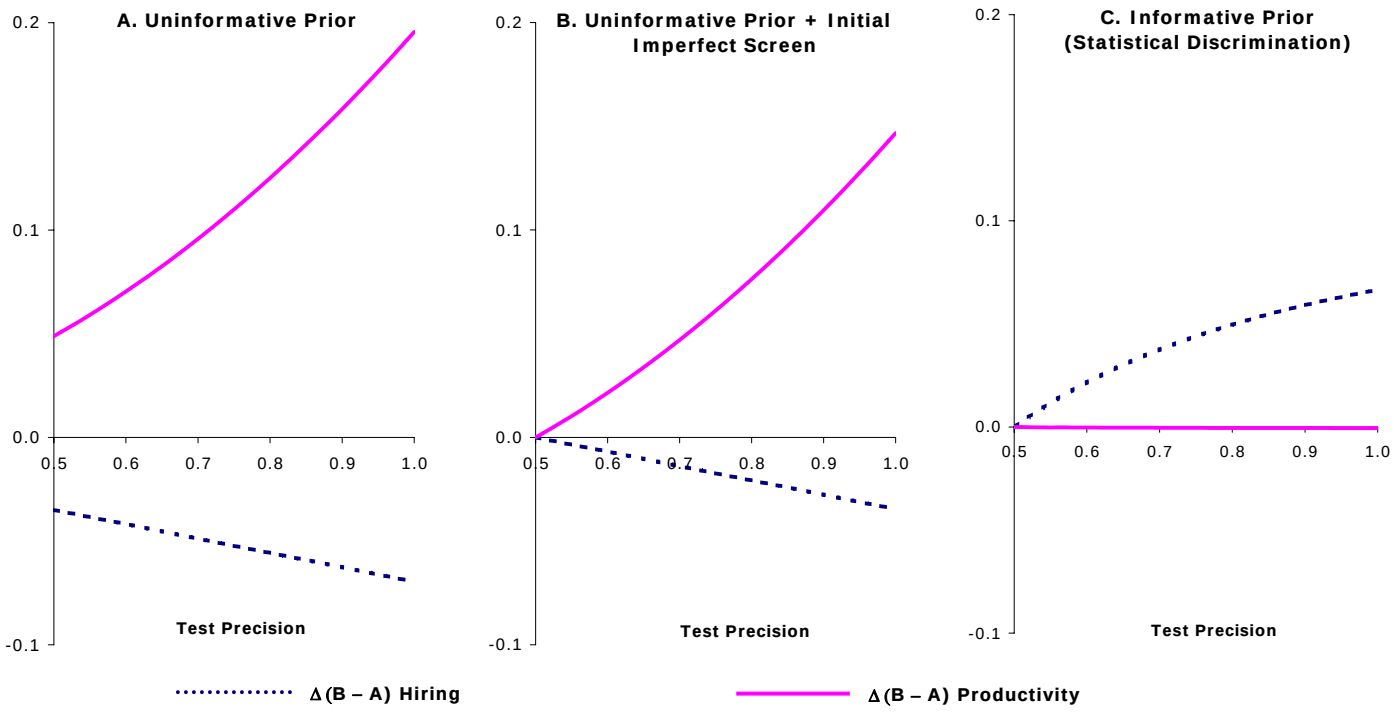

Figure 3. The Impact of Testing on the Hiring and Productivity Gap Between A and B Workers Under Three Priors.

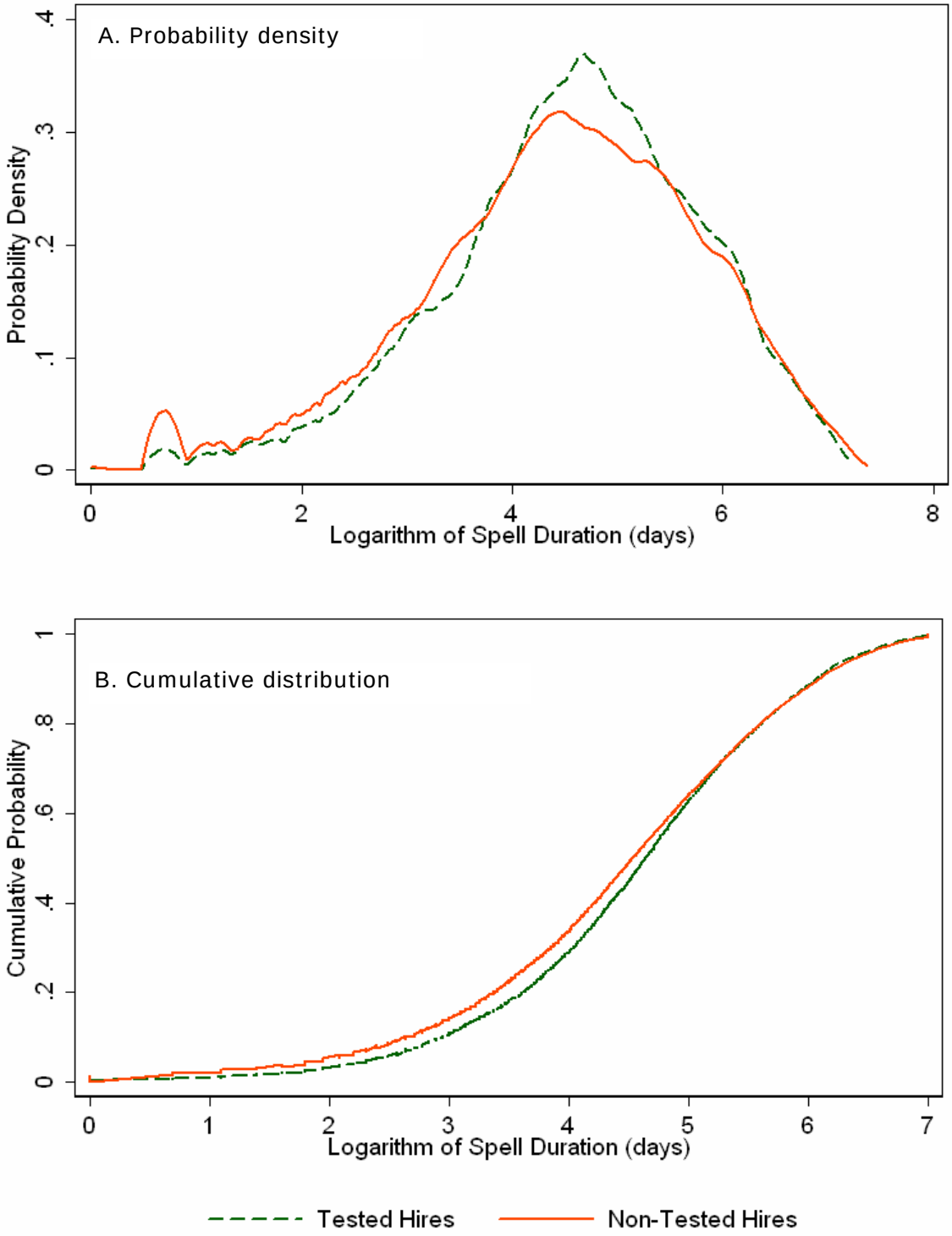

Figure 4. Completed Job Spell Durations of Tested and Non-Tested Hires.
Sample: Hires June 2000 - May 2001 with Valid Outcome Data (n = 33,266)

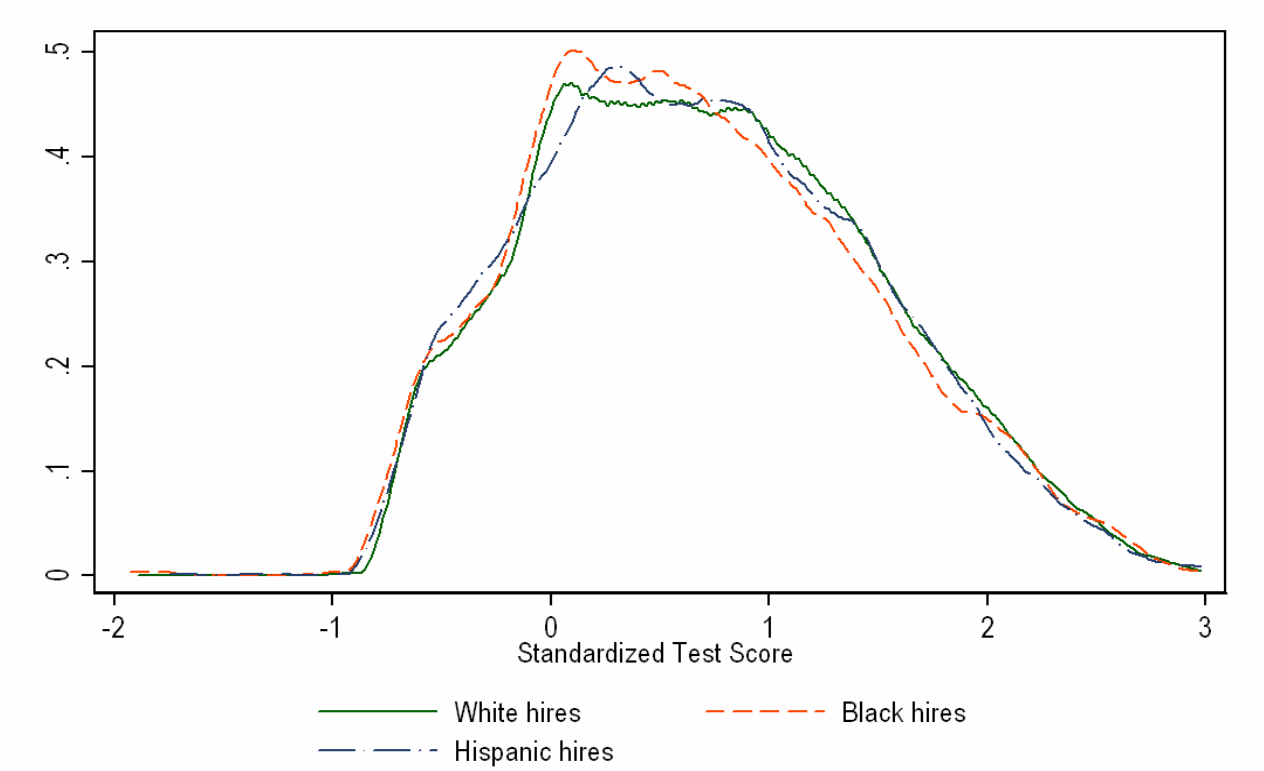

Figure 5. Test Score Densities of Hired Workers by Race

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

Panel A: Frequencies

	Full Sample		Non-Tested Hires		Tested Hires	
	Frequency	% of Total	Frequency	% of Total	Frequency	% of Total
All	33,924		25,561		8,363	
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

Panel B: Employment Spell Duration (days)

	Full Sample		Non-Tested Hires		Tested Hires	
	Mean	Median	Mean	Median	Mean	Median
All	173.7	99	173.3	96	174.8	107
	2	[97, 100]	(2.1)	[94, 98]	(2.9)	[104, 111]
White	184.0	106	183.0	102	187.1	115
	(2.1)	[103, 108]	(2.3)	[100, 105]	(3.6)	[112, 119]
Black	140.1	77	138.1	74	145.7	87
	(3.0)	[75, 80]	(3.5)	[71, 77.4]	(4.8)	[81.9, 92]
Hispanic	166.4	98	169.3	98	159.5	99
	(4.6)	[93, 103]	(5.4)	[92, 104]	(6.4)	[90, 106]

Panel C: Percent Terminated Voluntarily and Involuntarily Within 90 Days

	Full Sample		Non-Tested Hires		Tested Hires	
	Voluntary	Involuntary	Voluntary	Involuntary	Voluntary	Involuntary
All	30.3	16.4	32.0	16.0	24.9	17.5
	(0.3)	(0.3)	(0.4)	(0.4)	(0.6)	(0.6)
White	30.4	14.1	32.1	13.9	24.5	14.8
	(0.4)	(0.3)	(0.5)	(0.3)	(0.7)	(0.6)
Black	30.2	24.5	32.0	24.4	25.4	24.7
	(0.7)	(0.8)	(0.9)	(0.8)	(1.2)	(1.5)
Hispanic	30.0	17.0	31.5	16.0	26.4	19.6
	(0.8)	(0.7)	(1.0)	(0.8)	(1.4)	(1.3)

Table Notes:

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

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

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

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

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

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

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

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

			Percent in each category		
	Mean	SD	Red	Yellow	Green
All	71.9	20.5	0.18	16.06	83.76
White	72.3	20.4	0.14	15.71	84.15
Black	70.8	20.8	0.39	16.41	83.2
Hispanic	71.7	20.6	0.13	17.3	82.57
Male	72.9	20.6	0.23	14.93	84.84
Female	70.8	20.4	0.13	17.4	82.47

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 June 2000 and May 2001 at sites used in treatment sample.

Table 3. OLS and IV Estimates of the Effect of Job Testing on the Job Spell Duration of Hires
Dependent Variable: Length of Completed Employment Spell (days)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			A. OLS Estimates				B. IV Estimates			
Employment test			8.9	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 4. Quantile Regression Estimates of the Effect of Job Testing on Job Spell
Dependent Variable: Length of Employment Spell (days)

	(1)	(2)	(3)	(1)	(2)	(3)	(4)	(5)	(6)
	A. All Spells			B. Completed Spells					
	Median			Median		10th	25th	75th	90th
Employment test		9.0	8.0		10.2	3.0	5.5	17.0	-2.0
		(2.1)	(2.1)		(2.4)	(1.3)	(1.6)	(6.6)	(13.2)
Black	-24.0		-24.0	-22.0	-22.2	-2.0	-7.0	-55.7	-104.0
	(1.9)		(1.7)	(1.7)	(1.9)	(1.0)	(1.3)	(5.3)	(10.3)
Hispanic	-9.7		-10.0	-9.0	-9.3	-1.0	-3.8	-21.3	-39.0
	(2.3)		(2.0)	(2.0)	(2.3)	(1.2)	(1.5)	(6.3)	(12.4)
Male	2.3		2.0	3.0	2.9	2.0	3.3	-7.0	-13.0
	(1.4)		(1.2)	(1.2)	(1.4)	(0.7)	(0.9)	(3.9)	(7.7)
N	33,878	33,878	33,878	33,266	33,266	33,266	33,266	33,266	33,266

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

Table 5. 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)
	A. OLS Estimates			B. IV 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 6. OLS and IV Linear Probability Models for The Effect of Job Testing on Termination Status 90 Days Following Hire

Dependent Variable: Dichotomous Variable Equal to 100 if Worker has Indicated Termination Status

	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)
	Volun-tary	Involun-tary	Volun-tary	Involun-tary	Volun-tary	Involun-tary	Volun-tary	Involun-tary	Volun-tary	Involun-tary
			A. OLS Estimates				B. IV Estimates			
Employment test			-4.59	-1.59	-5.54	-1.78	-2.88	-0.74	-3.83	-0.90
			(1.00)	(0.84)	(1.04)	(0.88)	(1.12)	(0.97)	(1.18)	(1.04)
Black	-0.89	7.37	-0.90	7.37	-0.87	7.35	-0.89	7.37	-0.87	7.35
	(0.86)	(0.77)	(0.86)	(0.77)	(0.86)	(0.76)	(0.86)	(0.77)	(0.86)	(0.76)
Hispanic	-0.81	1.74	-0.81	1.75	-0.80	1.77	-0.81	1.74	-0.80	1.77
	(0.98)	(0.78)	(0.98)	(0.78)	(0.98)	(0.78)	(0.98)	(0.78)	(0.98)	(0.78)
Male	-3.68	2.12	-3.68	2.12	-3.66	2.10	-3.68	2.12	-3.66	2.10
	(0.54)	(0.44)	(0.54)	(0.44)	(0.54)	(0.44)	(0.54)	(0.44)	(0.54)	(0.44)
State trends	No	No	No	No	Yes	Yes	No	No	Yes	
R-squared	0.080	0.102	0.081	0.102	0.083	0.104				

Table Notes:

- N=32,933

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

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

-Sample includes workers hired Jan 1999 through May 2000.

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

Table 7. OLS and IV Linear Probability Models for The Effect of Job Testing on Termination Status 90 Days Following Hire
Dependent Variable: Dichotomous Variable Equal to 100 if Worker has Indicated Termination Status

	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
	Volun-tary	Involun-tary	Volun-tary	Involun-tary	Volun-tary	Involun-tary	Volun-tary	Involun-tary
	A. OLS Estimates				B. IV Estimates			
White x tested	-5.31	-1.65	-6.15	-1.81	-3.69	-0.83	-4.53	-0.94
	(1.09)	(0.87)	(1.13)	(0.91)	(1.22)	(1.03)	(1.27)	(1.09)
Black x tested	-3.46	-2.80	-4.52	-3.00	-1.33	-2.18	-2.37	-2.41
	(1.58)	(1.58)	(1.61)	(1.59)	(1.75)	(1.70)	(1.80)	(1.73)
Hispanic x tested	-3.05	0.38	-4.29	0.21	-1.50	1.62	-2.72	1.52
	(1.82)	(1.55)	(1.90)	(1.62)	(1.93)	(1.66)	(2.03)	(1.75)
Black	-1.33	7.65	-1.25	7.64	-1.45	7.70	-1.37	7.70
	(0.95)	(0.84)	(0.95)	(0.84)	(0.96)	(0.85)	(0.96)	(7.70)
Hispanic	-1.42	1.16	-1.30	1.19	-1.39	1.04	-1.28	1.07
	(1.12)	(0.84)	(1.13)	(0.84)	(1.14)	(0.84)	(1.15)	(0.84)
Male	-3.67	2.11	-3.66	2.09	-3.67	2.11	-3.66	2.09
	(0.54)	(0.44)	(0.54)	(0.44)	(0.54)	(0.44)	(0.54)	(0.44)
State trends	No	No	Yes	Yes	No	No	Yes	Yes
H_0: Race interactions jointly equal	0.292	0.264	0.422	0.273	0.258	0.187	0.363	0.181
R-squared	0.081	0.102	0.083	0.104				

Table Notes:
N=32,933
-Robust standard errors in parentheses account for correlation between observations from the same site hired under each screening method (testing or no testing).
-All models include 1,363 site fixed effects and controls for month-year of hire.
-Sample includes workers hired Jan 1999 through May 2000.
-Instrument for worker receiving employment test is an indicator variable equal to one if site has begun testing.

Table 8. The Relationship between Site-Level Applicant Mean Test Scores and Job Spell Duration of Hires

Dependent Variable: Length of Employment Spell (days)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Non-Tested Hires		Tested Hires		All Hires		
Mean applicant test score	2.70	3.20	1.07	1.61	2.82	3.24	
	(0.56)	(0.73)	(0.81)	(1.04)	(0.60)	(0.65)	
Mean applicant test score x tested						-1.51	-1.18
						(0.81)	(0.62)
Worker received employment test						8.04	18.32
						(4.79)	(4.03)
Log median income in store zip code		-14.76		-24.33	-16.57	-17.37	
		(7.29)		(11.14)	(6.13)	(6.16)	
Share non-white in store zip code		-2.04		-9.12	-1.88	-3.03	
		(12.43)		(17.67)	(10.20)	(10.23)	
State effects	Yes	No	No	No	Yes	Yes	No
Site effects	No	No	Yes	No	No	No	Yes
R-squared	0.024	0.024	0.025	0.026	0.022	0.022	0.110
N	25,089	25,089	8,177	8,177	33,266	33,266	33,266

Table Notes:

-Robust standard errors in parentheses account for correlation between observations from the same site (and, in columns 5 - 7, hired under each screening method: testing or no testing).

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

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

-Applicant test sample includes all applications submitted from June 2000 through May 2001 at treatment sites (189,067 applicants total).

Table 9. Conditional Logit and Linear Probability Models of The Effect of Job Testing on Applicant Hiring Odds by Race
Dependent Variable: An indicator variable equal to 100 if hired worker is of given race

	(1)	(2)	(3)	(4)	(5)	(6)
	Black vs. White		Hispanic vs. White		White vs. Non-White	
			Panel A: Fixed Effects Logit Estimates			
Employment test (logit coefficient)	-0.023	-0.005	-0.026	-0.050	0.029	0.021
	(0.071)	(0.075)	(0.078)	(0.081)	(0.056)	(0.059)
State trends	No	Yes	No	Yes	No	Yes
N	23,597	23,957	18,636	18,636	30,921	22,453
			Panel B: OLS Estimates			
Employment test (OLS coefficient)	-0.28	-0.07	-0.14	-0.10	0.41	0.24
	(0.77)	(0.81)	(0.73)	(0.79)	(0.84)	(0.89)
State trends	No	Yes	No	Yes	No	Yes
N	29,822	29,822	27,622	27,622	33,924	33,924
			Panel B: Instrumental Variables Estimates			
Employment test (IV coefficient)	-0.25	-0.07	-0.68	-0.70	0.78	0.69
	(0.88)	(0.92)	(0.83)	(0.92)	(0.95)	(1.02)
State trends	No	Yes	No	Yes	No	Yes
N	29,822	29,822	27,622	27,622	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 10. The Relationship Between Store Zip Code Demographics and Race of Hires
Before and After Use of Job Testing
Dependent Variable: An Indicator Variable Equal to 100 if Hired Worker is of Given Race

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Black vs. White				Hispanic vs. White				White vs. Non-White			
	Not Tested	Tested	All	All	Not Tested	Tested	All	All	Not Tested	Tested	All	All
Panel A: Race of Hires and Racial Composition of Store Zip-Code												
Share non-white in zip code	77.0 (3.1)	77.5 (4.3)	76.8 (2.9)		61.4 (3.3)	65.5 (4.6)	62.2 (3.1)		-87.4 (2.3)	-86.1 (3.4)	-87.6 (2.2)	
Share non-white x tested			2.2 (4.3)	2.1 (1.9)			0.6 (4.6)	-2.8 (2.2)			1.3 (3.3)	-0.3 (1.8)
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	0.222	0.264	0.232	0.390	0.205	0.199	0.202	0.350	0.231	0.253	0.236	0.353
N	22,648	7,174	29,822	29,822	20,970	6,692	27,662	27,662	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	-25.7 (2.7)	-32.5 (3.5)	-25.8 (2.6)		-18.9 (2.0)	-26.3 (3.0)	-19.3 (2.0)		32.0 (2.5)	39.5 (3.1)	32.2 (2.4)	
Log median income x tested			-5.9 (4.0)	-0.4 (1.6)			-4.7 (3.4)	0.5 (1.4)			5.9 (3.8)	0.6 (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	0.113	0.162	0.123	0.390	0.129	0.134	0.128	0.350	0.117	0.155	0.124	0.353
N	22,648	7,174	29,822	29,822	20,970	6,692	27,662	27,662	22,648	7,174	29,822	29,822

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

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

	(1)	(2)	(3)	(4)
Store has adopted test	0.888	0.862	0.863	0.852
	(0.008)	(0.010)	(0.007)	(0.008)
Black	0.002	0.004	-0.001	0.000
	(0.003)	(0.003)	(0.003)	(0.003)
Hispanic	0.008	0.006	0.003	0.003
	(0.003)	(0.003)	(0.003)	(0.003)
Male	0.000	0.001	0.000	0.000
	(0.002)	(0.002)	(0.002)	(0.001)
State trends	No	Yes	No	Yes
Site effects	No	No	Yes	Yes
R-squared	0.892	0.896	0.909	0.910

Table Notes:
-N=33,924 includes workers hired Jan 1999 through May 2000.
-Robust standard errors in parentheses account for correlation between observations from the same site hired under each screening method (testing or no testing).
-All models include controls for month-year of hire.

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

Month relative to adoption of testing	(1)	(2)
5 months prior	5.3	4.6
	(6.2)	(6.2)
4 months prior	7.9	7.3
	(5.9)	(6.0)
3 months prior	-8.0	-7.7
	(5.9)	(5.9)
2 months prior	-7.2	-6.5
	(5.8)	(5.8)
1 month prior	6.9	7.7
	(6.6)	(6.7)
Month of rollout	13.3	15.8
	(6.5)	(6.6)
1 month post	26.9	30.5
	(7.9)	(8.0)
2 months post	25.2	29.1
	(8.3)	(8.6)
3 months post	18.9	25.0
	(9.4)	(9.8)
4+ months post	19.7	31.5
	(8.5)	(9.8)
State Trends	No	Yes
R-squared	0.110	0.112

Table Notes:
-N=33,266
-Robust standard errors in parentheses account for correlation between observations from the same site.
-All models include controls for month-year of hire.
-Sample includes workers hired Jan 1999 through May 2000.