DISCRETION IN HIRING*

Mitchell Hoffman
University of Toronto & NBER

Lisa B. Kahn
Yale University & NBER

Danielle Li
MIT & NBER

September 2017

Abstract

Job testing technologies enable firms to rely less on human judgement when making hiring decisions. Placing more weight on test scores may improve hiring decisions by reducing the influence of human bias or mistakes but may also lead firms to forgo the potentially valuable private information of their managers. We study the introduction of job testing across 15 firms employing low-skilled service sector workers. When faced with similar applicant pools, we find that managers who appear to hire against test recommendations end up with worse average hires. This suggests that managers often overrule test recommendations because they are biased or mistaken, not only because they have superior private information.

JEL Classifications: M51, J24

Keywords: Hiring; rules vs. discretion; job testing

---

*We are grateful to Jason Abaluck, Ajay Agrawal, Ricardo Alonso, Pol Antras, Ian Ball, David Berger, Arthur Campbell, David Deming, Alex Frankel, Avi Goldfarb, Lawrence Katz, Harry Krashinsky, Peter Landry, Jin Li, Liz Lyons, Steve Malliaris, Mike Powell, Kathryn Shaw, Steve Tadelis, numerous seminar participants, and anonymous referees for helpful comments. We are grateful to the anonymous data provider for providing access to proprietary data. Hoffman acknowledges financial support from the Social Science and Humanities Research Council of Canada. All errors are our own. Correspondence: Mitchell Hoffman, University of Toronto Rotman School of Management, 105 St. George St., Toronto, ON M5S 3E6. Email: mitchell.hoffman@rotman.utoronto.ca. Lisa Kahn, Yale School of Management, 165 Whitney Ave, PO Box 208200, New Haven, CT 06511. Email: lisa.kahn@yale.edu. Danielle Li, MIT Sloan School of Management, 100 Main Street, Cambridge, MA 02142. Email: d_li@mit.edu.

I Introduction

Hiring the right workers is one of the most important and difficult problems that a firm faces. Resumes, interviews, and other screening tools are often limited in their ability to reveal whether a worker has the right skills or will be a good fit. Further, the managers that firms employ to gather and interpret this information may have poor judgement or preferences that are imperfectly aligned with firm objectives.[1] Firms may thus face both information and agency problems when making hiring decisions.

The increasing adoption of "workforce analytics" and job testing has provided firms with new hiring tools.[2] Job testing has the potential to both improve information about the quality of candidates and to reduce agency problems between firms and human resource (HR) managers. As with interviews, job tests provide an additional signal of a worker's quality. Yet, unlike interviews and other subjective assessments, job testing provides information about worker quality that is directly verifiable by the firm.

What is the impact of job testing on the quality of hires and how should firms use job tests? In the absence of agency problems, firms should allow managers discretion to weigh job tests alongside interviews and other private signals when deciding whom to hire. Yet, if managers are biased or if their judgment is otherwise flawed, firms may prefer to limit discretion and place more weight on test results, even if this means ignoring the private information of the manager. Firms may have difficulty evaluating this trade off because they cannot tell whether a manager hires a candidate with poor test scores because of private evidence to the contrary, or because he or she is biased or simply mistaken.

In this paper, we evaluate the introduction of a job test and analyze the consequences of making hiring decisions that deviate from test score recommendations. We use a unique personnel dataset consisting of 15 firms who employ workers in the same low-skilled service sector. Prior to the introduction of testing, firms employed HR managers who were involved in hiring new workers. After the introduction of testing, HR managers were also given access

[1] For example, a manager could have preferences over demographics or family background that do not maximize productivity. In a case study of elite professional services firms, Riviera (2014) shows that one of the most important determinants of hiring is the presence of shared leisure activities.

[2] See, for instance, *Forbes*: http://www.forbes.com/sites/joshbersin/2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age/.

to a test score for each applicant: green (high potential candidate), yellow (moderate potential candidate), or red (lowest rating). Managers were encouraged to factor the test into their hiring decisions, but were not required to hire strictly according to test recommendations.

We first estimate the impact of introducing the job test on the quality of hired workers. Exploiting the staggered introduction of job testing across sample locations, we show that cohorts of workers hired with job testing have substantially longer tenures than cohorts of workers hired without testing, holding constant a variety of time-varying location and firm variables. In our setting, job tenure is a key measure of quality because turnover is costly and workers already spend a substantial fraction of their tenure in paid training. This finding suggests that this job test contains useful information about the quality of candidates.

Next, we examine how managers use job test information. We propose a model in which firms rely on potentially biased HR managers who observe a private signal of worker quality in addition to the publicly observable job test. Managers can decide to hire workers with the best test scores or make "exceptions" by hiring against the test recommendation. In the absence of bias, managers make exceptions only when they have additional information, resulting in better hires for the firm. However, biased managers are also more likely to make exceptions, and these exceptions lead to worse hires on average. This model thus provides intuition for why the observed relationship between a manager's propensity to make exceptions and worker outcomes can be informative about the role of bias in hiring: a positive relationship suggests that managers primarily make exceptions when they are better informed, while a negative relationship suggests the presence of bias or mistaken beliefs.

Our data, which includes information on applicants as well as hired workers, allows us to explore this relationship empirically. We define an "exception" as hiring an applicant with a yellow test score when one with a green score had also applied but is not hired (or similarly, when a "red" applicant is hired while a "yellow" or "green" is not). Across a variety of specifications, we find that exceptions are strongly correlated with worse outcomes. Even controlling for the test scores of the applicant pools they hire from, managers who appear to make more exceptions systematically bring in workers who leave their jobs more quickly. This result suggests that managers make exceptions not only when they are better informed but also because they are biased or mistaken.

Finally, we show that our results are unlikely to be driven by the possibility that managers sacrifice job tenure in search of workers who have higher quality on other dimensions. If this were the case, limiting discretion may improve worker durations, but at the expense of other quality measures. To examine this possibility, we examine the relationship between hiring exceptions and a direct measure of individual productivity, daily output per hour, which we observe for a subset of firms in our sample. In this supplemental analysis, we find no evidence that exceptions are related to increased productivity; this makes it unlikely that managers trade off duration for productivity.

Our empirical approach differs from an experiment in which discretion is granted to some managers and not others. Rather, our analysis exploits differences across managers in the extent to which they appear to make exceptions by overruling test recommendations. Our approach uses this non-random variation in willingness to *exercise* discretion to infer whether discretion facilitates better hires. If managers use discretion only when they have better information, then managers who make more exceptions should have better outcomes than managers who do not. If exceptions are instead associated with worse outcomes, then it is likely that managers are also biased or mistaken.

The validity of this approach relies on two key assumptions. First, we must be able to isolate variation in exceptions that is reflective of managerial choices, and not driven by lower yield rates for higher quality applicants. A weakness of our data is that we do not observe job offers; because of this, managers who hire yellow or red workers only after green applicants have turned down job offers will mistakenly look as though they made more exceptions. Second, it must also be true that the unobserved quality of applicants are similar across low- and high-exception cohorts. For example, we want to rule out cases where managers make exceptions precisely because the pool of green applicants is idiosyncratically weak. We discuss both of these assumptions in more detail throughout the text and estimate specifications that either directly address or limit these concerns.

As data analytics is more frequently applied to human resource management decisions, it becomes increasingly important to understand how these new technologies impact the organizational structure of the firm and the efficiency of worker-firm matching. While a large theoretical literature has studied how firms should allocate authority, and a smaller

empirical literature has examined discretion and rule-making in other settings, empirical evidence on discretion in hiring is scant.[3] Our paper provides a first step towards an empirical understanding of the potential benefits of discretion in hiring. Our findings provide evidence that screening technologies may improve information symmetry between firms and managers. In this spirit, our paper is related to the classic Baker and Hubbard (2004) analysis of the adoption of on board computers in the trucking industry.

Our work is most closely related to Autor and Scarborough (2008), the first paper in economics to provide an estimate of the impact of job testing on worker performance.[4] The authors evaluate the introduction of a job test in retail trade, with a particular focus on whether testing will have a disparate impact on minority hiring. We also find positive impacts of testing, and, from there, focus on the complementary question of the consequences of overruling the job test. Our results are broadly aligned with findings in psychology and behavioral economics that emphasize the potential of machine-based algorithms to mitigate errors and biases in human judgement across a variety of domains.[5]

The remainder of this paper proceeds as follows. Section II describes the setting and data. Section III evaluates the impact of testing on job duration. Section IV presents a model of hiring with potentially biased managers. Motivated by the model, Section V empirically assesses whether managers use their discretion to improve hires. Section VI concludes. All appendix material can be found in the Online Appendix.

[3]For theoretical work, see the canonical Aghion and Tirole (1997), the Bolton and Dewatripont (2012) survey, and Dessein (2002) and Alonso and Matouschek (2008) for particularly relevant instances. For empirical work, see for example, Paravisini and Schoar (2012) and Wang (2014) for analyses of loan officers, Li (2017) on grant committees, Kuziemko (2013) on parole boards, and Diamond and Persson (2016) on teacher grading.

[4]We also contribute to the broader literatures on screening technologies (e.g., Autor (2001), Stanton and Thomas (2015), Horton (2017), Brown, Setren, and Topa (2016), Burks et al. (2015), and Pallais and Sands (2016)) and employer learning (Farber and Gibbons (1996), Altonji and Pierret (2001), and Kahn and Lange (2014)).

[5]See Kuncel et. al. (2013) for a meta-analysis of this literature, Kahneman (2011) for a behavioral economics perspective, and Kleinberg at al. (2018) for empirical evidence that machine-based algorithms outperform judges in deciding which arrestees to detain pre-trail.

II Setting and Data

Firms have increasingly incorporated testing into their hiring practices. One explanation for this shift is that the rising power of data analytics has made it easier to look for regularities that predict worker performance. We obtain data from an anonymous job testing provider that follows such a model. We hereafter term this firm the "data firm." In this section, we summarize the key features of our setting and dataset. More detail about both the job test and our sample can be found in Section A of the Online Appendix.

II.A Job test and testing adoption

Our data firm offers a test designed to predict performance for a particular job in the low-skilled service sector. We are unable to reveal the exact nature of the job, but it is similar to jobs such as data entry work, standardized test grading, and call center work (and is not a retail store job). The data firm sells its services to clients (hereafter, "client firms") that wish to fill these types of positions. We have 15 such client firms in our dataset.

Across locations, the workers in our data are engaged in a fairly uniform job and perform essentially a single task. For example, one should think of our data as comprised entirely of data entry jobs, entirely of standardized test grader jobs, or entirely of call center jobs. Workers generally do not have other major job tasks to perform. As with data entry, grading, or call center work, workers in our sample engage in individual production: they do not work in teams to create output nor does the pace of their output directly impact others.

The job test provided by our data firm consists of an online questionnaire comprising a large battery of questions, including those on computer/technical skills, personality, cognitive skills, fit for the job, and various job scenarios. The data firm matches applicant responses with subsequent performance in order to identify the various questions that are the most predictive of future workplace success in this setting. Drawing on these correlations, a proprietary algorithm delivers a *green-yellow-red* job test score. In our sample, 48% of applicants receive a green score, 32% score yellow, and 20% score red. See Section A.1 of the Online Appendix for more detail on the test itself.

Job testing was gradually rolled out across locations (establishments) within a given client firm. We observe the date at which test scores appear in our data, but not all workers are tested immediately. Our preferred measure defines test-adoption as the month at which the modal hire in a location had a test score. See Online Appendix A.2 for more discussion and robustness to other definitions.

The HR managers in our data are referred to as recruiters by our data provider and are unlikely to manage day-to-day production. Prior to the introduction of job testing, our client firms gave their HR managers discretion to make hiring recommendations based on interviews and resumes.[6] After adopting this job test, firms made applicant test scores available to managers and encouraged them to factor scores into hiring recommendations, but managers were still permitted to hire their preferred candidate.[7]

II.B Applicant and Worker Data

Our data contain information on hired workers, including hire and termination dates, job function, and worker location. This information is collected by client firms and shared with the data firm. Once a partnership with the data firm forms, we observe additional information, including applicant test scores, application date, and an identifier for the HR manager responsible for a given applicant.

Table I provides sample characteristics. We observe nearly 266,000 hires; two-thirds are observed before testing was introduced and one-third after. Our post-testing sample consists of 400,000 applicants and 91,000 hires assigned to 445 managers.[8]

Our primary worker outcome is job duration. We focus on turnover for three main reasons. Foremost, turnover is a perennial challenge for firms employing low skilled service sector workers. Hence, tenure is an important measure of worker quality for our sample

[6] Other managers may take part in hiring decisions as well. For example, in one firm, recruiters typically endorse a candidate to another manager (e.g., a manager in operations one rank above the frontline supervisor) who will make a "final call."

[7] We do not directly observe managerial authority in our data. However information provided to us by the data firm indicates that managers at client firms were not required to hire strictly by the test, and we see in our data that many workers with low test scores are hired. Also, some client firms had other forms of job testing before partnering with our data firm (see Online Appendix A.3 for details and robustness to restricting the sample to client firms that likely did not have pre-sample testing.).

[8] See Section A.7 of the Online Appendix for sample restrictions.

firms. To illustrate this concern, Figure I shows a histogram of job tenure for completed spells (79% of the spells in our data) among employees in our sample. The median worker (solid line) stays only 99 days, or just over 3 months. One in six workers leave after only a month. Despite these short tenures, hired workers in our sample spend the first several weeks of their employment in paid training.[9] Both our data firm and its client firms are aware of these concerns: in its marketing materials, our data firm emphasizes the ability of its job test to reduce turnover. Second, in addition to its importance for our sample firms, in many canonical models of job search (e.g., Jovanovic 1979), worker tenures can be thought of as a proxy for match quality. As such, job duration is a commonly used measure of worker quality. For example, it is the primary worker quality measure used by Autor and Scarborough (2008), who also study the impact job testing in a low-skilled service sector setting (retail). Finally, job duration is available for all workers in our sample.

For a subset of our client firms, we also observe a direct measure of worker productivity: output per hour.[10] Again, we are not able to reveal the exact nature of the job. That said, output per hour measures the number of primary tasks that an individual worker is able to complete. For example, this would be number of words entered per hour in data entry, number of tests graded in test grading, and number of calls handled in call centers. Recall that in our setting, individuals perform essentially one major task and engage in individual production. Because of the discretized nature of the work, output per hour is a very common performance metric for the type of job we study, and is easily measured. However, our data firm was only able to provide us with this measure for a subset of client firms (roughly a quarter of hired workers). We report these findings separately when we discuss alternative explanations.

The middle panel of Table I provides summary statistics for duration and output per hour. Job durations are censored for the 21% of hired workers who were still employed at the time our data was collected. In our analysis, we take censoring into account by estimating censored normal regressions whenever we use duration as an outcome measure.

[9] Reported lengths of paid training vary considerably, from around 1-2 weeks to around a couple months or more, but is provided by all client firms in our sample.

[10] A similar productivity measure was used in Lazear, Shaw, and Stanton (2015) to evaluate the value of bosses in a comparable setting to ours.

Table I shows that both censored and uncensored job durations increase in color score. For example, among those with completed spells, greens stay 12 days (11%) longer than yellows who stay 18 days (20%) longer than reds. These differences are statistically significant and provide initial evidence that test scores are predictive of worker performance. Further, if managers hire red and yellow applicants only when their unobserved quality is high, then tenure differences in the overall applicant population should be even larger. There is no difference across color score in the share of observations that are censored.[11]

Average output per hour in our dataset is 8.4 and is fairly similar across color. Red workers have somewhat higher productivity along this metric, although these differences are not significant; also, controlling for client firm fixed effects removes any difference in output per hour for red workers. Finally, the bottom panel of Table I shows that scores are predictive of hiring: greens are more likely to be hired than yellows, who are in turn substantially more likely to be hired than reds.

III The Impact of Testing

III.A Empirical Strategy

We first evaluate the impact of introducing testing itself. This analysis helps us understand whether the test has useful information that at least some managers take advantage of. To do so, we exploit the gradual roll-out of testing across locations and over time, and examine its impact on worker quality, as measured by tenure.

$$\text{Log(Duration)}_{ilt} = \alpha_0 + \alpha_1 \text{Testing}_{lt} + \delta_l + \gamma_t + \text{Position}_i\beta + \epsilon_{ilt} \quad (1)$$

Equation (1) compares outcomes for workers hired with and without job testing. We estimate censored normal regressions with individual-specific truncation points to account for the fact that not all workers are observed through the end of their employment spell. We regress log

[11] One thing to note in our table is that, somewhat counterintuitively, job durations are longer for workers hired before testing than afterwards. The main reason for this is mechanical: on average, pre-testing periods are earlier in the sample (by about 16 months), allowing hired workers more time to accrue more tenure. Hire cohort fixed effects account for this effect in our regression analysis.

duration ($\text{Log(Duration)}_{ilt}$) for a worker i, hired to a location l, at time t, on an indicator for whether the location, l had testing at time t (Testing_{lt}). Recall, we assign the test adoption date as the first month in which the modal hire at a location had a test score. After that point, the location is always assigned to the testing regime. We choose to define testing at the location, rather than individual, level in order to avoid the possibility that whether an individual worker is tested may depend on observed personal characteristics (Table A1 of the Online Appendix shows that our results are robust to defining testing at the individual level or at the first date in which any worker is tested).

All regressions include location (δ_l) and month-by-year of hire (γ_t) fixed effects to control for time-invariant differences across locations within our client firms, and for cohort and macroeconomic effects that may impact job duration or censoring probability. We also always include position-type fixed effects (the vector, Position_i, and associated coefficients β) that adjust for small differences in job function across individuals.[12] In some specifications, we also include additional controls, which we describe alongside the results. In all specifications, standard errors are clustered at the location level to account for correlated observations within a location over time.

Section A.3 of the Online Appendix discusses sample coverage of locations over time and shows robustness to using a more balanced panel; Section A.4 explores the timing of testing and assesses whether early testing locations look different on observable characteristics.

III.B Results

Table II reports regression results. Column 1 presents results with controls for location, cohort, and position. In the subsequent columns, we cumulatively add controls. Column 2 adds client firm-by-year fixed effects, to control for the implementation of any new strategies and HR policies that firms may have adopted along with testing.[13] The column 2 coefficient of 0.24 means that employees hired with the assistance of job testing stay, on average, 0.24 log points, longer. Column 3 adds location unemployment rate controls to account for the fact

[12]For example, in data entry, fixed effects would distinguish workers who enter textual data from those who transcribe auditory data, and those who enter data regarding images; in test grading, individuals may grade science or math tests; in call centers, individuals may engage in customer service or sales.

[13]Our data firm has indicated that it was not aware of other client-specific policy changes, though they acknowledge they would not have had full knowledge of whether such changes may have occurred.

that outside job options will impact turnover. In practice, we use education-specific state-level unemployment rates measured at an annual frequency, obtained from the American Community Survey.[14] Finally, Column 4 adds location-specific time trends to account for the possibility that the timing of the introduction of testing is related to trends at the location level, for example, that testing was introduced first to locations that were on an upward (or downward) trajectory.

With full controls, we find that testing improves completed job tenures by 0.23 log points, or just over 25%. These results are broadly consistent with previous estimates from Autor and Scarborough (2008).[15] These estimates reflect the treatment on the treated effect for the sample of firms that select into receiving the sort of test we study. Given that firms often select into receiving technologies based on their expected returns (e.g., Griliches (1957)), it is quite possible that other firms (e.g., those that are less open to new technologies) might experience less of a return.

In addition to log duration, we also examine whether testing impacts the probability that hires reach particular tenure milestones: staying at least three, six, or twelve months. For these samples, we restrict to workers hired three, six, or twelve months, respectively, before the data end date. We estimate OLS models because censoring for these variables is based only on start date and not survival time. The top panel of Online Appendix Table C1 reports results using these milestone measures. We find a positive impact of testing for all these variables, with the most pronounced effects at longer durations. For example, using our full set of controls, we find that testing increases the probability of workers surviving at least 6 months by 6 percentage points (13%) and one year by 7.5 percentage points (23%).

[14]For the 25% of locations that are international, we use aggregated (i.e., non-education-specific), annual, national unemployment rates obtained from the World Bank. For a small set of location identifiers in our data where state cannot be easily assigned (e.g., because workers typically work off-site in different US states), we use national education-specific unemployment rates from the Current Population Survey. We include one set of variables for education-specific unemployment rates (either national or state) and one variable for international unemployment rates. Values are replaced by zeros when missing because of location type and location fixed effects indicate type. Our results are robust to restricting to the 70% of locations with US state-level data.

[15]Although our estimates are larger, the Autor and Scarborough (2008) estimate of 12% is inside the range of our 95% confidence interval with full controls. We also estimated Cox proportional hazard models, and obtained coefficients a bit smaller in magnitude than those from censored normal models, but that were qualitatively similar. Our results are also robust to performing OLS on the length of completed job spells (as in Autor and Scarborough (2008)).

Figure II plots the accompanying event studies. The treatment effect of testing appears to grow over time, suggesting that HR managers and other participants might take some time to learn how to use the test effectively.[16]

Our results in this section indicate that job testing increases job durations relative to the sample firms' initial hiring practices. In the remainder of the paper, we focus on analyzing the consequences of overruling job test recommendations.

IV Model

In this section, we formalize a model in which a firm makes hiring decisions with the help of an HR manager and a job test. As in our sample firms, managers in this model observe job test recommendations, but are not required to hire strictly by the test. The main purpose of this model is to highlight the tradeoffs involved in granting discretion to managers: discretion enables HR managers to take advantage of their private information but also gives them scope to make hires based on biases or incorrect beliefs that are not in the interest of the firm. The model also provides intuition for how we might empirically assess the roles of information and bias/mistakes in hiring. All proofs are in Section B of the Online Appendix.

IV.A Setup

A mass one of applicants apply for job openings within a firm. The firm's payoff of hiring worker i is given by a_i. We assume that a_i is drawn from a distribution which depends on a worker's type, $t_i \in \{G, Y\}$; a share of workers p_G are type G, a share $1 - p_G$ are type Y, and $a|t \sim N(\mu_t, \sigma_a^2)$ with $\mu_G > \mu_Y$ and $\sigma_a^2 \in (0, \infty)$. This worker-quality distribution enables us to naturally incorporate the discrete test score into the hiring environment. We do so by assuming that the test publicly reveals t.[17]

[16] Figure II includes all controls except location time trends so that any pre-trends will be apparent in the figure. Estimates are especially large and noisy 10 quarters after testing, reflecting only a few locations that can be observed to that point. Online Appendix Figure A3 in the Online Appendix replicates Figure II while restricting to a balanced panel of locations that hire in each of the four quarters before and after testing. Impacts there are smaller, but are qualitatively similar.

[17] The values of G and Y in the model correspond to test scores green and yellow, respectively, in our data. We assume binary outcomes for simplicity, even though in our data the signal can take three possible values. This is without loss of generality for the mechanics of the model.

The firm's objective is to hire a proportion, W, of workers that maximizes expected quality, $E[a|Hire]$.[18] For simplicity, we also assume $W < p_G$.[19]

To hire workers, the firm must employ HR managers whose interests are imperfectly aligned with those of the firm. A manager's payoff for hiring worker i is given by:

$$U_i = (1-k)a_i + kb_i.$$

In addition to valuing the firm's payoff, managers also receive an idiosyncratic payoff b_i, which they value with a weight $k \in [0,1]$. We assume that b is independent of (a,t).

Managers may value the firm's payoff, a, because they are directly incentivized to; because they risk termination or desire promotion; or because they are simply altruistic.[20] The additional quality, b, can be thought of in two ways. First, it may capture managerial preferences for certain workers (e.g. for certain demographic groups or those with shared interests). Second, b can represent manager mistakes such as overconfidence that lead them to prefer the wrong candidates.[21]

The parameter k measures the manager's *bias*, i.e., the degree to which the manager's incentives are misaligned with those of the firm or the degree to which the manager is mistaken. An unbiased manager has $k = 0$, while a manager who makes decisions entirely based on bias or the wrong characteristics corresponds to $k = 1$.

The manager privately observes information about a_i and b_i. For simplicity, we assume that the manager directly observes b_i, which is distributed in the population by $N(0, \sigma_b^2)$

[18] In theory, firms should hire all workers whose expected value is greater than their cost. However, one explanation for the hire share rule is that a threshold rule is not contractible because a_i is unobservable. Nonetheless, a firm with rational expectations will know the typical share W of applicants that are worth hiring, and W itself is contractible. Assuming a fixed hiring share is also consistent with the previous literature, for example, Autor and Scarborough (2008).

[19] This implies that a manager could always fill a hired cohort with type G applicants. In our sample of applicant pools (see Table I), the average share of applicants who are green is 48%, the average share of green or yellow applicants who are green is 59%, and the average hiring rate is 19%. Thus, $W < p_G$ will hold for the typical pool.

[20] We do not have systematic data on manager incentives. However, a manager at the data firm told us that HR managers often face some targets and/or incentives. See Online Appendix A.8 for more detail.

[21] For example, a manager may genuinely have the same preferences as the firm but draw incorrect inferences from his or her interview. Indeed, work in psychology (e.g., Dana, Dawes, and Peterson, 2013) shows that interviewers are often overconfident about their ability to read candidates. Such mistakes fit our assumed form for manager utility because we can always separate the posterior belief over worker ability into a component related to true ability, and an orthogonal component resulting from their error.

with $\sigma_b^2 \in \mathbb{R}_+$. Second, we assume the manager observes a noisy signal of a_i:

$$s_i = a_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ is independent of a_i, t_i, and b_i, and attributes of applicants are independent across workers, i. The parameter $\sigma_\epsilon^2 \in \mathbb{R}_+$ measures the manager's *information*. A manager with perfect information on a_i has $\sigma_\epsilon^2 = 0$; as σ_ϵ^2 approaches ∞, a manager tends towards having no private information. This private information of managers can be thought of as their assessments of interviews or the worker's overall resume, etc. Unlike the job test, these subjective signals cannot be verifiably communicated to the firm.

Let M denote the set of managers in a firm. For a given manager, $m \in M$, his or her type is defined by the pair $(k, 1/\sigma_\epsilon^2)$, corresponding to the bias and precision of private information, respectively. These have implied subscripts, m, which we suppress for ease of notation. We assume firms do not observe manager type, s_i, or b_i.[22]

Managers form a posterior expectation of worker quality and hire a worker if and only if $E(U_i|s_i, b_i, t_i)$ exceeds some threshold. Managers thus wield "discretion" because they choose how to weigh various signals about an applicant when making hiring decisions. We denote the quality of hires for a given manager under this policy as $E[a|Hire]$ (where an m subscript is implied).

IV.B Model Discussion

This model illustrates the fundamental trade-off inherent in allowing managers discretion over hiring decisions: a manager's private information may be valuable to the firm, but worker quality is hurt by his or her bias.

In practice, firms cannot directly observe a manager's bias or information. However, firms generally do observe the hiring choices of managers and the quality of workers that are hired. We say that when a manager chooses to overrule the test by hiring a Y worker over a G one, the manager is making an "exception." The frequency of such exceptions is increasing

[22] We assume that managers observe the same distribution of other qualities, b, but value them differently, based on the parameter, k. In contrast, the distribution of s_i will vary across managers, based on the value of their private information.

in both a manager's bias and the precision of their private information (Proposition 1 in Online Appendix B). That is, managers are more likely to make exceptions both when their preferences differ from those of the firm and when they have information which is not captured by the test.

To assess whether bias plays a role in driving exceptions, we examine the relationship between a manager's propensity to make exceptions and the realized quality of his or her hires. In the absence of bias, managers only make exceptions when they are better informed, resulting in better hires. By contrast, bias reduces the quality of hires (Proposition 2 in Online Appendix B). Finding a negative relationship between exceptions and the quality of hires would therefore mean that managers make exceptions not only when they are better informed but also because they are biased or mistaken.

The presence of bias raises the possibility that firms may be able to improve outcomes by placing some limits on managerial discretion. To illustrate an example, Proposition 3 (in Online Appendix B) summarizes theoretical conditions under which a firm would prefer no discretion to full discretion. In general, greater bias pushes the firm to prefer no discretion, while better information pushes it towards allowing discretion. These extreme cases need not be optimal and firms may also wish to consider intermediate policies such as limiting the number of exceptions that managers can make.[23]

V Empirical Analysis on Discretion

In our data, we observe job applicants, hires, and the outcomes of hired workers. This allows us to assess how effectively managers exercise discretion, by examining whether worker outcomes are better when managers follow test recommendations more closely, or when they choose to hire more workers as exceptions. If we find that managers who make many exceptions tend to do worse than managers who make few exceptions (holding all else constant), then this suggests that managers are biased. Firms in our setting may then want to consider limiting discretion, at least for the managers that make such frequent exceptions.

[23]See Frankel (2017) for follow up theoretical work providing a discussion of optimal hiring in a related model. In particular, the author shows that there are conditions under which the optimal firm response is to cap the number of exceptions that managers are permitted.

To examine the consequences of overruling test recommendations, we must first define an "exception rate" that corresponds to a manager's choices, rather than his or her circumstances. For example, two managers with the same information and bias may nonetheless make different numbers of exceptions if they face different applicant pools or need to hire different numbers of workers. Our metric should also adjust for applicant pool characteristics that make exceptions mechanically more likely, for example, in pools with few green applicants relative to slots.

Second, we must compare outcomes for managers who have different exception rates despite facing similar applicant pools in similar labor market conditions. These potentially unobservable factors may also separately impact worker tenure, making it difficult to learn about the relationship between exceptions and worker outcomes. For example, we need to address the concern that because we do not observe job offers, applicant pools in which more green workers turn down offers may wrongly appear to have a higher exception rate.

We discuss how we address these issues in the next two subsections. We first define an exception rate that takes into account observable differences in applicant pools. Second, we discuss a range of empirical specifications that help deal with unobserved differences across applicant pools (i.e., differences within color or across locations).

V.A Defining Exceptions

To construct an empirical analogue to the exception rate, we use data on the test scores of applicants and hires in the post-testing period. First, we define an "applicant pool" as a group of applicants being considered by the same manager for jobs at the same location in the same month.[24]

We then measure how often managers overrule the recommendation of the test by either 1) hiring a yellow when a green had applied and is not hired, or 2) hiring a red when a yellow or green had applied and is not hired. We define the exception rate, for a manager m at a

[24] An applicant is under consideration if he or she applied in the last 4 months and had not yet been hired. Over 90% of workers are hired within 4 months of the date they first submitted an application.

location l in a month t, as follows:

$$\text{Exception Rate}_{mlt} = \frac{N_y^h * N_g^{nh} + N_r^h * (N_g^{nh} + N_y^{nh})}{\text{Maximum \# of Exceptions}}, \tag{2}$$

where N_{color}^h and N_{color}^{nh} are the number of hired and not hire applicants, respectively. These variables are defined at the pool level (m, l, t) though subscripts have been suppressed for notational ease.

The numerator of Exception Rate$_{mlt}$ counts the number of exceptions (or "order violations") a manager makes when hiring, i.e., the number of times a yellow is hired for each green that goes unhired plus the number of times a red is hired for each yellow or green that goes unhired. This definition assigns a higher exception rate to a manager when he or she hires a yellow applicant from a pool of 100 green applicants and 1 yellow applicant, than from a pool of 1 green applicant and 100 yellow applicants.

However, the total number of order violations in a pool depends on both the manager's choices and on factors related to the applicant pool, such as size and color composition. For example, if a pool has only green applicants, it is impossible to make any exceptions. Similarly, if the manager needs to hire all available applicants, then there can also be no exceptions. These variations were implicitly held constant in our model, but need to be accounted for in the empirics. To control for pool characteristics that may mechanically impact the number of exceptions, we normalize the number of order violations by the maximum number of violations that could occur, given the applicant pool that the recruiter faces and the number of hires required.[25] This results in an exception rate that ranges from 0 if the manager never made any exceptions, to 1, if the manager made all possible exceptions. Importantly, although the propositions in Section IV are derived for the probability of an exception, their proofs hold equally for this definition of an exception rate. In Online Ap-

[25] That is, we count the number of order violations that would occur if the manager first hired all available reds, then, if there are still positions to fill, all available yellows. Specifically,

$$\text{Maximum \# of Exceptions} = \begin{cases} N^h(N_g^A + N_y^A) & \text{if } N^h \leq N_r^A \\ N^h N_g^A + N_r^A(N_y^A - (N^h - N_r^A)) & \text{if } N_r^A < N^h \leq N_y^A + N_r^A \\ (N_r^A + N_y^A)(N_g^A - (N^h - N_r^A - N_y^A)) & \text{if } N_y^A + N_r^A < N^h \end{cases}$$

where N_{color}^h is the number of applicants of a given color and N^h is the total number of hires.

pendix A.6 we show that our results are also robust to alternative definitions of exception rates.

As described in Table I, we observe nearly 3,700 applicant pools consisting of, on average, 260 applicants.[26] On average, 19% of workers in a given pool are hired and this proportion is increasing in the score of the applicant. Despite this, exceptions are common: the average exception rate across applicant pools is 22%.

There is substantial variation in exception rates across both applicant pools and managers. Figure III shows histograms of the exception rate at the application pool level in the top panel. The left graph shows the unweighted distribution, while the right graph shows the distribution weighted by the number of hires. In either case, the median exception rate is about 20% of the maximal number of possible exceptions, and the standard deviation is about 15 percentage points.

We then aggregate exception rates to the manager level by averaging over all pools a manager hired in, weighting by the number of hires in the pool. The middle panels of Figure III show histograms of manager-level exception rates: these have the same same mean and a slightly smaller standard deviation of 10 percentage points. This means that managers very frequently make exceptions, and some managers consistently make more exceptions than others. The bottom panels of Figure III aggregate exception rates to the location level and show that there is also systematic variation in exception rates across locations.

When we examine the relationship between manager-level exception rates and worker outcomes, we require that variation in exception rates reflect differences in manager choices, driven by their information and biases. However, it is possible that other factors such as unobserved differences in applicant quality also influence exceptions rates. In the next section, we describe how our empirical analysis seeks to control for such confounders.

V.B Empirical Specifications

Post-testing Correlation Between Exception Rates and Outcomes. We first examine the relationship between the manager-level exception rate and the realized durations of hires in

[26] This excludes months in which no hires were made.

the post-testing period:

$$(3) \qquad \text{Log(Duration)}_{imlt} = a_0 + a_1 \text{Exception Rate}_m + \delta_l + \gamma_t + \text{Position}_i\beta + \epsilon_{imlt}$$

Our variation comes from differences in manager-level exception rates for managers employed at the same location. In our data, 99.1% of workers are hired at locations with more than one manager. The average location in our sample has nearly 7 managers and the average worker is at a location with 11 managers. The coefficient of interest is a_1. A negative coefficient, $a_1 < 0$, indicates that the quality of hires is decreasing in the manager's exception rate. Such a finding suggests that managers may be making exceptions because they are biased or mistaken, and not solely because they have useful private information. We cluster standard errors at the location level, again to take into account any correlation in observations within a location over time.[27]

We face two key concerns in interpreting a_1. First, exception rates may be driven by omitted variables that separately impact worker durations. For example, some locations may be inherently less desirable places that both attract more managers with biases or bad judgement and retain fewer workers. This would drive a negative correlation between exception rates and outcomes that is unrelated to discretion.

Second, as discussed in the introduction, we observe only hires and not offers. This means that we cannot tell the difference between a yellow that is hired even when a green applicant is available or a yellow that is hired after all green applicants have turned down the offer. One concern is that we may observe more "false exceptions" when green workers have better outside options. In such cases, we may also see lower durations simply because the manager was forced to hire second choice workers.

In both cases, accounting for controls may alleviate some concerns. For example, location fixed effects control for fixed differences across locations in unobserved applicant quality; location-specific time trends further control for smooth changes in these characteristics. Controlling for local labor market conditions reduces the likelihood that our results are driven by "false exceptions," because such exceptions may be more common when green

[27] If we instead cluster by manager, the level of variation underlying our key right hand side variable, we get very similar standard errors.

workers have better outside options. Our full set of controls includes location, time and position type fixed effects, client-year fixed effects, local labor market variables, location-specific time trends, and detailed controls for the quality and number of applicants in an application pool (fixed effects for each decile of: the number of applicants, hire rate, share of applicants that are green, and share that are yellow).[28]

In addition to these controls, examining manager-level exception rates has the benefit of smoothing idiosyncratic variation across individual pools that may drive both exception rates and outcomes. For example, in some pools, green applicants may be atypically weak. In this case, managers may optimally hire yellow applicants, but their hires would still have low average durations relative to workers the location is usually able to attract. Similarly, some pools may have more "false exceptions" because of an idiosyncratically low yield rate for green applicants. Averaging to the manager level (the average manager hires in 18 applicant pools) reduces the extent to which our measure of exceptions is driven by such sources of variation. Given our controls, this same concern would only apply if some managers systematically face idiosyncratically weaker pools or lower yield rates than other managers at the same location facing observably similar pools.

Differential Impact of Testing, by Exception Rates. We also consider how the *impact* of testing differs across exception rates. If managers exercise discretion because they are biased or misinformed, then we may expect the benefits of testing that we document in Section III.B to be lower for high exception managers. We estimate this using a similar specification to that described in Section III.A:

$$\begin{aligned} (4) \qquad \text{Log(Duration)}_{imlt} &= b_0 + b_1 \text{Testing}_{lt} \times \text{Exception Rate}_l + b_2 \text{Testing}_{lt} \\ &\quad + \delta_l + \gamma_t + \text{Position}_i \beta + \epsilon_{imlt} \end{aligned}$$

Equation (4) includes the main effect of testing but allows testing to interact with the location-specific exception rate. We consider location-level variation in this case because we do not observe manager identifiers in the pre-period. The coefficient of interest, b_1, thus estimates how the impact of testing differs at locations that subsequently make more or fewer

[28] We have also explored controls for the number of hires made in the several preceding months to take into account that applicant pools may be depleted over time. Results are very similar with these controls.

exceptions. $b_1 < 0$ indicates that exceptions attenuate gains from testing. Unlike Equation (3), which can only be estimated on post-testing data, this specification uses our full dataset, making it possible for us to more precisely identify location fixed effects and other controls.

V.C Results

Figure IV examines the correlation between exception rates and durations for hired workers after the introduction of testing. We divide managers into 20 equally sized bins based on their hire-weighted exception rate (x-axis) and regress duration measures on an exhaustive set of indicators for each bin, plus base controls. We plot the coefficients on these exception rate bin indicators (y-axis) against the average exception rate in each bin (x-axis). The top left panel summarizes the log duration regression, which adjusts for censoring with a censored-normal regression. The remaining panels plot the milestone measures for the probability that a worker stays at least 3, 6, or 12 months. For all outcomes, we see a negative relationship: job durations are shorter for workers hired by managers with higher exception rates.

Table III presents the accompanying regression analysis. In these regressions, we standardize exception rates to be mean 0 and standard deviation 1 so that the units are easier to interpret. Column 1 contains our base specification and indicates that a one standard deviation increase in the exception rate is associated with a 7% reduction in job durations, significant at the 5% level. Adding controls reduces the size of the standard error and the coefficient slightly. In our full-controls specification, a one standard deviation higher exception rate is associated with 6% shorter durations, still significant at the 5% level. This says that even when we analyze managers at the same location hiring the same number of workers out of pools that have the same share of red, yellow, and green applicants, we continue to find that managers who makes more exceptions do worse. The middle panel of Online Appendix Table C1 summarizes regressions for the milestone measures, where we also find significant negative relationships.

Next, Table IV examines how the impact of testing varies by the extent to which locations make exceptions. Estimates are based on Equation (4). Including the full set of controls (Column 4), we find that at the mean exception rate (recall that we standardize exception rates), testing increases durations by 0.23 log points, but that this effect is substantially offset

(by 0.14 log points) for each standard deviation increase in the exception rate, significant at the 5% level.[29] The bottom panel of Online Appendix Table C1 shows that these results are robust to OLS estimates using milestone measures as dependent variables.

Figure V illustrates how the impact of testing varies for locations with different average exception rates, using base controls.[30] For all tenure outcomes (log(duration) and milestones) we find a negative relationship that does not appear to be driven by any particular exception-rate bin.

Across a variety of specifications, we consistently find that worker tenure is lower for managers who made more exceptions to test recommendations. The magnitude of this estimate implies that a firm made up of managers at the 10th percentile of the exception distribution would have approximately 15% longer worker durations in the post testing period, relative to a firm made up of 90th percentile managers (higher exception rates are worse). Further, locations at the 90th percentile of the exception distribution experience duration improvements with the adoption of testing that are multiple times larger than improvements at 10th percentile locations.

Viewed in light of our theoretical predictions, these results suggest that managers often make exceptions because they are either biased or misinformed. Even if high exception managers were well informed about worker quality, the fact that their hiring outcomes are worse suggests that their biases lead them to make choices that do not maximize quality.

V.D Additional Robustness Checks

In this section we address several alternative explanations for our findings.

Quality of "Passed Over" Workers. There are some scenarios under which we may find a negative correlation between worker outcomes and exception rates, even when managerial discretion improves hiring. For example, as discussed earlier, a manager may tend to make more exceptions because he or she sees idiosyncratically weak green applicants, relative to the

[29] Here, we do not include controls for applicant pool quality because it is unavailable pre-testing.

[30] To construct this, we divide locations into 20 hire-weighted bins based on their average location-level exception rate post testing and augment Equation (4) with indicators for the interaction of exception rate bins and the post-testing dummy. We then plot the bin-specific impact of testing coefficient on the y-axis and the average exception rate in each bin on the x-axis. For the Log(duration) outcome (top left panel), we adjust for censoring with censored-normal regressions.

typical applicants at the location. As another example, locations with high exception rates may benefit less from the test because its managers always had better private information.

In these and other similar scenarios, it should still be the case that individual exceptions are correct: a yellow hired as an exception should perform better than a green who is not hired. To examine this, we would ideally compare the counterfactual duration of applicants who are not hired with the actual durations of those who were. While this is not generally possible, we can, in some cases, approximate such a comparison by exploiting the timing of hires. Specifically, we compare the tenure of yellow workers hired as exceptions to green workers from the same applicant pool who are not hired that month, but who subsequently begin working in a later month. If managers make exceptions when they have better information, then exception yellows should have longer tenures than "passed over" greens.

Table V shows that is not the case. The first panel compares durations of workers who are exception yellows (the omitted group) to greens whose application was active in the same month, but were hired only in a later month. Because these workers are hired at different times, all regressions control for hire month fixed effects to account for mechanical differences in duration. In Column 2, which includes applicant pool fixed effects, the coefficient on "passed over green" compares this group to yellow applicants from the *same* applicant pool who were hired before them.[31] The second panel of Table V repeats this exercise, comparing exception reds (the omitted group), to passed over yellows and greens.[32]

Both panels show that workers hired as exceptions have shorter tenures. Passed over greens stay 4% longer than yellows hired before them from the same pool (Column 2, top panel), though this estimate is noisy. We estimate a more precise relationship for exception-red workers: passed over greens and yellows stay roughly 14% and 12% longer, respectively. These results suggest it is unlikely that exceptions are driven by better information: high scoring workers who are initially passed over outperform low scoring workers chosen first.[33]

[31]The applicant pool fixed effect is at the location-manager-date level, where the date is the month in which both applications were active, the yellow was hired, and the green was hired only later. These fixed effects thus subsume a number of controls from our full specification from Table III.

[32]We restrict observations in both panels to pools in which there was both an exception and a passed over applicant (92% and 59% of hires in the top and bottom panels, respectively). To identify control variables, we further restrict to locations and pools that have at least 10 and 5 observations, respectively.

[33]An alternative explanation is that the applicants with higher test scores were not initially passed up, but were instead initially unavailable, for example because they were engaged in on-the-job search. However, Online Appendix Table C2 shows that delays are not correlated with worker quality.

"False Exceptions." As mentioned, one may be concerned that we do not observe job offers and thus cannot distinguish between cases in which yellow applicants are hired as true exceptions, or when they are hired because green applicants turned down offers. By analyzing manager or location-level exception rates, we aggregate over some of the idiosyncratic variation that may generate false exceptions, and our controls for local labor market conditions may help absorb some of the time-varying drivers of such exceptions.

As an additional test, Online Appendix Table C3 shows that our results hold when restricting to pools with at least as many green applicants as the total number of hires (84% of hires came from such a pool). In such pools, it is less likely that a yellow or red worker was hired because all green applicants received an offer and turned it down.

Heterogeneity Across Locations. Another possible concern is that the relevance of the test varies across locations and that this drives the negative correlation between exception rates and worker outcomes. For example, in very undesirable locations, green applicants might have better outside options and be more difficult to retain. In these locations, a manager attempting to avoid costly retraining may optimally decide to make exceptions in order to hire workers with lower outside options.

In Online Appendix A.5 we provide evidence that the apparent usefulness of the test does not systematically vary by location characteristics. There we explore the relationship between color score and job duration as a function of a wide range of location characteristics, such as exception rates and average durations. We robustly find that color score is predictive of worker quality, regardless of the location's characteristics on each of these dimensions.

Productivity. Our results show that high-exception managers hire workers with lower job duration. These exceptions may still benefit the firm if such workers are better on other dimensions. For example, managers may optimally hire workers who are more likely to turn over if their private signals indicate that those workers might be more productive while they are employed.

Our final set of results provides evidence that this is unlikely to be the case. For a subset of client firms, we observe a direct measure of worker productivity: output per hour. Recall that in our setting, individuals perform essentially one major task and engage in individual production. Some examples may include: the number of data items entered per hour, the

number of standardized tests graded per hour, and the number of phone calls completed per hour. As in these examples, output per hour is an important measure of productivity for the fairly homogenous task we study.

We define a worker-level output per hour metric as a worker's output per hour averaged over all the days where such a metric is observed for that worker. Across all workers with an available measure, output per hour has an average of 8.4 with a standard deviation of 4.7. There is thus a wide range of performance outcomes.[34] From Table I, average output per hour is slightly higher in the post-testing sample period and varies slightly by color score, though the differences are not significant.

This measure is available for 62,427 workers (one-quarter of all hires) in 6 client firms. The primary reason for missing output is that the metric is not made available to us for many locations, time periods, and end clients (i.e., the ones purchasing services from the client firms).[35] In addition, its availability depends on workers completing their training and being permitted to perform the job task: this is the period in which workers become valuable to client firms."

Relative to our main sample, the set of workers with output data is positively selected on duration. Despite this, output remains positively correlated with job duration. Online Appendix Figure C1 presents a binned scatter of output per hour for 20-evenly sized bins of Log(Duration).[36] Except for one outlier for workers with very low tenure, there is a strong positive relationship: workers with longer job durations have higher output per hour.

Table VI summarizes our main analyses using output per hour as the dependent variable. Columns 1-2 document the post-testing correlation between manager-level exceptions and output per hour. For both our base and full sets of controls, we obtain negative coefficients that are not significant. For example in Column 2, the estimate is -0.11 with a standard error of 0.095. Recall the manager-level exception rates are standardized to be mean 0 and standard deviation 1 in the full post-testing sample. The estimate therefore implies that a

[34] We also control for the number of tasks in a day that are used to measure a worker's output per hour. We aggregate this to the worker level by averaging indicators for count decile across all observations for a worker.

[35] We can account for half of the variation in whether an output measure is available for an individual worker with location, time, and position controls. According to the data firm, certain lines of business within a firm do not make their productivity data available.

[36] We control for location fixed effects to account for differences in average output per hour across locations.

one standard deviation higher exception rate manager hires workers who perform 0.11 fewer units of output per hour, on average. This is 2.3% of the standard deviation for output per hour (4.7, mentioned above). Based on the standard error, we can rule out positive effects beyond 1.7% and negative effects beyond -6.4% of a standard deviation with 95% confidence.

Columns 3-4 examine the differential impact of testing by location-level exception rate. The coefficient on testing gives the impact of testing for locations with the mean exception rate (based on the full sample). In the baseline specification, testing improves output per hour by 0.34 or 7% of a standard deviation. This effect is small in magnitude and is not statistically significant. We also find modestly sized and insignificant coefficients for the interaction term. Coefficients are similar in magnitude but opposite in sign across base and full controls. With full controls, the point estimate of -0.137 implies that the impact of testing in a location with a one standard deviation higher exception rate is offset by 0.137 output per hour units, or about 2.9% of a standard deviation. We can rule out positive effects outside of 6.3% and negative effects outside of -12% with 95% confidence.

Although noisy, these findings taken together suggest that output per hour does not appear to be strongly related to exception rates. We do not find evidence of a large negative association between exceptions and output per hour, as we did with job durations. However, in all cases we find no evidence that managerial exceptions *improve* output per hour by any sizeable amount. This is inconsistent with a model in which managers optimally sacrifice job tenure in favor of workers who perform better on other quality dimensions.

VI Conclusion

We evaluate the introduction of a hiring test for a low-skilled service sector job. Exploiting variation in the timing of adoption across locations within firms, we show that testing significantly increases the durations of hired workers. We then document substantial variation in how managers appear to use job test recommendations: some tend to hire applicants with the best test scores while others appear to make many more exceptions. Across a range of specifications, we show that hiring against test recommendations is associated with worse outcomes.

Firms in our setting may find this result useful as they decide how much discretion to grant their managers, and how much to rely on job tests or other signals of worker quality. For example, in cases where high-exception managers are associated with worse outcomes, firms may wish to decrease the rate of exceptions either by limiting managerial discretion (particularly for high exception managers) or by finding new managers who are less biased. More broadly, firms may be able to improve outcomes by adopting policies to influence manager behavior such as increasing feedback about the quality of hires or tying pay more closely to performance. Such policies may encourage managers to find ways to complement the test as they continue to learn.

There are several caveats one must keep in mind in interpreting our results. First, while our results suggest that high-exception managers make decisions with bias, limiting managerial discretion could have unintended consequences (such as demoralizing managers), and could be bad for some managers who do have valuable private information. Second, we emphasize that our findings may not apply to all firms. We focus on workers who perform low-skilled service sector tasks without a teamwork component. A manager's private signals of worker quality may be more valuable in higher skilled settings with more complex tasks.[37] Further, managers may have more opportunities to correct their mistaken beliefs in settings where they regularly interact with applicants on the job. The HR managers we study generally do not supervise applicants after they are hired, which also limits the scope for a manager-employee match component that might make discretion more useful. An additional contribution of our paper is that we present a way to assess the consequences of discretion using only data that would readily be available for many firms using workforce analytics.

More broadly, our findings highlight the role that technology can play in reducing the impact of managerial mistakes or biases by changing how decision-making is structured within the firm. As workforce analytics becomes an increasingly important part of human resource management, more work needs to be done to understand how such technologies

[37] In fact, Frederiksen, Kahn, and Lange (2017) show that managerial discretion over performance management can be valuable in the context of a high-skilled service profession. Li and Agha (2015) show that the judgement of human reviewers provides valuable information about the quality of scientific proposals that is not available from CVs and other quantitative metrics. Hoffman and Tadelis (2017) show that subordinates provide subjective assessments of managers that correlate with hard outcomes in another high skilled setting.

interact with organizational structure and the allocation of decisions rights within the firm. This paper makes an important step towards understanding and quantifying these issues.

UNIVERSITY OF TORONTO & NBER
YALE UNIVERSITY & NBER
MIT & NBER

References

[1] Aghion, Philippe and Jean Tirole, "Formal and Real Authority in Organizations," *Journal of Political Economy*, 105 (1997), 1-29.

[2] Altonji, Joseph and Charles Pierret, "Employer Learning and Statistical Discrimination," *Quarterly Journal of Economics*, 113 (2001), 79-119.

[3] Alonso, Ricardo and Niko Matouschek, "Optimal Delegation," *Review of Economic Studies*, 75 (2008), 259-3.

[4] Autor, David, "Why Do Temporary Help Firms Provide Free General Skills Training?," *Quarterly Journal of Economics*, 116 (2001), 1409-1448.

[5] Autor, David and David Scarborough, "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments," *Quarterly Journal of Economics*, 123 (2008), 219-277.

[6] Baker, George and Thomas Hubbard, "Contractibility and Asset Ownership: On-Board Computers and Governance in U.S. Trucking," *Quarterly Journal of Economics*, 119 (2004), 1443-1479.

[7] Bolton, Patrick and Mathias Dewatripont "Authority in Organizations." in Robert Gibbons and John Roberts (eds.), *The Handbook of Organizational Economics*, (Princeton, NJ: Princeton University Press, 2010).

[8] Brown, Meta, Elizabeth Setren, and Giorgio Topa,"Do Informal Referrals Lead to Better Matches? Evidence from a Firm's Employee Referral System," *Journal of Labor Economics*, 34 (2016), 161-209.

[9] Burks, Stephen, Bo Cowgill, Mitchell Hoffman, and Michael Housman, "The Value of Hiring through Employee Referrals," *Quarterly Journal of Economics*, 130 (2015), 805-839.

[10] Dana, Jason, Robyn Dawes, and Nathaniel Peterson, "Belief in the Unstructured Interview: The Persistence of an Illusion," *Judgment and Decision Making*, 8 (2013), 512-520.

[11] Dessein, Wouter, "Authority and Communication in Organizations," *Review of Economic Studies*, 69 (2002), 811-838.

[12] Diamond, Rebecca and Petra Persson, "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests," mimeo Stanford University, 2016.

[13] Farber, Henry and Robert Gibbons, "Learning and Wage Dynamics," *Quarterly Journal of Economics*, 111 (1996), 1007-1047.

[14] Frankel, Alexander, "Selecting Applicants," mimeo University of Chicago, 2017.

[15] Frederiksen, Anders, Lisa B. Kahn, and Fabian Lange, "Supervisors and Performance Management Systems," NBER Working Paper #23351, 2017.

[16] Griliches, Zvi, "Hybrid Corn: An Exploration in the Economics of Technological Change," *Econometrica*, 25 (1957), 501-522.

[17] Hoffman, Mitchell and Steven Tadelis, "People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis," working paper, University of Toronto, 2017.

[18] Horton, John, "The effects of algorithmic labor market recommendations: Evidence from a field experiment," *Journal of Labor Economics*, 35 (2017), 345-385.

[19] Jovanovic, Boyan, "Job Matching and the Theory of Turnover," *The Journal of Political Economy*, 87 (1979), 972-90.

[20] Kahn, Lisa B. and Fabian Lange, "Employer Learning, Productivity and the Earnings Distribution: Evidence from Performance Measures," *Review of Economic Studies*, 81 (2014), 1575-1613.

[21] Kahneman, Daniel. *Thinking Fast and Slow.* (New York: Farrar, Strauss, and Giroux, 2011).

[22] Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *Quarterly Jounral of Economics*, 2018, forthcoming.

[23] Kuncel, Nathan, David Klieger, Brian Connelly, and Deniz Ones, "Mechanical Versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis," *Journal of Applied Psychology*. 98 (2013), 1060–1072.

[24] Kuziemko, Ilyana, "How Should Inmates Be Released from Prison? an Assessment of Parole Versus Fixed Sentence Regimes," *Quarterly Journal of Economics*. 128 (2013), 371-424.

[25] Lazear, Edward, Kathryn Shaw, and Christopher Stanton, "The Value of Bosses," *Journal of Labor Economics*, 33 (2015), 823-861.

[26] Li, Danielle, "Expertise and Bias in Evaluation: Evidence from the NIH" *American Economic Journal: Applied Economics*, 9 (2017), 60-92.

[27] Li, Danielle and Leila Agha, "Big Names or Big Ideas: Do Peer Review Panels Select the Best Science Proposals?" *Science*, 348 (2016), 434-438.

[28] Pallais, Amanda and Emily Sands, "Why the Referential Treatment? Evidence from Field Experiments on Referrals," *Journal of Political Economy*, 124 (2016), 1793-1828.

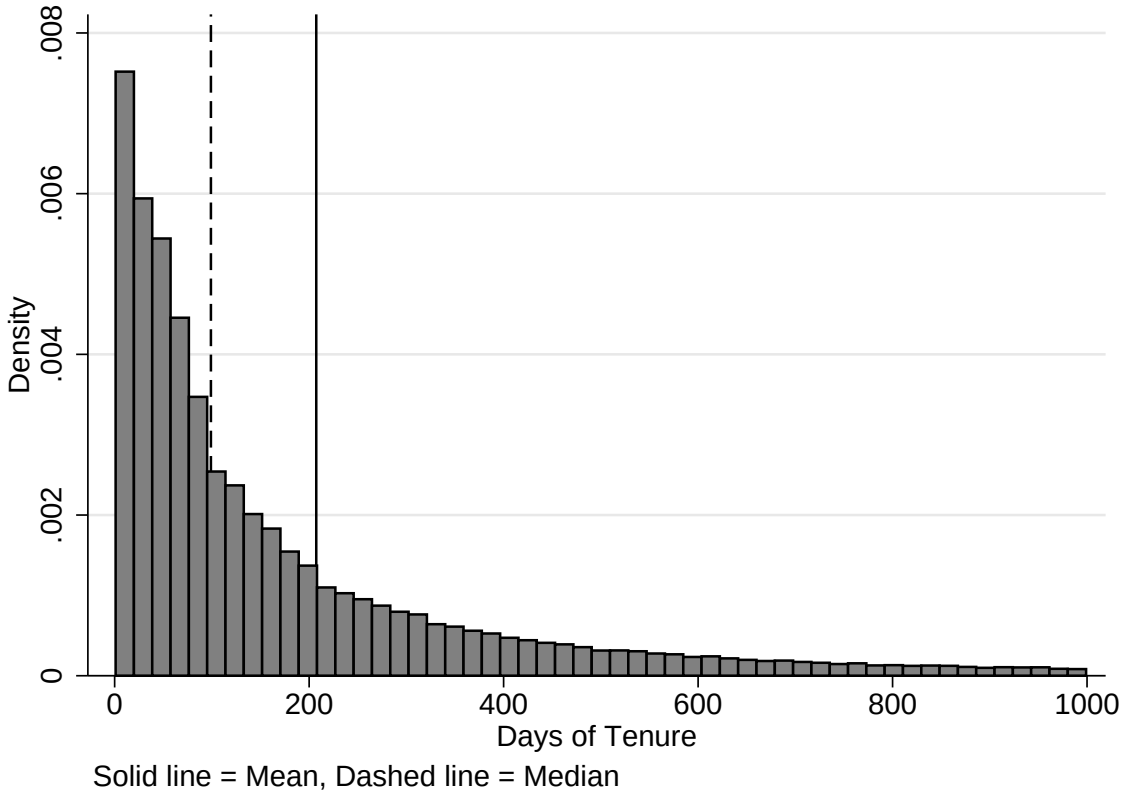
[29] Paravisini, Daniel and Antoinette Schoar, "The Incentive Effect of IT: Randomized Evidence from Credit Committees" NBER Working Paper #19303, 2013.

[30] Riviera, Lauren, "Hiring as Cultural Matching: The Case of Elite Professional Service Firms," *American Sociological Review*, 77 (2014), 999-1022

[31] Stanton, Christopher and Catherine Thomas, "Landing The First Job: The Value of Intermediaries in Online Hiring," *Review of Economic Studies*, 83 (2015), 810-854.

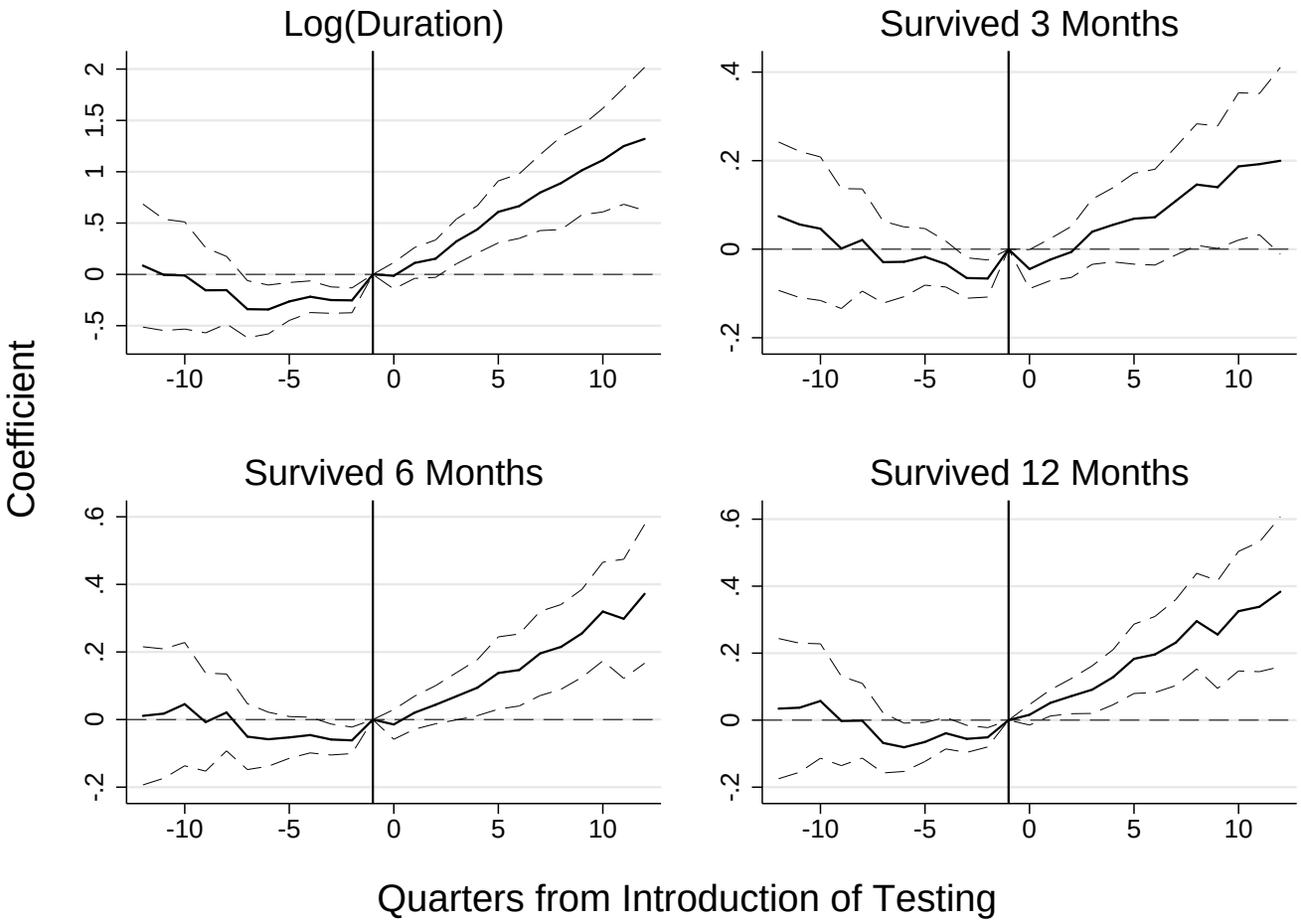
[32] Wang, James, "Why Hire Loan Officers? Examining Delegated Expertise," mimeo University of Michigan, 2014.

FIGURE I: DISTRIBUTION OF LENGTH OF COMPLETED JOB SPELLS



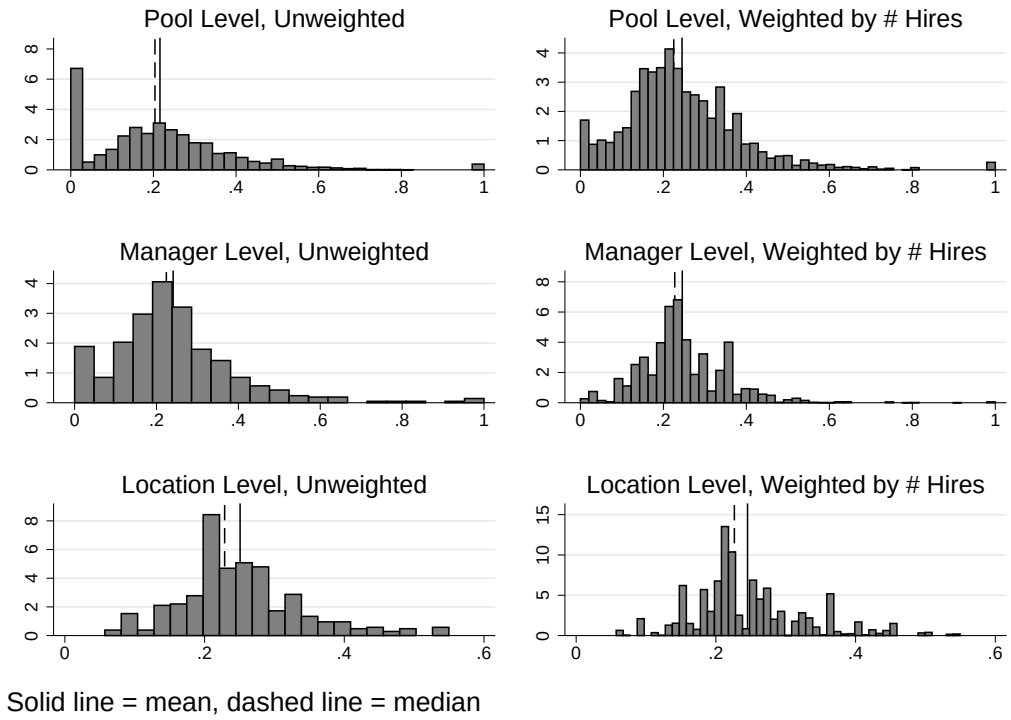
NOTES: Figure I plots the distribution of completed job spells at the individual level. For legibility, this histogram (though not the computed mean or median) omits 3% of observations with durations over 1000 days.

FIGURE II: EVENT STUDY OF DURATION OUTCOMES



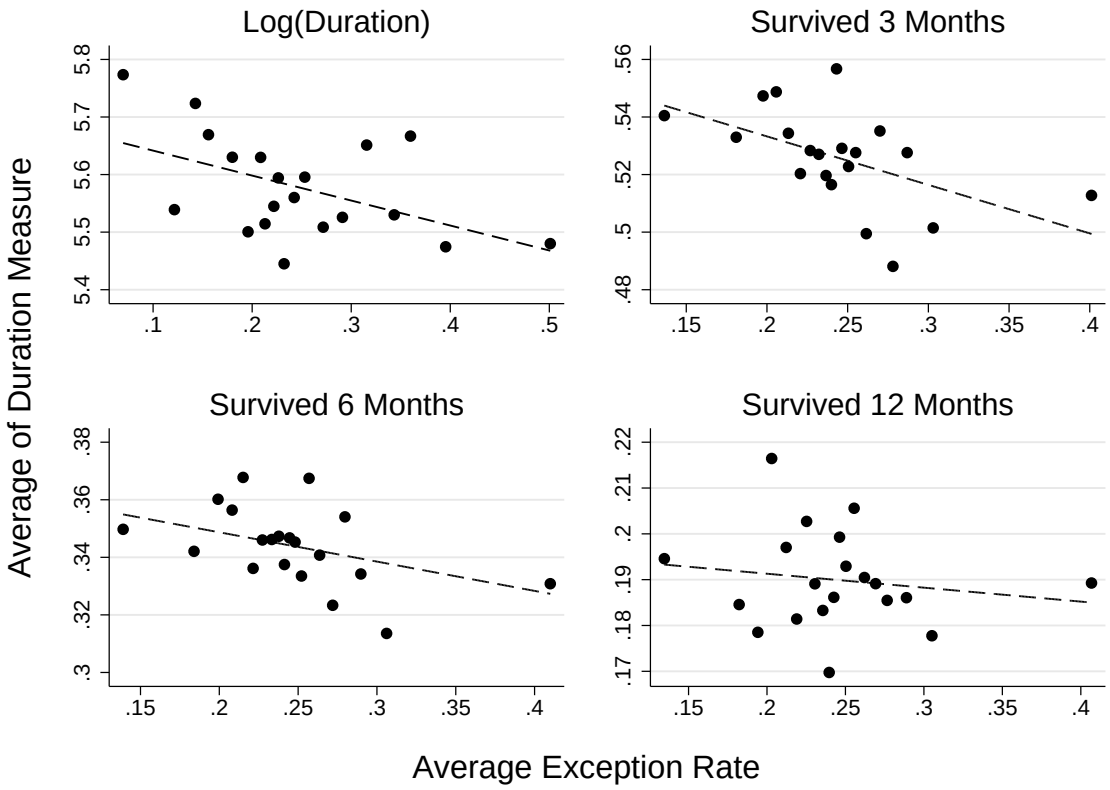
NOTES: These figures plot the impact of testing on worker durations as a function of event-time (in quarters) relative to testing adoption. The underlying estimating equation is given by $\text{Outcome}_{ilt} = \alpha_0 + I_{lt}^{\text{time since testing}}\alpha_1 + controls + \epsilon_{ilt}$, where $I^{\text{time since testing}}$ is a vector of event-time dummies in quarters, with the omitted category, -1, indicated with the vertical line. Controls include location, hire year-month, position, and client-by-year fixed effects, as well as local labor market variables. The top left panel is estimated using censored normal regression while the others are estimated using OLS for the sample of workers hired at least 3, 6, or 12 months before the data end date (measured for each of the 15 firms by the latest termination date or latest hire date in our data, whichever is later). Dashed lines indicate the 95% confidence interval. Online Appendix Figure A3 replicates this figure while restricting to a balanced panel of locations that hire in each of the four quarters before and after testing.

Figure III: Distributions of Application Pool Exception Rates



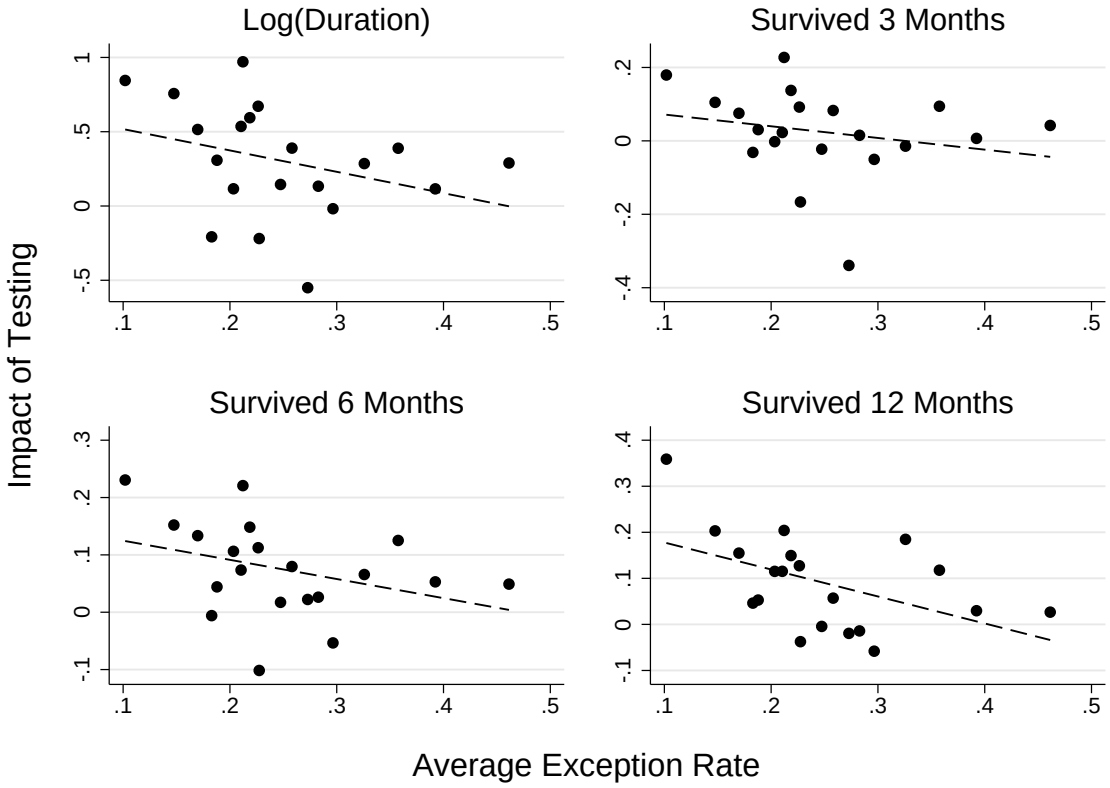
Notes: These figures plot the distribution of the exception rate, as defined by Equation (2) in Section V. The top panel presents results at the applicant pool level (defined to be a manager–location–month). The middle (bottom) panel aggregates these data to the manager (location) level. Figures on the left define the distribution giving applicant pools equal weight while figures on the right weight by number of hires. Exception rates are only defined for the post-testing sample.

Figure IV: Manager-Level Exception Rates and Post-Testing Job Durations



Notes: We relate post-testing job durations to manager-level exception rates across 20 equally sized bins (weighted by number of hires). In the top left panel, we estimate a censored normal regression of log duration on the 20 bins and control variables (location, hire month, and position fixed effects), where the x-axis represents the average exception rate within each bin. On the y-axis, we plot the coefficient estimates on the 20 bins. This is implemented in Stata using "cnreg" (with no constant/intercept term included). The panel also shows the line of best fit for the 20 points. In the top right, lower left, and lower right panels, we plot means of whether workers stay 3, 6, or 12 months as a function of exception rates (controlling for location, hire month and position fixed effects) using a binned scatter plot (using "binscatter" in Stata). In these three panels, we restrict attention to workers hired at least 3, 6, or 12 months before the data end date for each of the 15 client firms, and we also show the line of best fit.

FIGURE V: LOCATION-LEVEL EXCEPTION RATES AND THE IMPACT OF TESTING ON JOB DURATIONS



NOTES: We plot the impact of testing within 20 equally sized bins, based on the location-level exception rate, on the average exception rate in each bin. Estimates include base controls: location, hire month, and position fixed effects. The top left graph is estimated with censored-normal regressions, while the others are estimated using OLS for the sample of workers hired at least 3, 6, or 12 months before before the data end date for each of the 15 firms.

TABLE I: SUMMARY STATISTICS

	Sample Coverage		
	All	Pre-testing	Post-testing
Sample Coverage			
# Locations	127	113	97
# Hired Workers	265,648	174,329	91,319
# Applicants			403,006
# HR Managers			445
# Pools			3,698
# Applicants/Pool			260

	Worker Characteristics *mean (st dev)*				
	Pre-testing	Post-testing	Green	Yellow	Red
Duration of Completed Spell (Days) (N=209,808)	252 *(323)*	116 *(138)*	122 *(143)*	110 *(130)*	92 *(121)*
Duration of Censored Spell (Days) (N=55,840)	807 *(510)*	252 *(245)*	265 *(252)*	235 *(232)*	223 *(223)*
Share Censored	0.19 *(0.39)*	0.25 *(0.43)*	0.24 *(0.43)*	0.26 *(0.44)*	0.25 *(0.43)*
Output per Hour (N=62,427)	8.35 *(4.66)*	8.44 *(5.16)*	8.39 *(5.01)*	8.32 *(5.11)*	9.16 *(6.08)*

	Applicant Pool Characteristics			
	Post-testing	Green	Yellow	Red
Share Applicants		0.48	0.32	0.20
Hire Probability	0.19	0.23	0.18	0.08

NOTES: Post-testing is defined at the location-month level as the first month in which 50% of hires had test scores, and all months thereafter. An applicant pool is defined at the manager-location-month level and includes all applicants that had applied within four months of the current month and not yet hired. Number of applicants reflects the total number of applicants across all pools. Applicant pool characteristics are unweighted averages across pools, and are calculating using individuals with test scores in the post-testing sample.

TABLE II: IMPACT OF JOB TESTING ON JOB DURATIONS

Dependent Variable: Log(Duration)				
	(1)	(2)	(3)	(4)
Post-Testing	0.368***	0.244**	0.248***	0.233***
	(0.120)	*(0.113)*	*(0.0754)*	*(0.0637)*
N	265,648	265,648	265,648	265,648
Year-Month FEs	X	X	X	X
Location FEs	X	X	X	X
Position Type FEs	X	X	X	X
Client Firm X Year FEs		X	X	X
Local Unemployment Controls			X	X
Location Time Trends				X

*** p<0.01, ** p<0.05, * p<0.1

NOTES: We regress log durations on an indicator for testing availability (this equals 1 in the first month in which the modal hire at a location was tested, and in all months thereafter for that location) and the controls indicated. We use censored-normal regressions with individual-specific truncation points (using "cnreg" in Stata) to account for the fact that 21% of hired workers had not yet left their job at the end of our data collection. Standard errors are in parentheses and are clustered at the location level.

Table III: Exception Rates and Post-Testing Duration

	Dependent Variable: Log(Duration)				
	(1)	(2)	(3)	(4)	(5)
Manager Exception Rate	-0.0682** *(0.0346)*	-0.0658** *(0.0321)*	-0.0661** *(0.0322)*	-0.0607** *(0.0292)*	-0.0557** *(0.0283)*
N	91,319	91,319	91,319	91,319	91,319
Year-Month FEs	X	X	X	X	X
Location FEs	X	X	X	X	X
Position Type FEs	X	X	X	X	X
Client Firm X Year FEs		X	X	X	X
Local Unemployment Controls			X	X	X
Location Time Trends				X	X
Applicant Pool Controls					X

*** p<0.01, ** p<0.05, * p<0.1

Notes: This table reports censored normal regressions of manager-level exception rates and tenure outcomes restricted to the post-testing sample. The exception rate is defined as the number of times a yellow is hired above a green plus the number of times a red is hired above a green or yellow, divided by the maximum exceptions possible in that applicant pool. It is then aggregated to the manager level and standardized to be mean zero and standard deviation one. Applicant pool controls include fixed effects for deciles of each of the following variables: number of applicants, hire rate, share of applicants that are green, and share that are yellow. Standard errors are clustered by location.

TABLE IV: THE IMPACT OF TESTING BY EXCEPTION RATE

	Dependent Variable: Log(Duration)			
	(1)	(2)	(3)	(4)
Post-Testing	0.366***	0.234**	0.242***	0.234***
	(0.119)	*(0.107)*	*(0.0696)*	*(0.0589)*
Location Exception Rate Post-Testing*	-0.142**	-0.150**	-0.152**	-0.142**
	(0.0652)	*(0.0679)*	*(0.0655)*	*(0.0573)*
N	265,648	265,648	265,648	265,648
Year-Month FEs	X	X	X	X
Location FEs	X	X	X	X
Position Type FEs	X	X	X	X
Client Firm X Year FEs		X	X	X
Local Unemployment Controls			X	X
Location Time Trends				X

*** p<0.01, ** p<0.05, * p<0.1

NOTES: This table reports censored normal regressions of the differential impact of testing-adoption, by location-level exception rate. We use the same sample as defined by the notes to Tables II. The exception rate is defined as the number of times a yellow is hired above a green plus the number of times a red is hired above a green or yellow, divided by the maximum exceptions possible in that applicant pool. It is then aggregated to the location level and standardized to be mean zero and standard deviation one.

Table V: Tenure of Exceptions vs. Passed Over Applicants

Dependent Variable: Log(Duration)		
	(1)	(2)
Quality of Yellow Exceptions vs. Passed over Greens		
Passed Over Greens	0.0402*	0.0449
	(0.0220)	*(0.0357)*
N	53,166	53,166
Quality of Red Exceptions vs. Passed over Greens and Yellows		
Passed Over Greens	0.159***	0.143**
	(0.0543)	*(0.0634)*
Passed Over Yellows	0.143***	0.121**
	(0.0546)	*(0.0597)*
N	25,782	25,782
Base Controls	X	X
Comparison Pool FEs		X

*** p<0.01, ** p<0.05, * p<0.1

Notes: Regressions are restricted to the post-testing sample, adjust for censoring, and standard errors are clustered at the location level. The top (bottom) panel compares yellow (red) exceptions—the omitted category—to passed over greens (and yellows) who were available at the same time but hired in a later month. Observations are restricted to pools with at least one exception and one passed over worker, and are further restricted to locations and pools with at least 10 and 5 observations, respectively. Base controls are location, hire month, and position type fixed effects. Comparison pool fixed effects are defined by the manager-location-month for the applicant pool in which candidates were considered together.

Table VI: Testing, Exception Rates, and Output Per Hour

Dependent Variable: Output per Hour				
	(1)	(2)	(3)	(4)
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.343	0.156
			(0.366)	*(0.327)*
*Exception Rate*Post-Testing*	-0.0659	-0.111	0.137	-0.137
	(0.134)	*(0.0953)*	*(0.153)*	*(0.217)*
N	28,858	28,858	62,421	62,421
Base Controls	X	X	X	X
Full Controls		X		X

*** p<0.01, ** p<0.05, * p<0.1

Notes: See notes to Tables III and IV. The dependent variable in this case is output per hour and regressions are estimated with OLS. In columns 1 and 2, we examine the post-testing correlation between the manager-level exception rate and output per hour. In columns 3 and 4, we examine the differential impact of testing as a function of location-level exception rates. Base controls include location, hire month, and position fixed effects, as well as controls for the number of tasks in a day that are used to measure a worker's output per hour. Full controls add client-by-year effects, local unemployment rates, and location-specific time trends. For the post-testing sample regressions (columns 1 and 2), full controls also include applicant pool controls.

Discretion in Hiring: Online Appendix

Mitchell Hoffman
University of Toronto & NBER

Lisa B. Kahn
Yale University & NBER

Danielle Li
MIT & NBER

Section A is our Data Appendix. Section B is our Theory Appendix, accompanying Section IV in the main text. Section C contains supplemental tables and figures.

A Data Appendix

The 15 client firms in our sample each obtained testing services from our data provider. In this section, we describe the introduction of testing across locations within these firms, as well as give further information on the data. We first provide some details about the test itself (Section A.1). We then discuss how we assign the date at which testing is introduced to a location and demonstrate the robustness of our main results to this definition (Section A.2). We also describe sample coverage over time across client firms and show robustness to using more balanced panels (Section A.3). We then discuss the timing of test adoption (Section A.4). We further provide a discussion of heterogeneity in test accuracy across locations (Section A.5) and discuss alternative exception rate definitions (Section A.6). Finally, we provide details on sample restrictions (Section A.7), as well as provide additional information about the setting and data set (Section A.8).

A.1 The Job Test

The test is designed to take around 30-60 minutes, though its intended length varies by firm (e.g., according to whether the test covers multiple positions) and consists of several sections. Applicants generally take the test in addition to submitting standard application information (such as a resume). The test includes an introductory section describing the job and work environment, and asks the applicant if he/she thinks they are well-suited for the job and about eligibility. Following this section, there are questions on many dimensions, including those on work experience, computer/technical skills, personality traits, cognitive skills, hypothetical job scenarios, and workplace simulations. The hypothetical job scenarios reflect issues that may arise in performing the specific task we study: for example, if this were a data entry job, it may ask what the employee would do if she were unable to understand the data entry interface. In the workplace simulations, applicants are asked to perform part of the job itself. For example, if this were a data entry job, the applicant may be asked to read an input file and enter the relevant data.

Our data firm uses a proprietary algorithm based on candidates' responses to generate test scores. This algorithm varies somewhat by client firm, but there are commonalities, and the algorithm is updated over time as more data arrives. The algorithm used by a given firm will include data from that particular firm, as well as data from other firms. Correlations are analyzed between various questions and employee attrition (a key outcome), as well as between the various questions and other outcomes (depending on the client firms), particularly output per hour, as well as output quality. In its promotional materials as well

as in its conversations with us, our data provider has stressed the importance of attrition as a key outcome.

The central output of the test is a Red/Yellow/Green score (or scores if the test covers multiple positions) for each applicant. Recruiters observe overall job test scores, but do not observe underlying information on data such as cognitive skills, personality, or how applicants would handle various job scenarios.[1]

A.2 Assigning Testing Adoption Dates to Locations

We observe the date at which test scores appear in our data, but not all workers are tested immediately. Our preferred definition assigns testing to begin at a location when the modal hire in a cohort has a test score. At this point, testing "turns on" for the location for the remainder of our sample period.

Within locations, testing appears to be adopted quickly. Appendix Figure A1 plots the share of hires who are tested as a function of time relative to when the modal hire at that location is tested. This shows that testing ramps up very quickly within a location, reaching roughly 80% coverage almost immediately and continuing to increase to nearly 100% by the end of our sample period.[2] This supports our defining test-adoption as the first month in which the modal hire at a location is tested.

Appendix Table A1 shows that our results are robust to this testing definition. Column 1 replicates our base specifications from Tables II and IV in the main text for the introduction of testing (top panel) and the differential impact of testing across exception rates (bottom panel).[3] These results are very similar when the alternative testing definitions used in Columns 2 and 3. Column 2 defines testing adoption as the first date in which any hire is tested, while column 3 assigns testing at the individual level.

[1]Beyond the central Red/Yellow/Green score (or scores), recruiters observe information on typing speed and accuracy, and, for some firms and time periods, information on an additional job-related skill, but these do not enter into the scoring of Red/Yellow/Green. Results are robust to controlling for typing variables where available, which accounts for the possibility that some locations may have had typing threshold hiring rules. In addition, recruiters could observe information on responses submitted during the introductory section (e.g., whether applicants may have a work schedule issue). Further, recruiters had the option to observe several performance prediction scores that go into the final Red/Yellow/Green score; however, these also represent overall job test scores (as opposed to underlying information on data such as cognitive skills, personality, and job scenarios).

[2]According to the data firm, non-tested individuals are primarily those hired from job fairs. Also, our data contain a small number of non-frontline workers (such as managers and professionals) who are not tested. These workers are distinguished in our position controls. Last, it is possible that testing could be rolled out to hiring for particular end clients within a location (but not for others).

[3]Results from Table III from the main text on the correlation between manager-level exception rates and outcomes of hires do not rely on a comparison of pre and post-testing data so are not included.

A.3 Sample Coverage within Locations over Time

Based on our preferred definition of testing, 97 out of 127 locations receive testing at some point during our sample period; 83 locations are observed both before and after testing. Locations observed only before or after testing are included in our regressions and help identify coefficients on controls. However, Column 4 of Appendix Table A1 shows that our results on the impact of testing are robust to restricting to a balanced panel of locations that are observed both before and after testing.

Appendix Figure A2 provides a summary of our sample coverage over time for all locations. We collect locations by client firm on the y-axis and plot a dot for each month the location hires in, with calendar time indicated on the x-axis. Hollow circles indicate that testing had not yet been introduced to the location, based on our preferred measure; filled in circles indicate the post-testing period. A gap between circles indicates no hires were made in that month.

This figure indicates that we observe cohorts of workers for many periods both before and after testing for most locations. Specifically, among the 83 locations that hire both before and after testing, the average observation window post-testing is 15 months and the average pre-testing observation window is 3.5 years (worker weighted). Furthermore, 90% of hires in this sample are to a location that can be observed for at least 6 months before and after testing, 60% are hired to locations with at least a 1 year window around testing. Of course, the panel is highly unbalanced and there is a range of observation windows for clients and locations.[4]

From the figure, locations also appear to hire in most months during their observation window. In fact, of the locations that can be observed for at least a full year before and after testing, three-quarters (worker weighted) hire in every single quarter in that window. Column 5 of Appendix Table A1 shows that results on the impact of testing are robust to restricting to this very balanced panel of locations. Results are similar across a wide range of balanced panels.

Furthermore, Appendix Figure A3, replicates the event study for the impact of testing, restricting to locations that hire in each of the four quarters before and after testing, and shows a very similar picture to Figure II of the main text.

Finally, as noted in the main text, the data firm informed us that a number of client firms had some other form of testing before the introduction of our data firm's test. While information about whether a client firm had testing before our data provider is not part of our dataset, we asked our data provider to collect information about this on our behalf by surveying managers and executives at the data firm. From this, the data firm reported

[4]For example, client #13 has no pre-testing data.

that 5 firms had pre-sample testing (and not just in one part of its business), 1 firm had pre-sample testing in one part of its business, 1 firm was believed to have pre-sample testing (but our data firm was not certain), and 8 firms were regarded as either not having testing or believed not to have testing.

This survey does not provide certainty for all 15 client firms in our data. However, column 6 of Appendix Table A1 shows that key coefficients are larger on the sample of firms who likely did not have pre-sample testing. This is consistent with testing being more of an improvement for firms that had no alternative test in the pre-period, as well as it being more useful for managers to follow test recommendations rather than make exceptions at these firms.

A.4 Timing of Testing and Location Observables

Appendix Figure A4 describes how testing enters our sample across both client firms and locations. Circles indicate the date at which testing is adopted for the 97 locations that ever receive testing during our sample (x-axis). Locations are collected by client firm and lined up on the y-axis in the order of their specific test adoption date. The size of each circle reflects the location's size.[5] Among client firms with more than one location (11 out of 15 firms, accounting for 94% of hires in our data), most firms adopt testing across all their locations in our sample in under 2 years. There does not appear to be a systematic relationship between the size of a location and the time at which it receives testing.

In Section III of the main paper, we exploited this gradual roll-out of testing across locations within client firms to estimate the impact of testing on job durations, while controlling for location and hire date fixed effects. Naturally, one may be concerned about factors leading clients to introduce testing in some locations before others. However, based on qualitative and quantitative information, we see no evidence that the timing of this roll out would bias our results.

On the qualitative side, we had discussions involving different individuals from our data provider (including one person who worked closely with different firms on rolling out testing), as well as managers from a large client firm in our dataset. Representatives mentioned several possible drivers of testing adoption, including the availability/"bandwith" of managers to oversee the adoption of testing, considerations of geography, the openness of end clients (i.e., the ones paying for the services provided by our client firms) to testing, and whether a location had historically high attrition. Importantly, representatives did not say that

[5] We define the size of the location as the number of workers currently employed in July 2013. For one location we must use July 2012 instead. This snapshot date avoids overweighting locations that have high churn.

firms may have adopted testing in ways that reflect time-varying differences in a location's attrition risk. For example, no one mentioned bringing in testing to a location that was recently experiencing or expecting a retention problem.

On the quantitative side, we have examined the correlation between location-level observables and the timing of testing adoption. For example, Appendix Figure A5 plots location characteristics as a function of testing adoption date for several key variables. Circles and the fitted regression line are again weighted by location size, and durations are censoring adjusted.

The top panels show relationships for pre-testing characteristics at the location level. In the top left panel, we find no systematic relationship between a location's average pre-testing duration (censoring adjusted) and the date at which it adopts testing. The top middle panel considers a location-specific time trend in censoring-adjusted durations pre-testing.[6] This gradient is also quite flat: testing does not arrive earlier or later for locations that are on a stronger or weaker trend in worker duration. Finally, the top left panel plots the average unemployment rate among workers with exactly a High School Diploma pre-testing. Here, there is again no relationship between the testing date and local labor market conditions. We use the state-level unemployment rate for high school graduates (a high school diploma is a typical educational requirement for many low-skill service jobs), but the graph looks similar for unemployment rates for other groups.[7]

The bottom panel of Appendix Figure A5 focuses on variables that are available only after testing: the share of applicants with a green test score, the average number of applicants per month, and the average exception rate across HR managers at that location (see Equation 2). Again, we do not find a clear pattern for these dependent variables.

We also point out that the linear relationships in these graphs tend to be statistically insignificant and small in magnitude. For example, we can rule out a plus or minus 1.5% change in pre-testing average durations with each month that testing is delayed with 95% confidence. We can similarly rule out a plus or minus 0.2% change in the share of applicants that are green. We can also rule out a plus or minus 0.004 variation in the location exception rate. We have examined a wide range of location characteristics and similarly find little systematic or robust relationships with timing of testing. Notably, these include pre-testing averages for the share of months that the location is active in hiring and the location-specific churn rate.

[6]Specifically, we estimate a censored normal regression of job durations on location fixed effects and location-specific time trends for the pre-testing sample.

[7]The graph also looks similar when using aggregated unemployment rates for the 25% of international locations and when using U.S.-level unemployment rates for each education group for the non-standard location identifiers where state cannot be easily assigned.

A.5 The Accuracy of Test Scores Across Locations

One may be worried that the test does not predict worker quality equally well across locations. For example, worse establishments may be especially undesirable for more skilled workers, resulting in lower durations among greens.

Appendix Figure A6 plots the relationship between manager-level exception rates and worker duration, separately by color. We estimate censored normal regressions of log duration on indicators for each of 20 equally sized (based on number of hires) exception rate bins and base controls (location, hire month, and position fixed effects, and with no constant/intercept term included), separately by color. There are two main patterns to notice. First, greens perform better than yellows who in turn perform better than reds across exception rate bins. Second, the overall quality of hired yellows and greens is broadly stable across exception rates. This means that, among workers a manager is able to hire, color score is predictive of performance, even across varying manager exception rates.

We do see some evidence that reds hired by managers who make many exceptions appear worse than reds hired by managers who make few exceptions. This could be because reds in these locations are worse or because managers with high exception rates are especially bad at picking out reds. In either case, this reinforces the point that, in high exception locations, managers may do better by hiring more greens and yellows, relative to reds: the greens and yellows they are able to hire are broadly comparable to the quality of greens and yellows in low exception locations, while the reds they hire appear somewhat worse.

Appendix Figure A7 provides more information along these lines. We plot the relationship between color score and job duration as a function of the same set of location-level characteristics reported above in Appendix Figure A5. We divide locations into 20 equally sized bins (based on number of hires post-testing).[8]

Across the bins for the panels in Appendix Figure A7, we find that color score is predictive of job durations. For example, the top left panel plots the relationship for the average duration of the location pre-testing and shows three upward sloping parallel lines. This means that average job durations are increasing in the average quality of the location pre-testing, naturally. However, the gap between job durations by color is roughly constant across locations. The other panels suggest a similar conclusion: we fail to see systematic evidence that the predictiveness of the color score varies much with location characteristics.[9]

[8] We estimate these figures in a similar manner as Appendix Figure A6, with 20 bins for location-level characteristics instead of manager-level exception rates (and we exclude location fixed effects because they are collinear with the location characteristics).

[9] For each subfigure in Figure A7 (as well as for Figure A6), if the 3 regressions are run without control variables, or if the 3 regressions are run instead as one regression with color x ventile interactions and control variables, the differences in duration between the 3 colors become smaller. However, the same qualitative

A.6 Alternative exception rate definitions

In Appendix Table A2, we examine the robustness of our main results in Tables III and IV from the main text to alternative ways of defining an exception rate. Recall that we construct our exception rate by counting the number of order violations (the number of greens that are passed over by each hired yellow, plus the number of greens and yellows that are passed over by each hired red) and normalizing by the maximum number possible, given the same color composition of applicants and total number of hires.

First, we consider an alternative normalization: the number of order violations that would occur if managers hired at random. The random benchmark is interesting because this is the number of exceptions that would occur if managers ignored the test and the test were uninformative for quality. In our data, 86% of workers are hired from application pools in which this exception rate is less than 1, indicating that, in the vast majority of pools, managers' decisions align with test recommendations to some extent. Next, we consider a different way of conceptualizing the exception rate, using the idea of a "score" rather than a violation: 2 points for every green hire, 1 point for every yellow hire, and no points for red hires. We count up scores per applicant pool and normalize by either the maximum possible score, or the score that would obtain under random hiring. The score measure differs conceptually from the order violation approach because it is less sensitive to the number of unhired applicants. For example, the score is the same if a single yellow worker is hired over 20 greens, or over only one green.[10] We negate the score metrics so that a larger number means more exceptions, to align with the order violation measure. All three of these measures are aggregated to the manager and location-levels and then standardized. Appendix Table A2 shows that all of these metrics tell similar stories. Results are robust quantitatively and generally in terms of statistical significance as well.

A.7 Sample Restrictions

For the post-testing period, we make the following restrictions:

1. We drop roughly one third of applicants because they have a missing identifier for their HR manager.[11]

message remains that color is predictive of duration, and this holds across varying levels of manager exception rates or location characteristics.

[10]The maximum score for one hire is 2 in both cases, but the random score will differ.

[11]To assess the possibility of selection bias, we regressed whether HR manager is missing on duration (or log duration), a dummy for being censored, location controls, month-year of hire dummies, and position dummies, using the full sample of tested hires. In the two regressions, the coefficients on duration and log duration are statistically insignificant, suggesting that selection bias is not a main concern for our analysis. In forming the applicant pool, we also drop about 200 people (almost all from one firm) where there appears

2. We drop 2% of hires that are part of pools with less than 3 applicants.

3. We drop locations that do not have at least two managers because part of our exception rate analysis (Equation (3)) relies on within-location variation in manager-level exception rates. This drops 2% of remaining managers associated with 0.9% of remaining hires.

4. We drop pools that hire only exceptions because we worry that an idiosyncratic shock drives the lack of matriculation of higher scoring applicants. This reflects 8% of the remaining pools associated with 0.6% of remaining hires.

5. We drop managers that hire in only 1 pool to clean out some noise in the manager-level exception rates. This reflects 16% of the remaining managers associated with 0.55% of remaining hires.

6. We drop observations with missing manager-level exception rates, which occur when all pools a manager hires to have a value of 0 for the maximum number of possible exceptions. This reflects 1.5% of the remaining pools associated with 0.06% of remaining hires.

We implement these restrictions for the post-testing period in all analyses, even those that do not use exception rates, to keep the sample consistent. However, results from Section III on the impact of testing (which do not use exception rates) are similar without the restrictions. We do not exclude observations in the pre-testing period on the basis of being associated with locations that do not meet our post-testing criteria, as these observations help identify cohort, client, and position controls. Further, for all analyses, we drop the four locations (reflecting 0.04% of remaining hires) with less than 50 hires over the sample period. Section A.8 below describes a few further sample restrictions.

A.8 Further Information on Setting and Data

Firms in the Data. The data were assembled for us by the data firm from records of the individual client firms. The client firms in our sample employ workers who are engaged in the same job, but there are some differences across the firms along various dimensions. For example, at one firm, workers engage in a relatively high-skilled version of the job we study.[12]

to be an error in the application date; also, for about 1,500 people, we fill in missing application month using the hire month minus one.

[12]As such, the work performed at this firm is fairly different compared to our other firms.

At a second firm, the data firm provides assistance with recruiting (beyond providing the job test). Our baseline key results are similar when individual firms are excluded one by one.[13]

Pre-testing Data. In the pre-testing data at some client firms there is information not only on new hires, but also on incumbent workers. This may generate a survivor bias for incumbent workers, relative to new workers. For example, consider a hypothetical firm that provided pre-testing data on new hires going back to Jan. 2010. For this firm, for workers hired before Jan. 2010, we would only observe the subset of workers who survived to a later date. We do not explicitly observe the date at which the firm began providing information on new hires; instead, we conservatively proxy this date using the date of first recorded termination. We label all workers hired before this date as "stock sampled" because we cannot be sure that we observe their full entry cohort. We drop these workers from our primary sample, but have experimented with including them along with flexible controls for being stock sampled in our regressions. In forming our regression sample, we also drop a couple thousand hired workers who have a missing duration variable in the data, most of whom are from the pre-testing period.

Productivity. In addition to the information on job durations, some client firms provide data on output per hour. This is available for about a quarter of hired workers in our sample. We trim instances where average transaction time in a given day is less than 60 seconds.[14]

Test Scores. As described in the text, applicants are scored as Red, Yellow, or Green. Applicants may receive multiple scores (e.g., if they are being considered for multiple roles). In these cases, we assign applicants to the maximum of their scores.[15]

For candidates in our data with at least one Red/Yellow/Green score, roughly one third have one score, roughly half have two scores, and the remainder have more than two scores in our data. Among candidates with multiple scores, the scores are very highly correlated with one another. For example, scores for the two most common positions have a correlation

[13]Specifically, we estimated base specifications of Tables 2, 3, and 4 from the main text excluding each firm one by one.

[14]This is about one percent of transactions. Some other productivity variables are also shared with our data provider, but each variable is only available for an even smaller share of workers than is output per hour. Such variables would likely face significant statistical power issues if subjected to the analyses in the paper (which involve clustering standard errors at the location level).

[15]For 1 of the 15 client firms, the Red/Yellow/Green score is missing for non-hired applicants in the dataset provided for this project. Our conclusions are substantively unchanged if that firm is removed from the data. For another 1 of the 15 client firms, we fill in about 500 missing observations using a constructed test score variable from the data firm that exists primarily for hires. Our conclusions are also substantively unchanged if these observations are omitted.

coefficient of 0.88 (for Red=0, Yellow=1, Green=2).[16] Our focus on the maximum of scores thus seems without much loss of generality.[17]

HR Manager. The HR managers we study are referred to as recruiters by our data provider. We do not have data for this project on the characteristics of HR managers (we only see an individual identifier).

Other managers may take part in hiring decisions as well. As noted in footnote 6 of the main text, one firm said that its HR managers will typically endorse candidates to an additional manager (e.g., a manager in operations one rank above the frontline supervisor) to make a "final call." That said, HR managers play a critical role in deciding who gets hired. For low-skilled jobs of the type we study, past work suggests that HR managers play an active role in hiring; for example, in a detailed study by sociologists of a call-center at a bank, Fernandez, Castilla, and Moore (2000) report that HR managers played an important role in the recruiting process, even though there was a second interview that was done by line managers during their study period. In fact, the importance of HR managers at this particular firm happened to grow after the study: HR managers were granted authority to make hiring decisions on their own.

Also, applicants may interact with more than one HR manager during the recruitment process. In such cases, we assign an applicant to the HR manager with whom they have the most interactions.[18] Most managers are primarily associated with one location, but some are associated with multiple locations.

We do not observe manager incentives in our data. However, a manager from our data provider informed us that recruiters in our setting often receive a financial incentive to meet or exceed several targets (while pointing out that such pay structures are highly variable by firm). He said that recruiters always have targets with respect to fill rate (e.g., a requisition of 20 new hires to begin work on March 1st), and often have targets with respect to short-term tenure (e.g., a certain share of people graduating training, or of staying some length of time, such as 90 days) or activities (e.g., conducting X interviews or reaching out to Y candidates).

Race, Gender, Age. Data on race, sex, and age are not available for this project. However, Autor and Scarborough (2008) show that job testing does not seem to affect worker

[16] This is the correlation in the raw data before imposing data restrictions.

[17] Applicants may be considered for multiple positions and it is difficult to discern which is the most relevant score (or scores) for a given applicant.

[18] This excludes interactions where information on the HR manager is missing. If there is a tie for most interactions, we assign an applicant to only one manager. Our main results are also qualitatively robust to setting the HR manager identifier to missing in cases of ties for most interactions.

race, suggesting that changes in worker demographics such as race are not the mechanism by which job testing improves durations.

Location Identifiers. In our dataset, we do not have a common identifier for workplace location for workers hired in the pre-testing period and applicants applying post-testing. Consequently, we develop a crosswalk between anonymized location names (used for workers in the pre-testing period) and the location IDs in the post-testing period. We drop workers from our sample where the merge did not yield a clean location variable.[19]

As explained in the main text, there are a small number of non-standard location identifiers (e.g., those where workers generally work off-site in different states). We assign these locations to unemployment rates using education-specific, US national unemployment data. We do so even though a small share of workers associated with non-standard location identifiers may be outside the US.

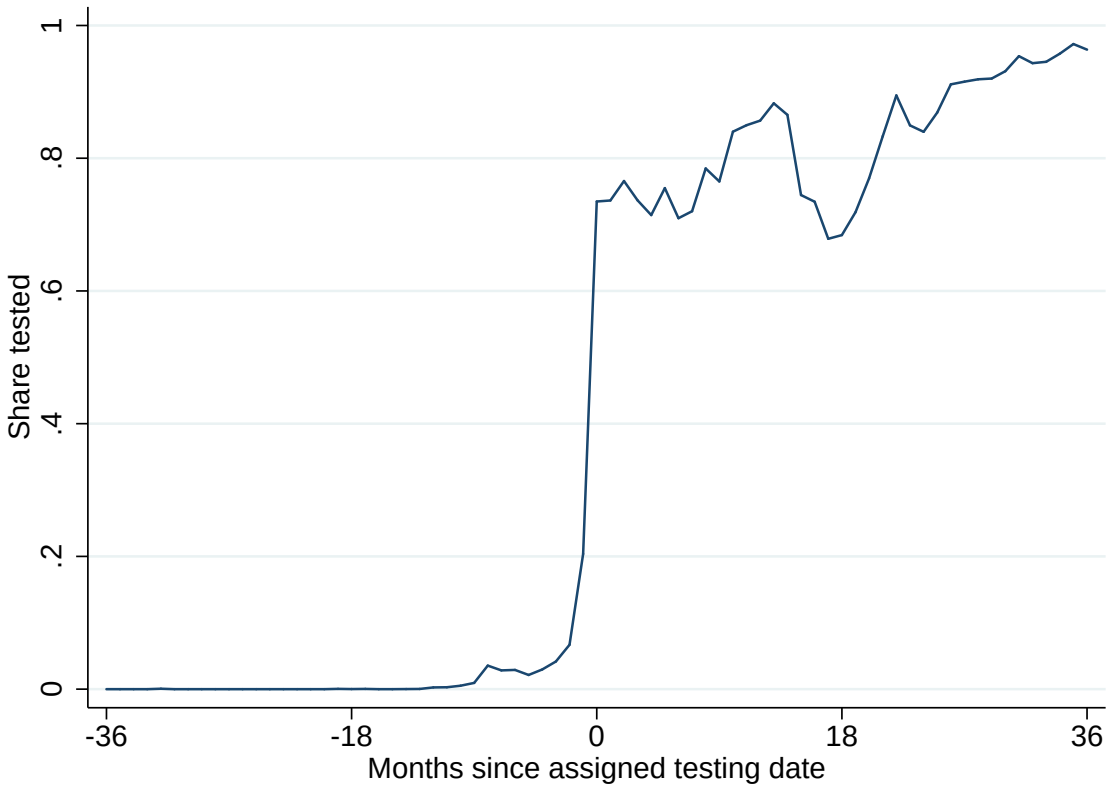
Hiring Practices Information. For several client firms, our data firm surveyed its account managers (who interact closely with the client firms regarding job testing matters), asking them to provide us with information on hiring practices once testing was adopted. The survey indicated that firms encouraged managers to hire workers with higher scores (and some firms had policies on not hiring low-scored candidates), but left substantial leeway for managers to overrule testing recommendations. Information from this survey is referenced in footnote 7 of the main text.

Job Offers. As discussed in the main text, our data for this project do not include information on the receipt of job offers, only on realized job matches. The data firm has a small amount of information on offers received, but is only available for a few firms and a small share of the total applicants in our sample, so it would likely be of little use for this project.

Position Controls. Position is measured in our data using the last position that a worker held when the data file was created.

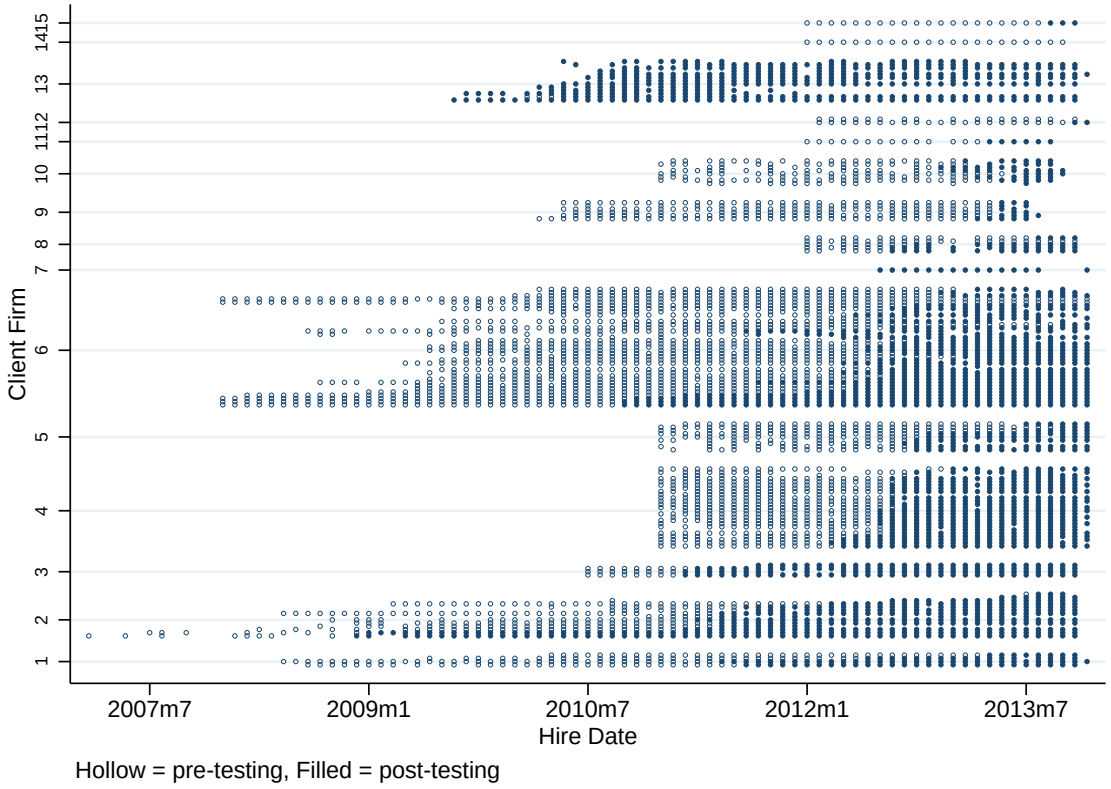
[19] This includes some locations in the pre-testing data where testing is never later introduced. Our sample is not all locations within the firms.

Appendix Figure A1: Share of hired workers tested by time since assigned test-adoption date



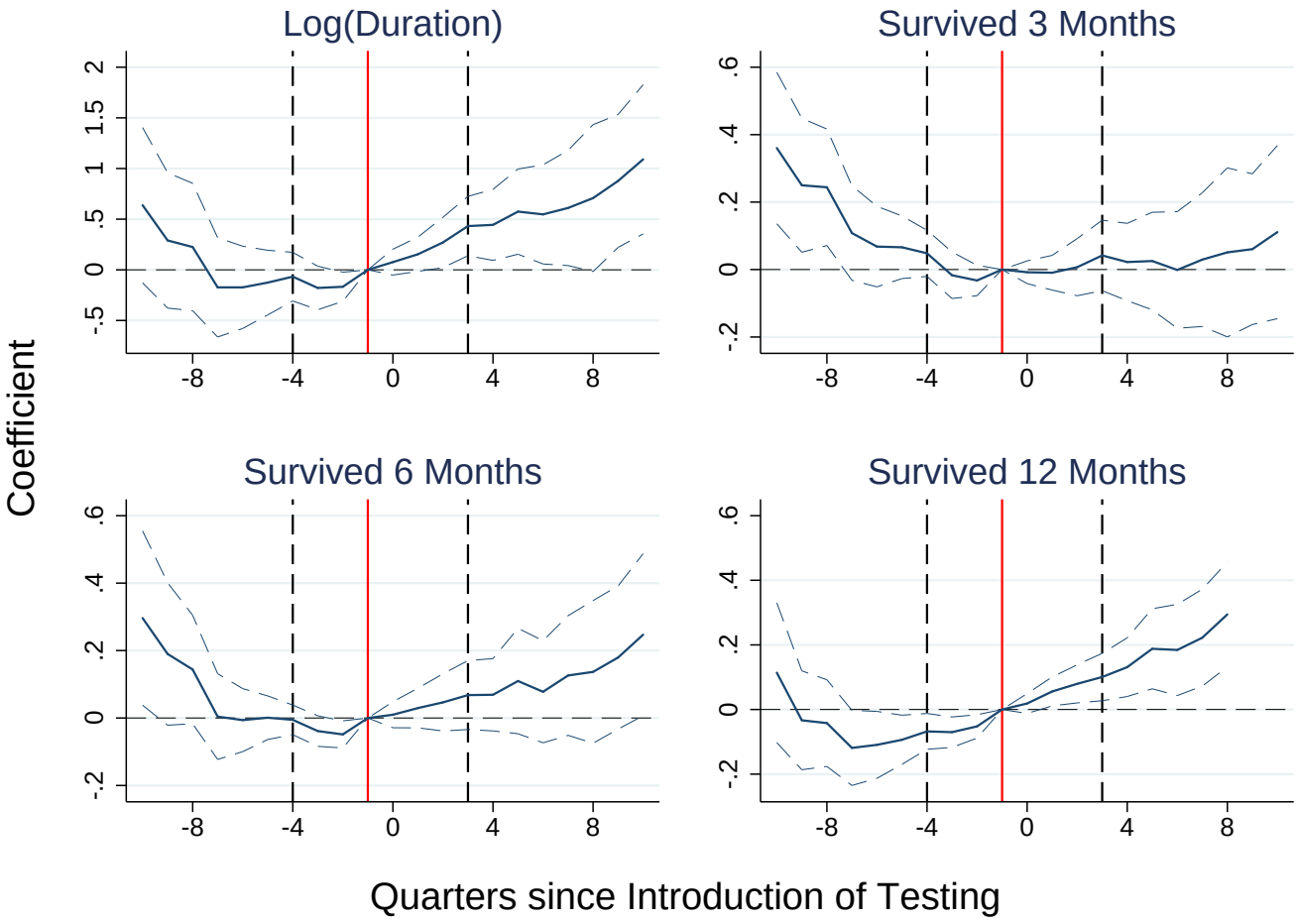
Notes: Figure A1 plots the share of hired workers with a test score as a function of time since the location-specific assigned testing date, averaged across locations. The testing date is defined at the location-month level as the first month in which the modal hire is tested. This graph is restricted to locations that receive testing. For figure clarity, we further restrict to the 89% of workers hired within 3 years of the introduction of testing.

Appendix Figure A2: Location Coverage by Date



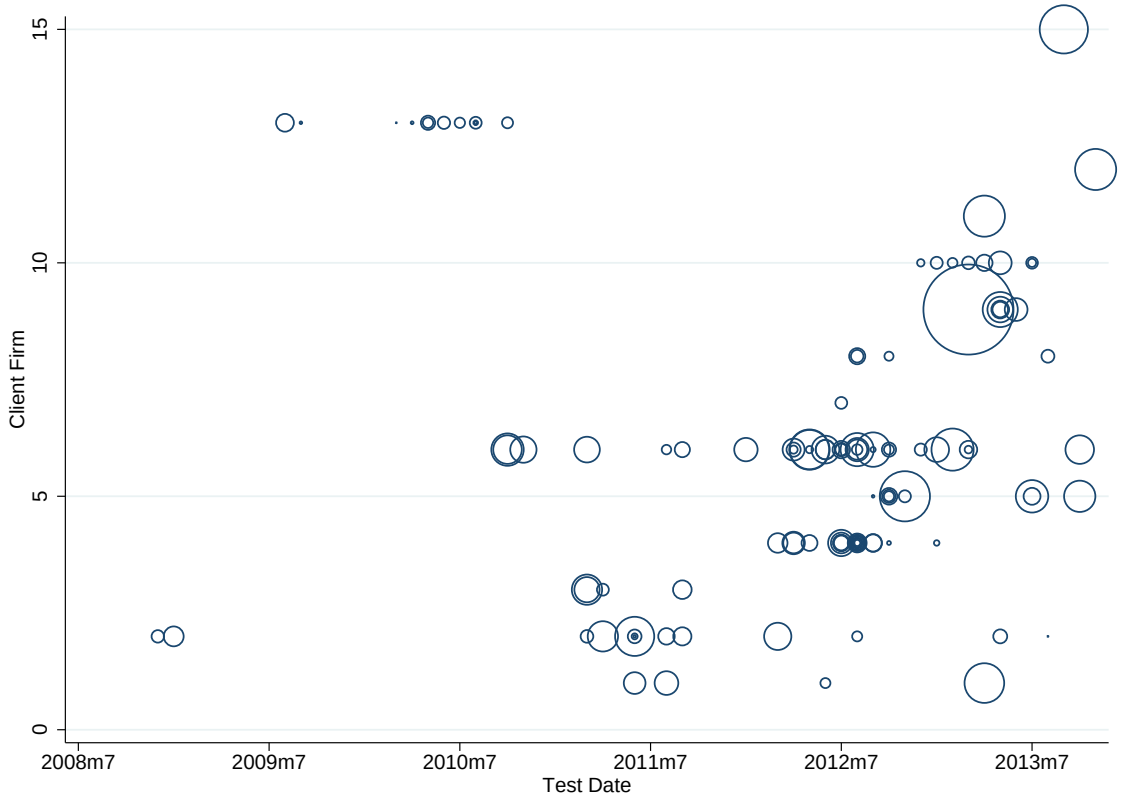
Notes: Locations are lined up on the y-axis, grouped by client firm. Dots indicate that the location hired in a given month, while a gap means no hires were made that month. Filled circles refer to periods after testing is adopted, using our definition (the modal hire was tested), while hollow circles refer to periods before testing. Dates are restricted to a 3 year window around testing adoption, covering 89% of hires. All dots are hollow for Firm 14 because it does not have a location meeting our definition of testing.

Appendix Figure A3: Event Study of Duration Outcomes, Balanced Panel



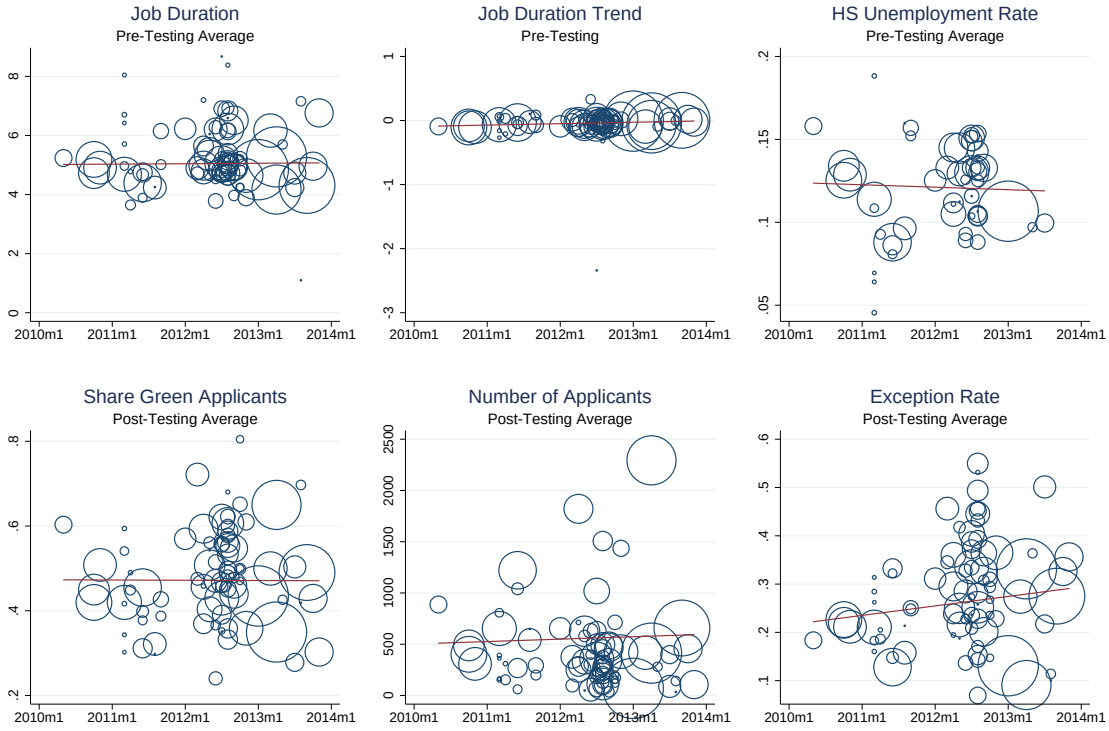
Notes: See notes to Figure II of the main text. The sample is restricted to locations with observations in each quarter from 4 lags before testing to 4 leads after (indicated with vertical dashed lines). The graph window is restricted to 10 quarters before and after testing. Dashed lines indicate the 95% confidence interval.

Appendix Figure A4: Date of Location Testing Adoption, by Client Firm



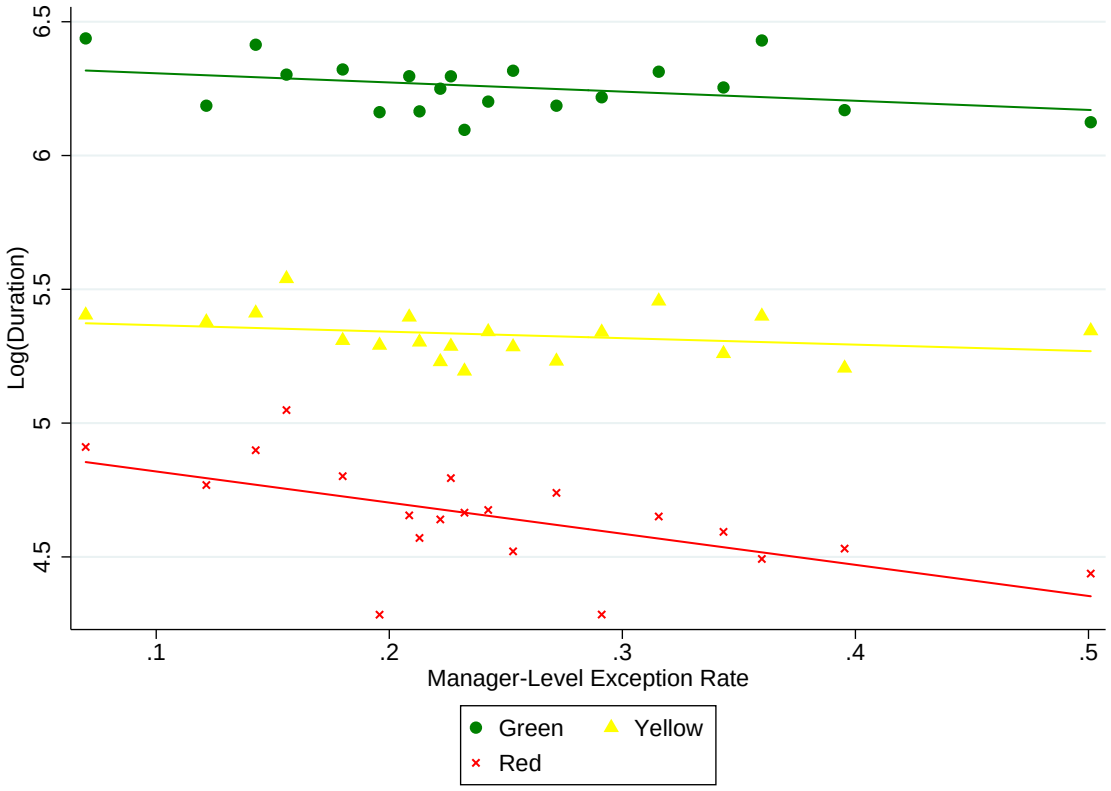
Notes: Figure A4 plots location-specific assigned testing dates on the x-axis, organized by client firm on the y-axis. Circles are weighted by location size, as defined by the number of workers currently employed in our data on July, 2013. As noted in Figure A2, Firm 14 does not appear on the graph because it does not have a location that meets our definition of testing.

Appendix Figure A5: Location Observables and Date of Testing Adoption



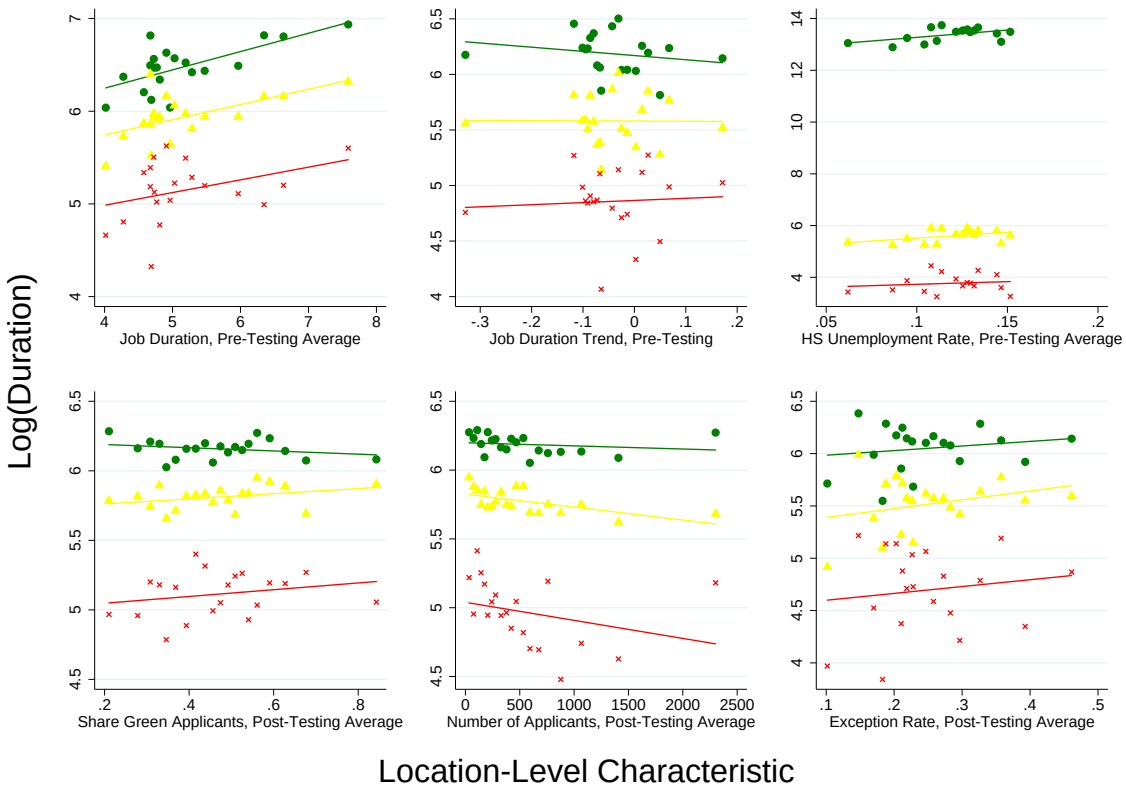
Notes: Figure A5 plots the relationship between various location-level variables (y-axis) and date of test adoption (x-axis). Circles and fitted lines are weighted by location size. In the top left panel, pre-testing durations are obtained from a censored normal regression of log durations on an exhaustive set of location fixed effects estimated on the pre-testing sample (and with no constant/intercept term included). The top middle panel plots location-specific time trends estimated from a censored normal regression of log durations on location fixed effects and location-specific time trends in the pre-testing sample (and with no constant/intercept term included). The remaining variables are raw averages at the location-level either pre- (top right) or post- (bottom panels) testing.

Appendix Figure A6: Manager-Level Exception Rates and the Color Score-Job Duration Relationship



Notes: This graph shows the relationship between color score and job duration for 20 equally sized manager-level exception rate bins. Specifically, we estimate censored normal regressions of log duration on 20 exhaustive indicators for exception rate bin and location, hire month, and position fixed effects (and with no constant/intercept term included), separately by color score. We plot the coefficients on the exception rate bins as well as the line of best fit.

Appendix Figure A7: Location Observables and The Color Score-Job Duration Relationship



Notes: This graph shows the relationship between color score and job duration for 20 equally sized bins based on the location-level characteristic specified on the x-axis. Specifically, we estimate censored normal regressions of log duration on 20 exhaustive indicators for the location characteristic bin and hire month and position fixed effects (and with no constant/intercept term included), separately by color score. (We exclude location fixed effects from these regressions because they are collinear with the location characteristics.) We plot the coefficients on the bins as well as the best linear fit.

Appendix Table A1: Robustness for Results on the Impact of Testing

	Dependent Variable: Log(Duration)					
	(1)	(2)	(3)	(4)	(5)	(6)
Impact of Testing						
Post-Testing	0.368***	0.316***	0.316***	0.296**	0.261**	0.516**
	(0.120)	*(0.121)*	*(0.119)*	*(0.150)*	*(0.117)*	*(0.245)*
Differential Impact of Testing by Exception Rates						
Post-Testing	0.366***	0.321***	0.320***	0.299**	0.294***	0.495***
	(0.119)	*(0.120)*	*(0.119)*	*(0.151)*	*(0.110)*	*(0.161)*
*Exception Rate*Post-Testing*	-0.142**	-0.151**	-0.146**	-0.127*	-0.178***	-0.444**
	(0.0652)	*(0.0696)*	*(0.0687)*	*(0.0663)*	*(0.0556)*	*(0.202)*
N	265,648	265,648	265,648	216,676	96,273	83,910
Base Controls	X	X	X	X	X	X
Testing Definition:						
Modal Worker Tested	X			X	X	X
Any Worker Tested		X				
Individual Worker Tested			X			
Location Restrictions:						
Observed Both Pre/Post Testing				X	X	
Observed in Balanced 4 Quarter Window					X	
Client Had No Pre-Sample Testing						X

*** p<0.01, ** p<0.05, * p<0.1

Notes: This table reports censored normal regressions with standard errors clustered at the location level. The top panel provides estimates of the impact of testing on log durations (see Table II), while the bottom panel estimates the differential impact of testing by location-level exception rates (see Table IV). Column 1 reproduces baseline specifications from the preceding tables. Column 2 defines the test adoption date as the first time a hire is observed with a test score at a location. Column 3 defines test adoption as whether the individual hire has a test score. Column 4 restricts to the 83 locations that are observed both before and after testing. Column 5 further restricts to locations that are observed in each of the four quarters prior and post testing. Column 6 restricts to locations that likely did not have job testing before partnering with our data firm. Base controls include location, hire month, and position fixed effects.

Appendix Table A2: Robustness to Alternative Exception Rates

Dependent Variable: Log(Duration)				
	(1)	(2)	(3)	(4)
# Exceptions Relative to Random				
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.359***	0.232***
			(0.119)	*(0.0588)*
*Exception Rate*Post-Testing*	-0.0730**	-0.0635**	-0.157**	-0.125**
	(0.0327)	*(0.0258)*	*(0.0713)*	*(0.0556)*
Exception Score Relative to Max Score				
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.356***	0.230***
			(0.120)	*(0.0656)*
*Exception Rate*Post-Testing*	-0.0237	-0.0707***	-0.190**	0.0420
	(0.0261)	*(0.0190)*	*(0.0933)*	*(0.0676)*
Exception Score Relative to Random Score				
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.353***	0.222***
			(0.117)	*(0.0609)*
*Exception Rate*Post-Testing*	-0.0585	-0.0149	-0.155**	-0.0999**
	(0.0364)	*(0.0241)*	*(0.0763)*	*(0.0498)*
N	91,319	91,319	265,648	265,648
Base Controls	X	X	X	X
Full Controls		X		X

*** p<0.01, ** p<0.05, * p<0.1

Notes: Columns 1 and 2 estimate the post-testing correlation between manager-level exception rates and log duration (see Table III). Columns 3 and 4 estimate the differential impact of testing by location-level exception rates (see Table IV). The top panel defines the exception rate as the number of order violations divided by the number of order violations under random hiring. The next panels use an exception score (1 point for yellow and 2 points for green hires) divided by the maximum possible score (middle panel) or the score under random hiring (bottom panel). Base controls include location, hire month, and position fixed effects. Full controls add client-by-year effects, local unemployment rates, and location-specific time trends (and applicant pool controls in columns 1 and 2).

B Theory Appendix

B.1 Preliminaries

We first provide more detail on the firm's hiring problem, to help with the proofs that follow.

Under Discretion, the manager hires all workers for whom $E[U_i|t_i, s_i, b_i] = (1-k)E[a|s_i, t_i] + kb_i > \underline{u}$ where $\underline{u}$ is chosen so that the total hire rate is fixed at W.

We assume b_i is perfectly observable, that $a_i|t_i \sim N(\mu_t, \sigma_a^2)$, and that $s_i = a_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and is independent of a_i, b_i, and t_i.

Define $U_t \equiv E[U_i|t_i, s_i, b_i]|t_i$. Based on standard projection formulas for the Gaussian distribution, we know that $U_t \sim N((1-k)\mu_t, \Sigma)$, where $\Sigma = (1-k)^2\sigma^2 + k^2\sigma_b^2$ and $\sigma^2 = \frac{\sigma_a^4}{\sigma_a^2+\sigma_\epsilon^2}$.

Let Φ and ϕ denote the standard normal cdf and pdf, respectively. Therefore, the hire probability and $\underline{u}$ are pinned down by equation B1:

$$W = p_G(1 - \Phi(z_G)) + (1 - p_G)(1 - \Phi(z_Y)) \tag{B1}$$

where $z_t = \frac{\underline{u}-(1-k)\mu_t}{\sqrt{\Sigma}}$.

The firm's payoff under Discretion is $E[a|Hire]$, i.e., the expected quality conditional on being hired. By using the properties of the bivariate normal distribution, this can be expressed as follows, where $\lambda(\cdot)$ is the inverse Mills ratio of the standard normal.

$$\begin{aligned} W * E[a|Hire] = {} & p_G(1 - \Phi(z_G)) \left[\mu_G + \frac{(1-k)\sigma^2}{\sqrt{\Sigma}}\lambda(z_G)\right] \\ & (1 - p_G)(1 - \Phi(z_Y)) \left[\mu_Y + \frac{(1-k)\sigma^2}{\sqrt{\Sigma}}\lambda(z_Y)\right] \end{aligned} \tag{B2}$$

Under No Discretion, the firm hires based solely on the test. Since we assume there are plenty of type G applicants, the firm will hire among type G applicants at random. Thus, the expected quality of hires equals μ_G.

B.2 Propositions

Propositions 1 and 2, formalized below, provide intuition for our empirical analysis. Proposition 1 states that the exception rate (the probability that a Y is hired above a G applicant) is increasing in both the precision of a manager's private information and his or her bias. Proposition 2 says that the quality of hired workers, $E[a|Hire]$, is decreasing in manager bias. It also shows that absent bias, the quality of hires is increasing in the precision of a manger's private information.

Proposition 1 *The exception rate is increasing in managerial bias, k, as well as weakly increasing in the precision of the manager's private information, $1/\sigma_\epsilon^2$.*

Proof Because the hiring rate is fixed at W, $E[Hire|Y]$ is a sufficient statistic for the probability that an applicant with $t = Y$ is hired *over* an applicant with $t = G$, i.e., an exception is made.

Recall from above that U_t is normally distributed with mean $(1-k)\mu_t$ and variance $\Sigma = (1-k)^2\sigma^2 + k^2\sigma_b^2$. A manager will hire all applicants for whom U_t is above $\underline{u}$ where the latter is chosen to keep the hire rate fixed at W.

Consider the difference in expected utility across G and Y types. If $\mu_G - \mu_Y$ were smaller, more Y types would be hired, while fewer G types would be hired. This is because, at any given quantile of U_G, there would be more Y types above that threshold.

Let us now define $\tilde{U}_t = \frac{U_t}{\sqrt{\Sigma}}$. This transformation is still normally distributed but now has mean $\frac{(1-k)\mu_t}{\sqrt{\Sigma}}$ and variance 1. Under this rescaling, it will still be the case that the probability of an exception is decreasing in the difference in expected utilities across $\tilde{U_G}$ and $\tilde{U_Y}$: $\Delta_U = \frac{(1-k)(\mu_G - \mu_Y)}{\sqrt{\Sigma}}$.

One can show (with some algebra) that $\frac{\partial \Delta U}{\partial k} = \frac{-k(\mu_G - \mu_Y)\sigma_b^2}{\Sigma^{3/2}}$, which is clearly negative for $k \neq 0$. When k is larger, the expected gap in utility between a G and a Y narrows so the probability of hiring a Y increases.

Similarly, one can show that $\frac{\partial \Delta U}{\partial \sigma_\epsilon^2} = \frac{(1-k)^3(\mu_G - \mu_Y)(\sigma_a^2)^2}{2\Sigma^{3/2}(\sigma_\epsilon^2 + \sigma_a^2)^2}$, which is clearly positive for $k < 1$ (and $\frac{\partial \Delta U}{\partial \sigma_\epsilon^2} = 0$ for $k = 1$). The gap in expected utility between G and Y widens when managers have less information. It thus narrows when managers have better private information, as does the probability of an exception. ■

Proposition 2 *Holding constant information, the quality of hires is decreasing in managerial bias, k. When $k = 0$, the expected quality of hires for a given manager, $E[a|Hire]$, is increasing in the precision of the manager's private information, $1/\sigma_\epsilon^2$.*

Proof The expected quality of hires, $E[a|Hire]$, is given in equation B2, above, and is a function of manager type. With some messy algebra, we obtain that:

$$\frac{\partial E\left[a|Hire\right]}{\partial k} = -\frac{k\sigma_b^2}{W\Sigma^{3/2}}\left[A\sigma^2 + \frac{k^2\sigma_b^2 B\left(\mu_G - \mu_Y\right)^2}{A\Sigma}\right] \tag{B3}$$

and

$$\frac{\partial E\left[a|Hire\right]}{\partial \sigma^2} = \frac{1-k}{2W\Sigma^{3/2}}\left[A\left((1-k)^2\sigma^2 + 2k^2\sigma_b^2\right) - \frac{(1-k)^2 k^2\sigma_b^2 B\left(\mu_G - \mu_Y\right)^2}{A\Sigma}\right] \tag{B4}$$

where $A \equiv p_G\phi\left(\tilde{z}_G\right) + \left(1-p_G\right)\phi\left(\tilde{z}_Y\right)$ and $B \equiv p_G\left(1-p_G\right)\phi\left(\tilde{z}_G\right)\phi\left(\tilde{z}_Y\right)$.

First, note that the derivative wrt k is negative; expected quality of hires is strictly decreasing in bias. Second, note that when setting $k=0$, the derivative wrt σ^2 is positive. Also, recall that σ^2 moves in the same direction as $1/\sigma_\epsilon^2$. Therefore, expected quality of hires is strictly increasing in the precision of private information, when the manager is unbiased.

We next provide intuition for these results by summarizing the logic for why these inequalities should at least weakly hold.

Since the family of normal distributions is Blackwell-ordered by precision, Blackwell's Theorem tells us that an increase in the precision of information must weakly increase the manager's utility. For $k=0$, the manager maximizes $U = E[a|Hire]$, which is a function of the precision of the manager's private information. Therefore, absent bias, the expected quality of hires is increasing in manager information.

For k, consider two managers, with bias k^H and k^L, respectively, where $k^L < k^H$. Each manager chooses a group of hires that maximizes $(1-k)E[a|Hire]+kE[b|Hire]$. Let $\left(a^H, b^H\right)$ denote the realized expectation of a and b, conditional on being hired by the manager with high bias, and let $\left(a^L, b^L\right)$ denote the same for the manager with low bias. Because $\left(a^H, b^H\right)$ and $\left(a^L, b^L\right)$ are chosen optimally, we have the following incentive compatibility (IC) constraints:

$$\begin{aligned}\left(1-k^H\right)a^H + k^H b^H &\geq \left(1-k^H\right)a^L + k^H b^L \\ \left(1-k^L\right)a^L + k^L b^L &\geq \left(1-k^L\right)a^H + k^L b^H\end{aligned}$$

We would like to prove that $a^H \leq a^L$. Suppose to the contrary that $a^H > a^L$.

First, note that it cannot be that $a^H > a^L$ *and* $b^H > b^L$ because the choice, $\left(a^L, b^L\right)$, would violate the IC of the low-bias manager.

Second, consider where $a^H > a^L$ and $b^H \leq b^L$. In this circumstance, the high-bias manager is choosing candidates with higher a and lower b than the low-bias manager. We can rearrange and sum the IC constraints to show that they imply the following:

$$\left(k^H - k^L\right)\left(\left(b^H - b^L\right) - \left(a^H - a^L\right)\right) \geq 0$$

However, this expression is false because, by assumption, we have $k^H > k^L$, $a^H > a^L$, and $b^H \leq b^L$.

Therefore, by contradiction, we have shown that $a^H \leq a^L$. ■

Discussion on Propositions 1 and 2. From Propositions 1 and 2, we observe that if high-exception managers achieve worse outcomes than low-exception managers, this must be because high-exception managers are biased or mistaken.

Formally, consider two managers, manager 1 and manager 2. The two managers have type (k_i, h_i), where k_i is bias and h_i is the precision (i.e., inverse variance) of each manager's private information, for $i \in \{1, 2\}$. The two managers have exception rates, R_i, and quality of hires, a_i, for $i \in \{1, 2\}$. We claim that if $R_1 > R_2$ and $a_1 < a_2$, then it must be that $k_1 > 0$.

To see this, suppose to the contrary that $k_1 = 0$. Then by Proposition 1, it must be that $h_1 > h_2$. That is, if manager 1 is weakly less biased than manager 2 but still has more exceptions, manager 1 must have more precise private information. Now consider a third manager with bias, $k_3 = 0$, precision of private information $h_3 = h_2$, and outcomes a_3. That is, manager 3 has no bias and the same information as manager 2. By Proposition 2, it must be that $a_1 > a_3 > a_2$.[20] But this is a contradiction.

Having discussed how Propositions 1 and 2 help frame our empirical work, we now present Proposition 3. Proposition 3 illustrates the fundamental tradeoff firms face when allocating authority: managers have private information, but they are also biased. Greater bias pushes the firm to prefer No Discretion, while better information tends to push it towards Discretion. Specifically, the first finding states that when bias, k, is low, firms prefer to grant discretion, and when bias is high, firms prefer No Discretion. Part 2 states that for any level of bias, there is a precision of private information small enough that firms prefer No Discretion. Uninformed managers would at best follow test recommendations and, at worst deviate because they are mistaken or biased. Finally, part 3 states that, for any fixed information precision threshold, there exists an accompanying bias threshold such that if managerial information is greater and bias is smaller, firms prefer to grant discretion. Put simply, Discretion beats out No Discretion when a manager has very precise information, but only if the manager is not too biased.

Proposition 3 *We formalize conditions under which the firm will prefer Discretion or No Discretion.*

1. *For any given precision of private information, $1/\sigma_\epsilon^2 > 0$, there exists a $k' \in (0, 1)$ such that if $k < k'$ worker quality is higher under Discretion than No Discretion and the opposite if $k > k'$.*

[20] Manager 1 should have better outcomes than manager 3 because they both have no bias but manager 1 has better information. Manager 3 should have weakly better outcomes than manager 2 because they have the same information but manager 3 is unbiased (so therefore weakly less biased than 2).

2. *For any given bias, $k > 0$, there exists $\underline{\rho}$ such that when $1/\sigma_\epsilon^2 < \underline{\rho}$, i.e., when precision of private information is low, worker quality is higher under No Discretion than Discretion.*

3. *For any precision of information $\overline{\rho} \in (0, \infty)$, there exists a bias, $k'' \in (0, 1)$, such that if $k < k''$ and $1/\sigma_\epsilon^2 > \overline{\rho}$, i.e., high precision of private information and low bias, worker quality is higher under Discretion than No Discretion.*

We next prove each item of Proposition 3:

1. *For any given precision of private information, $1/\sigma_\epsilon^2 > 0$, there exists a $k' \in (0, 1)$ such that if $k < k'$ worker quality is higher under Discretion than No Discretion and the opposite if $k > k'$.*

Proof When $k = 1$, the manager hires based only on b, which is independent of a. So $E[a|Hire] = p_G\mu_G + (1 - p_G)\mu_Y$. The firm would do better under No Discretion (where quality of hires equals μ_G).

When $k = 0$, under No Discretion, the quality of hires remains equal to μ_G. Write $\rho = 1/\sigma_\epsilon^2$ for the precision of the manager's information, and write $\mu(k, \rho)$ for average quality of hire under Discretion for a manager with bias k and precision ρ. Consider a precision ρ' with $0 < \rho' < \rho$. By Blackwell's Theorem, it must be that $\mu_G \leq \mu(0, \rho')$. That is, an unbiased manager with $\rho' > 0$ will do better than hiring by the test score alone. Further, since quality of hire is increasing in precision (Proposition 2), we know that $\mu(0, \rho') < \mu(0, \rho)$. By transitivity, $\mu_G < \mu(0, \rho)$, and the firm would do better under Discretion.

Thus, Discretion is better than No Discretion for $k = 0$ and the opposite is true for $k = 1$. Proposition 2 shows that the firm's payoff is strictly decreasing in k. There must therefore be a single cutpoint, k', where, below that point, the firm's payoff for Discretion is larger than that for No Discretion, and above that point, the opposite is true. ■

2. *For any given bias, $k > 0$, there exists $\underline{\rho}$ such that when $1/\sigma_\epsilon^2 < \underline{\rho}$, i.e., when precision of private information is low, worker quality is higher under No Discretion than Discretion.*

Proof Fix bias $k > 0$. One can show that as $1/\sigma_\epsilon^2$ approaches 0 and therefore σ^2 approaches 0, the expected quality of hires (equation (B2)) is strictly less than μ_G. To see this, note that equation (B2) can be rearranged to obtain:

(B5) $$W * E[a|Hire] = p_G(1-\Phi(z_G))\mu_G + (1-p_G)(1-\Phi(z_Y))\mu_Y + \frac{(1-k)\sigma^2}{\sqrt{\Sigma}}\left[p_G\phi(z_G) + (1-p_G)\phi(z_Y)\right]$$

As σ^2 approaches 0, the second term vanishes. This term is bounded a way from μ_G since the gap $z_G - z_Y$ approaches $\frac{(1-k)(\mu_G-\mu_Y)}{k\sigma_b}$, which is bounded away from ∞.

Thus, as precision of private information tends towards ∞, the firm will prefer No Discretion (which has a payoff of μ_G) to Discretion.

We also point out that the firm's payoff under Discretion, expressed above in equation (B2), is clearly continuous in σ (which is continuous in $1/\sigma_\epsilon^2$).

Thus, as the manager tends towards no information, the firm prefers No Discretion and the firm's payoff under Discretion is continuous in the manager's information. Therefore there must be a point $\underline{\rho}$ such that, for precision of manager information below that point, the firm prefers No Discretion to Discretion. ■

3. *For any precision of information $\overline{\rho} \in (0,\infty)$, there exists a bias, $k'' \in (0,1)$, such that if $k < k''$ and $1/\sigma_\epsilon^2 > \overline{\rho}$, i.e., high precision of private information and low bias, worker quality is higher under Discretion than No Discretion.*

Proof Define $\Delta(\sigma_\epsilon^2, k)$ as the difference in quality of hires under Discretion, compared to No Discretion, for fixed manager type (σ_ϵ^2, k). Since the firm's payoff under Discretion is continuous in both k and σ_ϵ^2 (see Equation (B2) above), Δ must also be continuous in these variables. By Proposition 2, the payoff of Discretion is strictly decreasing in the bias k, and so $\Delta(\sigma_\epsilon^2, k)$ is strictly decreasing in k. Moreover, when $k = 0$, Discretion is strictly preferable to No Discretion for $\sigma_\epsilon^2 < \infty$ (see proof to Proposition 3 part 1), so $\Delta(\sigma_\epsilon^2, k) > 0$ for $\sigma_\epsilon^2 < \infty$. Finally, Proposition 2 shows that when $k = 0$, the firm's payoff under Discretion is increasing in $\frac{1}{\sigma_\epsilon^2}$, i.e., $\Delta(\sigma_\epsilon^2, 0)$ is decreasing in σ_ϵ^2.

Fix any $\overline{\rho} \in (0,\infty)$ and let $\overline{\sigma_\epsilon^2} = 1/\overline{\rho}$. We seek to show that there exists k'' such that $\Delta(\sigma_\epsilon^2, k) > 0$ for all $\sigma_\epsilon^2 < \overline{\sigma_\epsilon^2}$ and all $k < k''$.

Let $y = \Delta(\overline{\sigma_\epsilon^2}, 0) > 0$. This is the difference in the firm's payoff for Discretion compared to No Discretion for an unbiased manager, with some minimal amount of information, $\overline{\sigma_\epsilon^2}$. We know $y > 0$ because for an unbiased manager, Discretion strictly improves upon No Discretion (see part 1 of Proposition 3).

Since $\Delta(\sigma_\epsilon^2, 0)$ is decreasing in σ_ϵ^2, it holds that $\Delta(\sigma_\epsilon^2, 0) > y$ for all $\sigma_\epsilon^2 < \overline{\sigma_\epsilon^2}$. Therefore, it suffices to show that there exists k'' such that $\Delta(\sigma_\epsilon^2, 0) - \Delta(\sigma_\epsilon^2, k) < y$ for all $k < k''$ and

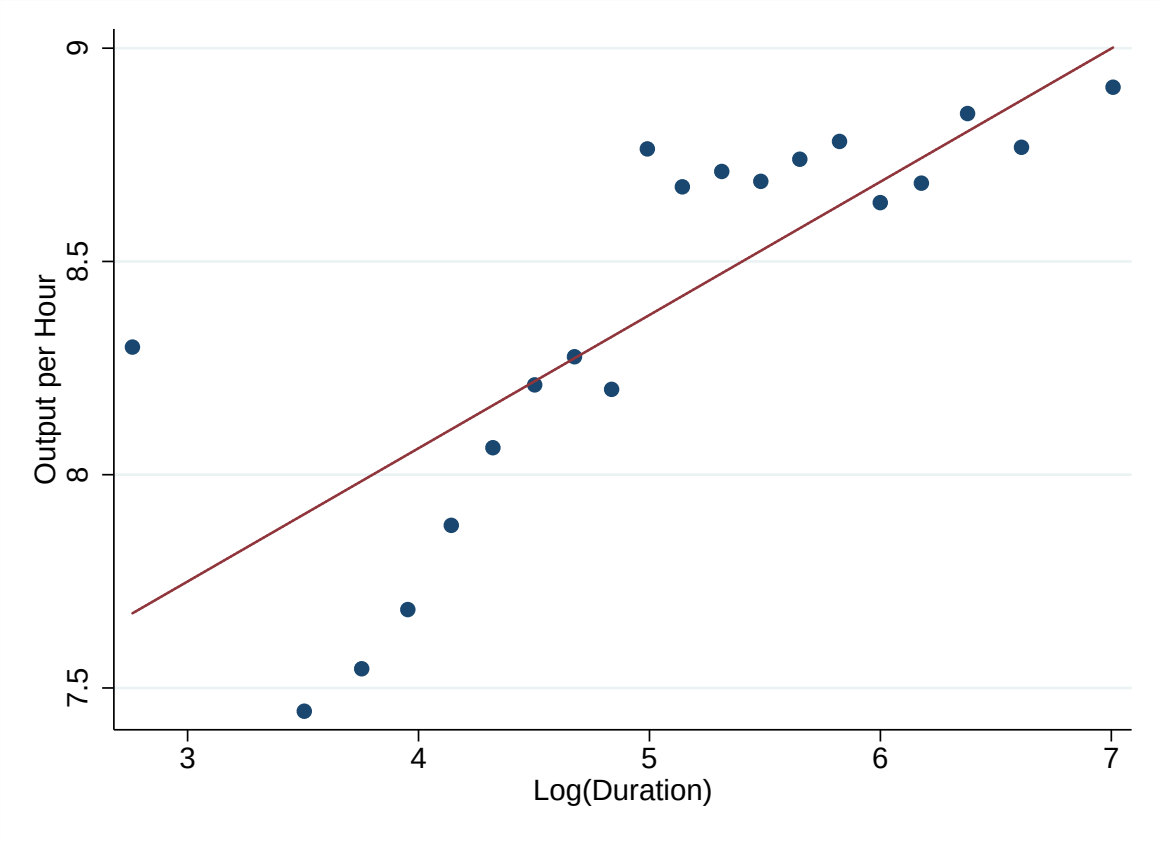
$\sigma_\epsilon^2 < \overline{\sigma_\epsilon^2}$. That is, for bias less than k'' and info more precise than the minimal case $\overline{\sigma_\epsilon^2}$, we want to show that there is some small enough bias, such that removing this bias improves the firm's payoff by only a small amount, relative to the case of no bias and minimal information, y. If this is true, then the bias amount is not enough to make the firm prefer No Discretion, because $y > 0$.

Let $d(k) = \max_{\sigma_\epsilon^2 \in [0, \overline{\sigma_\epsilon^2}]} \Delta(\sigma_\epsilon^2, 0) - \Delta(\sigma_\epsilon^2, k)$. We know $d(k)$ exists because $\Delta()$ is continuous wrt σ_ϵ^2 and the interval over which we take the maximum is compact. In words, $d(k)$ finds the largest possible improvement from eliminating bias, for bias k, across all values of information in the range.

It now suffices to show that there exists k'' such that $d(k) \leq y$ for all $k < k''$. This holds because $d(0) = 0$ (by definition) and $y > 0$, and because $d(k)$ is continuous in k (since Δ is). ■

C Supplemental Tables and Figures

Appendix Figure C1: Output per hour and Job Durations



Notes: Figure C1 plots average output per hour within 20 evenly sized bins, based on log(duration). It controls for location fixed effects to account for differences in average output per hour across locations. We use "binscatter" in Stata.

Appendix Table C1: Testing and job durations
Additional outcomes

	>3 Months (Mean=0.62; SD=0.49)		>6 Months (Mean=0.46; SD=0.50)		>12 Months (Mean=0.32; SD=0.47)	
	(1)	(2)	(3)	(4)	(5)	(6)
Introduction of Testing						
Post-Testing	0.0427*	0.0259	0.0919**	0.0597***	0.106***	0.0750***
	(0.0220)	*(0.0200)*	*(0.0371)*	*(0.0228)*	*(0.0369)*	*(0.0198)*
N	256,641	256,641	243,580	243,580	217,514	217,514
Post-Testing Correlations						
Manager Exception Rate	-0.0261***	-0.0171**	-0.0158**	-0.0101*	-0.00471	-0.0127**
	(0.00940)	*(0.00780)*	*(0.00638)*	*(0.00602)*	*(0.00496)*	*(0.00483)*
N	82,365	82,365	71,388	71,388	56,436	56,436
Differential Impact of Testing by Location-Level Exception Rate						
Post-Testing	0.0420*	0.0258	0.0912**	0.0581***	0.101***	0.0738***
	(0.0215)	*(0.0187)*	*(0.0369)*	*(0.0218)*	*(0.0350)*	*(0.0191)*
*Location Exception Rate*Post-Testing*	-0.0271*	-0.0372**	-0.0297	-0.0318*	-0.0548***	-0.0248
	(0.0160)	*(0.0173)*	*(0.0206)*	*(0.0179)*	*(0.0196)*	*(0.0172)*
N	256,641	256,641	243,580	243,580	217,514	217,514
Base Controls	X	X	X	X	X	X
Full Controls		X		X		X

*** p<0.01, ** p<0.05, * p<0.1

Notes: See notes to Tables II, III, and IV of the main text. The dependent variables are the probability that a worker survives 3, 6, or 12 months, respectively, among those who are not right-censored, i.e., those hired at least that many months before the data end date for each of the 15 firms. We use OLS regressions. Base controls include location, hire month, and position fixed effects. Full controls add client-by-year effects, local unemployment rates, and location-specific time trends. Full controls in the middle panel also include applicant pool characteristics. The top panel provides estimates of the impact of testing on job duration outcomes. The middle panel estimates the post-testing correlation between job duration and manager-level exception rates. The bottom panel estimates the differential impact of testing by location-level exception rates.

Appendix Table C2: Job Duration of Workers, by Length of Time in Applicant Pool

	Dependent Variable: Log(Duration)					
	(1)	(2)	(3)	(4)	(5)	(6)
	Green Workers		**Yellow Workers**		**Red Workers**	
Waited 1 Month	0.00545	-0.0276	-0.0271	-0.0139	-0.0338	-0.0449
	(0.0281)	*(0.0263)*	*(0.0320)*	*(0.0242)*	*(0.0622)*	*(0.0752)*
Waited 2 Months	-0.0352	-0.0714	-0.0204	-0.0542	0.00713	0.0467
	(0.0586)	*(0.0632)*	*(0.0647)*	*(0.0663)*	*(0.144)*	*(0.174)*
Waited 3 Months	0.00486	-0.0941	0.112	0.120	0.0338	0.0493
	(0.0673)	*(0.0851)*	*(0.0855)*	*(0.0867)*	*(0.220)*	*(0.242)*
N	47,809	47,809	24,496	24,496	4,098	4,098
Base Controls	X	X	X	X	X	X
Initial Applicant Pool FEs		X		X		X

*** p<0.01, ** p<0.05, * p<0.1

Notes: Regressions are restricted to the post-testing sample, adjust for censoring, and cluster standard errors at the location level. Each panel compares applicants who started working in the month they applied (omitted category) to those who started 1, 2, or 3 months later, separately by color. Panels restrict to applicant pools (location-recruiter-initial application month) with variation in wait time, and further restrict to locations and pools with at least 10 and 5 observations, respectively. Base controls are location, hire month, and position type fixed effects. Initial applicant pool fixed effects are defined by the manager-location-month for the pool when candidates first applied.

Appendix Table C3: Exception Rates and Duration Outcomes Applicant Pools with at Least as Many Green Applicants as Total Hires

	Dependent Variable: Log(Duration)			
	(1)	(2)	(3)	(4)
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.323***	0.262***
			(0.117)	*(0.0609)*
*Exception Rate*Post-Testing*	-0.112***	-0.111***	-0.243**	-0.116
	(0.0355)	*(0.0303)*	*(0.0949)*	*(0.0774)*
N	76,425	76,425	250,754	250,754
Base Controls	X	X	X	X
Full Controls		X		X

*** p<0.01, ** p<0.05, * p<0.1

Notes: Columns 1 and 2 estimate the post-testing correlation between manager-level exception rates and log duration (see Table III). Columns 3 and 4 estimate the differential impact of testing by location-level exception rates (see Table IV). Columns 1 and 2 include only hires from applicant pools with at least as many green applicants as total hires, in the post-testing sample. Columns 3 and 4 add all pre-testing observations. Base controls include location, hire month, and position fixed effects. Full controls add client-by-year effects, local unemployment rates, and location-specific time trends (and applicant pool controls in columns 1 and 2). In order to identify these controls, we must further restrict this subsample to locations that hire in at least 2 months in the post-testing period (all but 0.2% of observations).

References

[1] Autor, David and David Scarborough, "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments," *Quarterly Journal of Economics*, 123 (2008), 219-277.

[2] Fernandez, Roberto M., Emilio J. Castilla, and Paul Moore, "Social Capital at Work: Networks and Employment at a Phone Center," *American Journal of Sociology*, 105 (2000), 1288-1356.