Discretion in Hiring

Mitchell Hoffman
University of Toronto
& NBER

Lisa B. Kahn
Yale University & NBER

Danielle Li
Harvard University
& NBER

June 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 because they have superior private information. The firms in our setting may therefore be able to improve outcomes of hires by limiting managerial discretion and relying more on test recommendations.

JEL Classifications: M51, J24

Keywords: Hiring; rules vs. discretion; job testing

---

*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, Harvard Business School, 211 Rock Center, Boston, MA 02163. Email: dli@hbs.edu. We are grateful to Jason Abaluck, Ajay Agrawal, Ricardo Alonso, David Berger, Arthur Campbell, David Deming, Alex Frankel, Avi Goldfarb, Harry Krashinsky, Jin Li, Liz Lyons, Steve Malliaris, Mike Powell, Kathryn Shaw, Steve Tadelis, and numerous seminar participants. 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.

1 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 assess how firms should incorporate it into their hiring decisions. 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 to a test score for each

[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 (2012) 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/.

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 firms should use this information. In particular, we ask whether firms should limit discretion by relying more on test recommendations, relative to the status quo. We propose a model in which firms rely on potentially biased HR managers who observe both public and private signals of worker quality. Managers can decide to hire workers with the best test scores or make "exceptions" by hiring against the test recommendation. Managers with more precise signals of worker quality are both more likely to make exceptions and to hire workers who are a better fit; firms would therefore benefit from continuing to grant discretion to such managers. By contrast, managers who are biased or have poor judgment are also more likely to make exceptions, but the workers they hire will have worse outcomes on average. The model thus implies that these cases can be distinguished by examining the relationship between a manager's propensity to make exceptions and worker outcomes: a positive relationship suggests that managers make exceptions when they are better informed, while a negative relationship suggests that exceptions are driven by biases or mistakes. In the latter case, firms may be able to improve worker outcomes by limiting managerial discretion.

Our data, which includes information on applicants as well as hired workers, allows us to apply this diagnostic 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 exercise discretion because they are biased or have poor judgement, not because they are better informed.

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. 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. Based on this supplemental analysis, we see no evidence that managers trade off duration for productivity. Taken together, our findings suggest that placing more weight on job test recommendations may result in better hires for the firms in our sample.

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 managers make exceptions because they are biased or have poor judgement, then we should see exceptions associated with worse outcomes.

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 how job testing should be used. 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 2 describes the setting and data. Section 3 evaluates the impact of testing on job duration. Section 4 presents a model of hiring and derives an empirical diagnostic for assessing the relationship between discretion and turnover. Section 5 applies the diagnostic to our empirical setting. Section 6 concludes.

[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 (2014), Horton (2013), Brown et al. (2015), Burks et al. (2015), and Pallais and Sands (2015)) 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. (2017) for empirical evidence that machine-based algorithms outperform judges in deciding which arrestees to detain pre-trail.

2 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 Appendix A.

2.1 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, 46% of applicants receive a green score, 33% score yellow, and 21% score red. See Appendix A.1 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 date at which the modal hire had a test score. See 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]

2.2 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 1 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 firms. To illustrate this concern, Figure 1 shows a histogram of job tenure for completed

[6] Other managers may take part in hiring decisions as well. For example, in one firm, recruiters often endorse a candidate to an operations manager 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 Appendix A.3 for details and robustness to restricting the sample to client firms that likely did not have pre-sample testing.).

[8] See Appendix A.6 for sample restrictions.

spells (79% of the spells in our data) among employees in our sample. The median worker (solid red 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 specifically 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 1 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 et al., (2015) to evaluate the value of bosses in a comparable setting to ours.

Table 1 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 8.3 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 1 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.

3 The Impact of Testing

3.1 Empirical Strategy

Before examining whether firms should grant managers discretion over how to use job testing information, we first evaluate the impact of introducing testing itself. 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.

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

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 date 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 (Appendix Table A1 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_{ilt}, 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.

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

3.2 Results

Table 2 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 adept at retaining their data or 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 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, we find significant effects on the order of 14% when estimating the impact of testing on the length of *completed* job spells, which is similar to what Autor and Scarborough (2008) find using that same outcome. Further, the Autor and Scarborough estimate is inside the range of our confidence intervals. 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 very similar.

Figure 2 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 over-ruling job test recommendations.

4 Model

We formalize a model in which a firm makes hiring decisions with the help of an HR manager. This model has two purposes. First, it builds intuition for the rules-vs-discretion tradeoff that a firm faces. Granting discretion enables firms to take advantage of a manager's private information but comes at the cost of allowing for managerial biases and mistakes. Second, this model lets us derive an empirical diagnostic for whether firms that currently allow for discretion can improve hiring outcomes by relying more on test recommendations. We then apply this test in Section 5.

In Section 4.1, we describe a "Discretion" regime, where managers are allowed to make hiring decisions, weighing applicants' test scores against other attributes as they choose. In Section 4.2, we generate a diagnostic for when an alternative regime, "No Discretion", will dominate. We define the No Discretion regime, as hiring applicants in order of their test scores, randomizing within score to break ties. That is, eliminating discretion means hiring only the basis of the test recommendation.

Neither regime need be the optimal policy response after the introduction of testing. Firms may, for example, consider hybrid policies such as requiring managers to hire lexicographically by the test score before choosing preferred candidates, and these may generate

[16] Figure 2 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. Appendix Figure A3 replicates Figure 2 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.

better outcomes. Rather than solving for the optimal hiring policy, we focus on the extreme of eliminating discretion entirely, doing so for simplicity.[17] All proofs are in Appendix B.

4.1 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.[18]

The firm's objective is to hire a proportion, W, of workers that maximizes expected quality, $E[a|Hire]$.[19] For simplicity, we also assume $W < p_G$.[20]

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 $a \perp b$.

Managers may value the firm's payoff, a, because they are directly incentivized to, because they risk termination, or because they are simply altruistic.[21] The additional quality, b, can

[17] We also abstract away from other policies the firm could adopt, for example, directly incentivizing managers based on the productivity of their hires or fully replacing managers with the test. See Frankel (2016) for a theoretical discussion of optimal hiring in a similar setting.

[18] 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.

[19] In theory, firms should hire all workers whose expected value is greater than their cost (wage). 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).

[20] This implies that a manager could always fill a hired cohort with type G applicants. In our data, 0.46 of applicants are green and 0.58 of the green or yellow applicants are green, while the hire rate is 19%, so this will be true for the typical pool.

[21] 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 Appendix A.7 for more detail.

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.[22]

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 $N(0, \sigma_b^2)$ with $\sigma_b^2 \in (0, \infty)$. 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. The parameter $\sigma_\epsilon^2 \in \mathbb{R}_+ \cup \{\infty\}$ measures the manager's *information*. A manager with perfect information on a_i has $\sigma_\epsilon^2 = 0$, while a manager with no private information has $\sigma_\epsilon^2 = \infty$. The 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.

Managers form a posterior expectation of worker quality and hire a worker if and only if the expected value of their own utility U_i conditional on s_i, b_i, and 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).

[22]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 et al., 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.

4.2 Model Predictions

The Discretion regime, as described above, allows managers to weigh both the test and their private signals in making ultimate hiring decisions. As an alternative, we define the No Discretion regime as hiring based solely on the test recommendation, randomizing within score to break ties. In this subsection, we generate a diagnostic for when one policy will dominate the other.

First, Proposition B.1, presented in Appendix B provides formal conditions under which the firm will prefer Discretion or No Discretion. In general, it states that the quality of hires $E[a|Hire]$ is decreasing in the bias of the manager and increasing in their information. Greater bias pushes the firm to prefer No Discretion, while better information pushes it towards Discretion.

Firms, however, cannot directly observe a manager's bias or information, making it difficult to apply this result in practice. Instead, it is easier to observe (as we do in our setting) 1) the choice set of applicants available to managers and 2) the performance outcomes of workers hired from those applicant pools. We now discuss predictions about how a manager's type influences his or her observable hiring practices.

We define a hired worker as an "exception" if the worker would not have been hired under No Discretion (i.e., based on the test recommendation alone): any time a Y worker is hired when a G worker applied but is not hired. Denote the probability of making an exception for a given manager as R_m. Note that $R_m = E_m[Pr(Hire|Y)]$: the probability of an exception is simply the probability that a Y type is hired, because this is implicitly also equal to the probability that a Y is hired *over* a G.

Proposition 4.1 *The exception rate, R_m, is increasing in both managerial bias, k, and the precision of the manager's private information, $1/\sigma_\epsilon^2$.*

Intuitively, managers with better information make more exceptions because they place relatively more weight on their own signal of a. Managers with more bias also make more exceptions, but because they place more weight on maximizing other qualities, b. It is therefore difficult to discern whether granting discretion is beneficial to the firm simply by examining how often managers make exceptions.

However, while exceptions (R_m) are increasing in both bias and information, quality ($E[a|Hire]$) is decreasing only in bias. This suggests that we can assess whether exceptions are primarily driven by bias (in which case firms may want to limit discretion) by examining the relationship between exceptions and the quality of hires. This is formalized in the following result:

Proposition 4.2 *If the quality of hired workers is decreasing in the exception rate,* $\frac{\partial E[a|Hire]}{\partial R_m} < 0$*, then firms can improve outcomes by eliminating discretion. If quality is increasing in the exception rate then Discretion is better than Do Discretion.*

Proposition 4.2 states that if $E[a|Hire]$ is negatively correlated with R_m, then it is likely that exceptions are being driven primarily by managerial bias (because bias increases the probability of an exception and decreases the quality of hires). In this case, eliminating discretion can improve outcomes. If the opposite is true, then exceptions are primarily driven by private information and discretion is valuable.

For intuition, consider a manager who never makes exceptions. This manager's type must then have no additional information or preferences relative to the test. As such, the quality of this manager's hires is equivalent to that of workers hired under No Discretion, that is, using only the test and randomizing within score. If managers with types that lead them to make more exceptions turn out to do better, then allowing for discretion can improve outcomes relative to No Discretion. If they do worse, then firms can improve outcomes by moving to a regime with no exceptions—that is, by eliminating discretion and using only the test.

5 Empirical Analysis on Discretion

When testing is available, does granting managerial discretion result in better or worse outcomes than a hiring regime based solely on the test? In our data, we only observe managers hiring under discretion, and therefore cannot directly compare the two regimes. However, our model motivates the following empirical test to answer the same question: is worker tenure increasing or decreasing in the probability of an exception? That is, are

outcomes better when a manager hires more exceptions or when a manager follows test recommendations more closely?

In our model, variation in the use of exceptions across managers is driven exclusively by manager type (their information and bias). In reality, however, variation in exception rates may be driven by factors other than a manager's type. For example, two managers of the same type may nonetheless make different numbers of exceptions if they face different applicant pools, need to hire different numbers of workers, or if they face different labor market conditions. These factors may also separately impact worker tenure, making it difficult to learn about the impact of discretion from the relationship between exceptions and worker outcomes.

We must therefore address two key empirical issues in order to implement the test suggested by Proposition 4.2. First, we need to carefully define an "exception rate" that corresponds to a manager's choice to exercise discretion. 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. 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. A concern is that applicant pools in which managers make more exceptions may differ in their unobservable applicant characteristics in ways that also impact worker durations.

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).

5.1 Defining Exceptions

To construct an empirical analogue to the exception rate R_m, 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.[23]

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 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

[23] 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.

and the number of hires required.[24] 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 4 are derived for the probability of an exception, their proofs hold equally for this definition of an exception rate. In Section 5.4.4 we show that our results are also robust to alternative definitions of exception rates.

As described in Table 1, we observe nearly 3,700 applicant pools consisting of, on average, 260 applicants.[25] 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 worker is hired from an applicant pool in which 24% of possible exceptions are made.

There is substantial variation in exception rates across both applicant pools and managers. Figure 3 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.

To test Proposition 4.2, we 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 bottom panels of Figure 3 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.

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

[24] 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^h_{color} is the number of applicants of a given color and N^h is the total number of hires.

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

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.

5.2 Empirical Specifications

5.2.1 Post-testing correlation between exception rates and outcomes

The most direct implementation of Proposition 4.2 examines the correlation between the manager-level exception rate and the realized durations of hires in the post-testing period:

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

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. In our model, such a finding suggests that firms may improve worker outcomes by relying more on test recommendations. We cluster standard errors at the location level, again to take into account any correlation in observations within a location over time.[26]

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.

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

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

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 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).[27]

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.

5.2.2 Differential impact of testing, by exception rates

Our second test considers how the *impact* of testing differs across managers by exception rate. If managers exercise discretion because they are biased or misinformed, then we may expect the benefits of testing that we document in Section 3.2 to be lower for high exception

[27] 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.

managers. We estimate this using a similar specification to that described in Section 3.1:

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

Equation (4) includes the main effect of testing but allows testing to interact with the manager-specific exception rate. The coefficient of interest, b_1, estimates how the impact of testing differs when managers make exceptions. We cannot control for manager fixed effects because our data do not contain manager identifiers in the pre-testing period. Instead, as usual, we control for location fixed effects. The interpretation of b_1 is thus the differential change in durations for a manager with a higher exception rate, relative to the average duration across all managers at a location pre-testing. $b_1 < 0$ indicates that exceptions attenuate gains from testing.

Unlike Equation (3), which can only be estimated on post-testing data, this test uses our full dataset, making it possible for us to more precisely identify location fixed effects and other controls in addition to manager exception rates.

5.3 Results

Figure 4 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 plot the average tenure outcome against the average exception rate (y-axis). The top left panel plots average log duration, adjusted for censoring.[28] 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 3 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

[28] We estimate a censored-normal regression of log(duration) on indicators for the 20 exception rate bins and plot the coefficients. In these and in the milestone plots, we include only base controls to illustrate as close as possible the raw data: location, hire month, and position fixed effects.

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 applicant 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 Appendix Table C1 summarizes regressions for the milestone measures, where we also find significant negative relationships.

Next, Table 4 examines how the impact of testing varies by the extent to which managers 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 0.25 log points, but that this effect is substantially offset (by 0.10 log points) for each standard deviation increase in the exception rate, significant at the 1% level.[29] The bottom panel of Appendix Table C1 shows that these results are robust to OLS estimates using milestone measures as dependent variables.

Figure 5 illustrates how the impact of testing varies for locations with different manager exception rates, using our full set of 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 20% longer worker durations in the post testing period, relative to a firm made up of 90th percentile managers (higher exception rates are

[29]For these specifications we do not include controls for applicant pool quality, since pool quality is unavailable pre-testing. However, results are similar when we incorporate these controls by adding zeroes in the pre-testing period, effectively controlling for the interaction of testing and pool quality.

[30]To construct this, we divide locations into 20 hire-weighted bins based on their average manager-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. Observations in the graph are weighted by the inverse variance of the estimated testing impact for each bin. For the Log(duration) outcome (top left panel), we adjust for censoring with censored-normal regressions.

worse). Such firms would experience increased durations with the adoption of testing that are more than double the high exception rate manager's.

Viewed in light of our theoretical predictions, these results suggest that managers make exceptions primarily because they are either biased or misinformed, not because they have superior private information about a worker's potential duration. In this case, firms may be able to improve retention by limiting discretion and relying more on test recommendations.

5.4 Additional Robustness Checks

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

5.4.1 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, managers may make more exceptions when green applicants in a given applicant pool are idiosyncratically weak. If yellow workers in these pools are weaker than green workers in our sample on average, it will appear that more exceptions are correlated with worse outcomes even though managers are making individual exceptions to maximize worker quality. Because we include location fixed effects in our regressions and aggregate exception rates to the manager level, this scenario would only be of concern if green applicants were idiosyncratically weak over the entire time a manager is hiring, relative to the typical applicants at the location. As another example, another explanation for our finding that locations with more exceptions see fewer gains from testing may be that these locations are ones in which managers have always had better information about applicants: they see fewer improvements from testing because they simply do not need the test.

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 5 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 5 repeats this exercise, comparing exception reds (the omitted group), to passed over yellows and greens.[32]

In both panels, we find that workers hired as exceptions have shorter tenures. Passed over greens stay roughly 4% longer than the yellows hired before them from the same pool (column 2, top panel), though this estimate is noisy. We estimate a more precise relationship when considering exception-red workers: greens and yellows stay roughly 14% and 12% longer, respectively, than the reds they were passed over for (bottom panel). The results in Table 5 suggest that it is unlikely that exceptions are driven by better information. When workers with better test scores are at first passed over and then later hired, they still outperform the workers chosen first.

An alternative explanation is that the applicants with higher test scores were not initially passed up, but were instead initially unavailable. For example, higher quality workers may be more likely to engage in on-the-job search, and therefore require more time to matriculate after receiving an offer. In such cases, we may expect durations to differ for workers who take more or less time between application and start date, even within color score. However, Appendix Table C2 shows that, within color, job durations of workers hired immediately do not differ from durations for those hired after a lag. While this analysis is not meant

[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 3.

[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 post-testing hires in the top and bottom panels, respectively). We further restrict to locations and pools that have at least 10 and 5 observations, respectively, to be able to identify control variables.

to prove that green workers hired later were not initially unavailable, our results do suggest that delays are not correlated with worker quality.[33]

5.4.2 "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. In Section 5.2, we discussed how our specification alleviate concerns that such false exceptions drive our results: 1) our use of manager-level exception rates means we aggregate over some of the idiosyncratic variation that may generate false exceptions; 2) our controls for local labor market conditions may proxy for drivers of low yields among green applicants. We further note that pools with exception rates above 50% make up fewer than 2% of hires and that pools where *only* exceptions are hired constitute only 0.6% of hires. We might be especially worried that these cases are driven by false exceptions, so it is comforting that they are rare.

We also consider an additional test, based on the relative plentifulness of green applicants. In pools with many green applicants, it is less likely that a yellow or red worker was hired because all green applicants received an offer and turned it down. In such pools, a yellow or red hire may indicate a more active choice on the part of the manager, rather than a false exception. In Appendix Table C3, we restrict our analysis to applicant pools in which there are at least as many green applicants as the total number of hires. These pools represent a majority (84%) of hires. For this subsample, we find, if anything, a stronger negative relationship between exceptions and worker duration. While we cannot rule out the presence of false exceptions in our data, it is comforting that we find consistent results on a sample in which they possibly constitute a smaller fraction of hires.

[33] Appendix Table C2 also provides insight about how much information managers have, beyond the job test. If managers have useful private information about workers, then we would expect them to be able to distinguish quality within test-color categories: greens hired first should be better than greens who are passed up. The fact that we find no such gradient suggests the value of managerial private information is small, relative to the test.

5.4.3 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.

Our results on passed over workers already suggest that such explanations may not be likely: if managers were taking into account local heterogeneity to make better exceptions, then we should see that exception yellows stay longer.

We also provide more direct evidence that the apparent usefulness of the test does not systematically vary by location characteristics. Figure 6 plots the relationship between manager-level exception rates and worker duration, separately by color.[34] There are two main patterns to notice. First, greens perform better than yellows who in turn perform better than reds across all 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, regardless of the manager's exception rate.

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.

In Appendix A.5 we explore the relationship between color score and job duration as a function of a wide range of location characteristics. We robustly find that color score is predictive of worker quality, regardless of the location's characteristics on each of these dimensions.

[34] We regress log duration on indicators for each of 20 equally sized (based on number of hires) exception rate bins, separately by color, adjusting for censoring and including base controls (location, hire month, and position fixed effects).

5.4.4 Alternative exception rate definitions

In Appendix Table C4, we examine the robustness of our main results in Tables 3 and 4 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 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.[35] 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-level and then standardized. Appendix Table C4 shows that all of these metrics tell similar stories. Results are robust quantitatively and generally in terms of statistical significance as well.

5.4.5 Productivity

Our results suggest that firms can improve worker retention by relying more on test recommendations. However, firms may not want to pursue this strategy if their HR managers exercise discretion in order to improve worker quality on other metrics. For example, man-

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

agers may optimally choose to 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 in this paper.

To simplify our analysis, and to clean up some of the day-to-day variation in this measure, we define a worker-level output per hour metric that averages over all the days where a metric is available for the worker. This measure has an average of 8.4 with a standard deviation of 4.7. There is thus a wide range of performance outcomes.[36]

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 and time periods.[37] 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 durations.[38] Despite this, we still find that output is positively correlated with job durations. Appendix Figure C1 presents a binned scatter of output per hour and Log(Duration) of the worker for 20 evenly-sized bins.[39] 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.

[36] 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.

[37] 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.

[38] For example, the probability of having a missing output measure falls in half over the first month of employment.

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

Table 6 summarizes our main analyses using output per hour as the dependent variable. Columns 1-2 document the post-testing correlation between 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 of -0.11 means that workers hired to a manager with a one standard deviation higher exception rate performs 0.11 fewer output units per hour, or 2.3% of a standard deviation worse. This allows us to rule out a *positive* effect beyond 1.7% with 95% confidence.

Columns 3-4 examine the differential impact of testing by manager exception rate. The coefficient on testing gives the impact of testing for the manager with the mean exception rate. In the baseline specification, we find that testing improves output per hour by 0.42 or nearly 10% of a standard deviation. This effect is smaller in magnitude than the impact on duration, and is not statistically significant. Examining the coefficient on the interaction, we again find modestly sized and insignificant coefficients. With full controls, the point estimate translates to a 3.4% smaller impact of testing for a one standard deviation higher manager. We can rule out positive effects outside of 1.9%.

Because of the noise in our estimates, we do not view these results as strong evidence that exceptions are associated with decreases output per hour. However, in all cases, we find no evidence that managerial exceptions *improve* output per hour. This is inconsistent with a model in which managers optimally sacrifice job tenure in favor of workers who perform better on other quality dimensions.

6 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. Viewed in light of our model, these results suggest that managers may be exhibit-

ing bias or poor judgement. In this case, firms may be able to improve worker retention by relying more on test recommendations.

There are several caveats one must keep in mind in interpreting our results. First, while our results suggest that firms may benefit from limiting discretion (relative to the status quo), we do not claim that eliminating discretion or eliminating HR managers is the optimal hiring strategy for a firm. For example, intermediate policies such as restricting the frequency of exceptions may yield better outcomes. 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.

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.[40]

Further, the HR managers we study do not supervise applicants after they are hired. Managers may have more opportunities to correct their mistaken beliefs in settings where they interact with applicants on the job. In such settings, there may also be a manager-employee match component that makes managerial discretion more useful. An additional contribution of our paper is that we present a way to assess the potential for managerial discretion to improve worker outcomes using only data that would readily be available for many firms using workforce analytics.

More broadly, our findings highlight the role new technologies 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 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.

[40]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.

References

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

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

[3] Alonso, Ricardo and Niko Matouschek (2008), "Optimal Delegation," *Review of Economic Studies*, 75(1): pp 259-3.

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

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

[6] Baker, George and Thomas Hubbard (2004), "Contractibility and Asset Ownership: On-Board Computers and Governance in U.S. Trucking," *Quarterly Journal of Economics*, 119(4): pp. 1443-1479.

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

[8] Brown, Meta, Elizabeth Setren, and Giorgio Topa (2015),"Do Informal Referrals Lead to Better Matches? Evidence from a Firm's Employee Referral System," *Journal of Labor Economics*, forthcoming.

[9] Burks, Stephen, Bo Cowgill, Mitchell Hoffman, and Michael Housman (2015), "The Value of Hiring through Employee Referrals," *Quarterly Journal of Economics*, 130(2): pp. 805-839.

[10] Dana, Jason, Robyn Dawes, and Nathaniel Peterson (2013), "Belief in the Unstructured Interview: The Persistence of an Illusion," *Judgment and Decision Making*, 8(5), pp. 512-520.

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

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

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

[14] Frankel, Alexander (2016), "Selecting Applicants," mimeo University of Chicago.

[15] Fernandez, Roberto M., Emilio J. Castilla, and Paul Moore, (2000), "Social Capital at Work: Networks and Employment at a Phone Center," *American Journal of Sociology*, 105(5): pp. 1288-1356.

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

[17] Griliches, Zvi (1957), "Hybrid Corn: An Exploration in the Economics of Technological Change," *Econometrica*, 25(4), pp. 501-522.

[18] Hoffman, Mitchell and Steven Tadelis (2017), "How Do Managers Matter? Evidence from Performance Metrics and Employee Surveys in a Firm," working paper, University of Toronto.

[19] Horton, John (2013), "The Effects of Subsidizing Employer Search," mimeo New York University.

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

[21] Kahn, Lisa B. and Fabian Lange (2014), "Employer Learning, Productivity and the Earnings Distribution: Evidence from Performance Measures," *Review of Economic Studies*, 81(4) pp.1575-1613.

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

[23] Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2017). "Human Decisions and Machine Predictions," NBER Working Paper #23180.

[24] Kuncel, Nathan, David Klieger, Brian Connelly, and Deniz Ones (2013), "Mechanical Versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis," *Journal of Applied Psychology.* Vol. 98, No. 6, 1060–1072.

[25] Kuziemko, Ilyana (2013), "How Should Inmates Be Released from Prison? an Assessment of Parole Versus Fixed Sentence Regimes," *Quarterly Journal of Economics.* Vol. 128, No. 1, 371–424.

[26] Lazear, Edward, Kathryn Shaw, and Christopher Stanton (2015), "The Value of Bosses," *Journal of Labor Economics*, forthcoming.

[27] Li, Danielle. (Forthcoming 2017), "Expertise and Bias in Evaluation: Evidence from the NIH" *American Economic Journal: Applied Economics*

[28] Li, Danielle and Leila Agha. (2016), "Big Names or Big Ideas: Do Peer Review Panels Select the Best Science Proposals?" *Science*, Vol. 348 no. 6233 pp. 434-438.

[29] Pallais, Amanda and Emily Sands (2015), "Why the Referential Treatment? Evidence from Field Experiments on Referrals," *The Journal of Political Economy*, forthcoming.

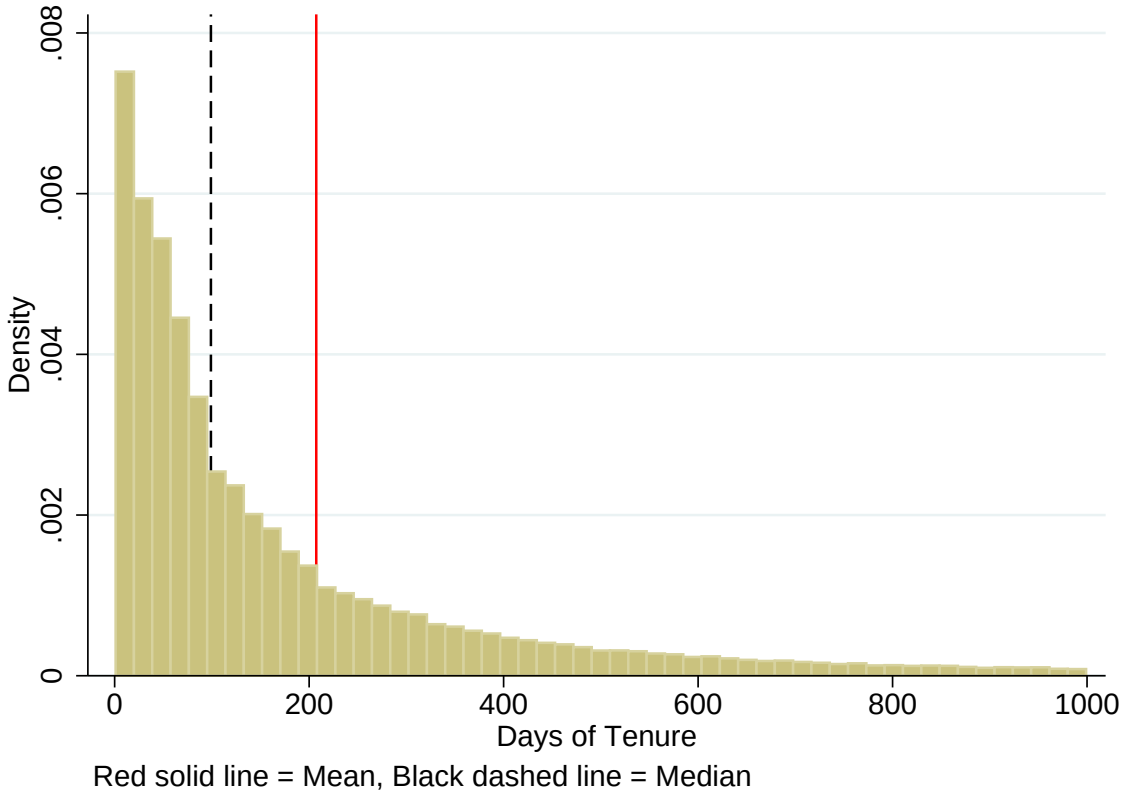
[30] Paravisini, Daniel and Antoinette Schoar (2013) "The Incentive Effect of IT: Randomized Evidence from Credit Committees" NBER Working Paper #19303.

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

[32] Stanton, Christopher and Catherine Thomas (2014), "Landing The First Job: The Value of Intermediaries in Online Hiring," mimeo London School of Economics.

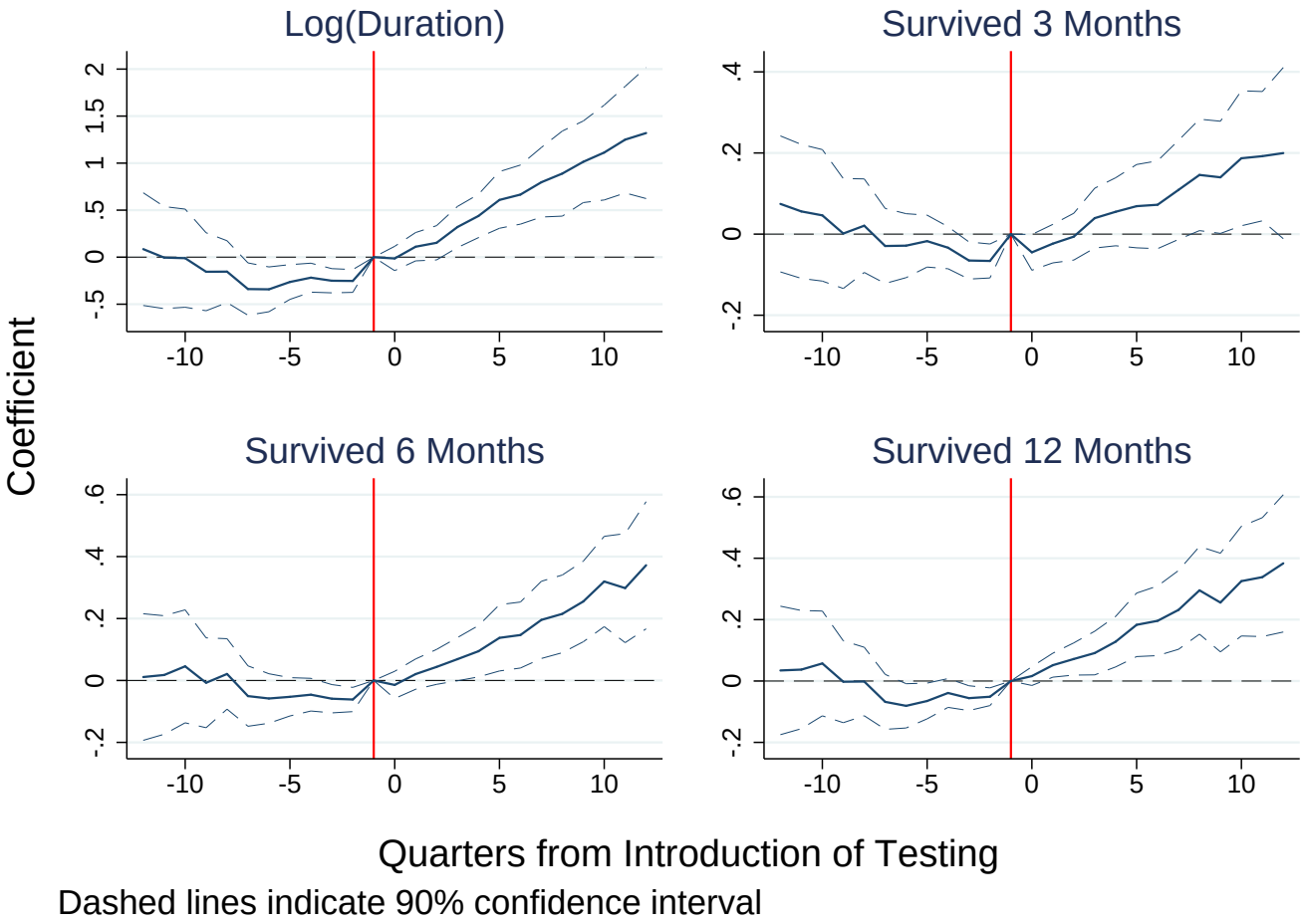
[33] Wang, James (2014), "Why Hire Loan Officers? Examining Delegated Expertise," mimeo University of Michigan.

FIGURE 1: DISTRIBUTION OF LENGTH OF COMPLETED JOB SPELLS



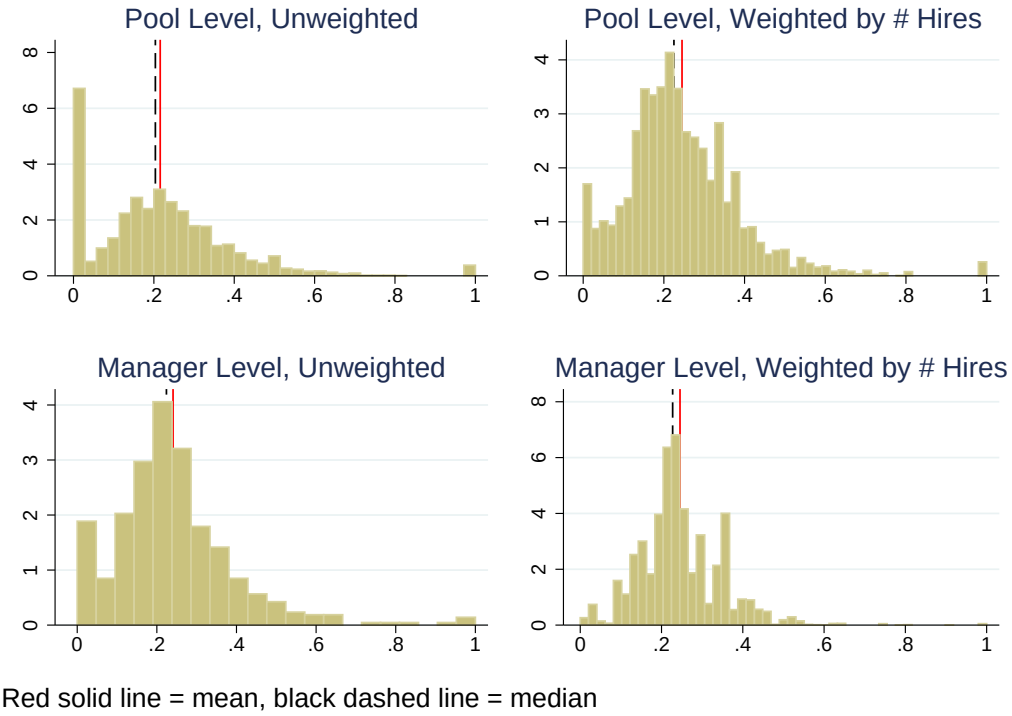
NOTES: Figure 1 plots the distribution of completed job spells at the individual level. For legibility, this histogram (though not the computed mean or median) omits 3% of observations with durations over 1000 days.

FIGURE 2: EVENT STUDY OF DURATION OUTCOMES



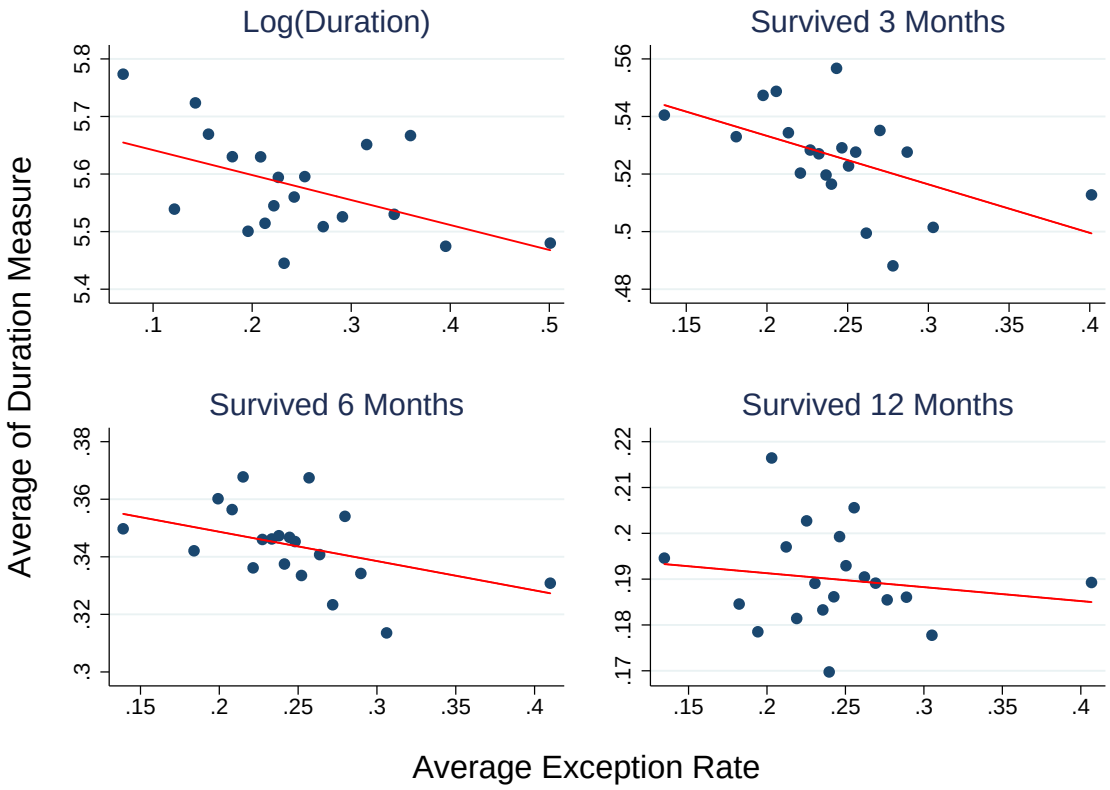
NOTES: These figures plot the impact of testing on worker durations as a function of event-time (in quarters) relative to testing adoption, adjusting for base controls. The underlying estimating equation is given by $\text{Outcome}_{ilt} = \alpha_0 + I_{lt}^{\text{time since testing}}\alpha_1 + controls + \epsilon_{ilt}$, where $I^{\text{time since testing}}$ is a vector of event-time dummies in quarters, with the omitted category, -1, indicated with the vertical red 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 end of our data. Dashed lines indicate the 95% confidence interval. 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 3: DISTRIBUTIONS OF APPLICATION POOL EXCEPTION RATES



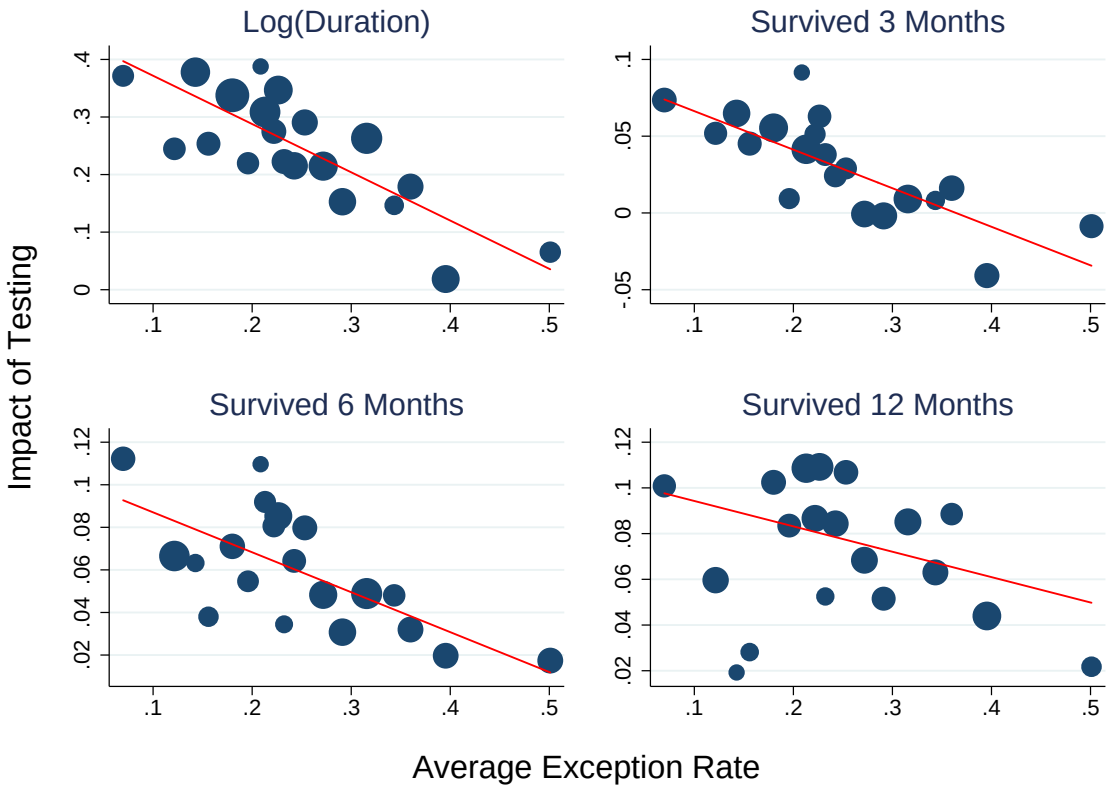
NOTES: These figures plot the distribution of the exception rate, as defined by Equation (2) in Section 5. The top panel presents results at the applicant pool level (defined to be a manager–location–month). The bottom panel aggregates these data to the manager level. Figures on the left define the distribution 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 4: MANAGER-LEVEL EXCEPTION RATES AND POST-TESTING JOB DURATIONS



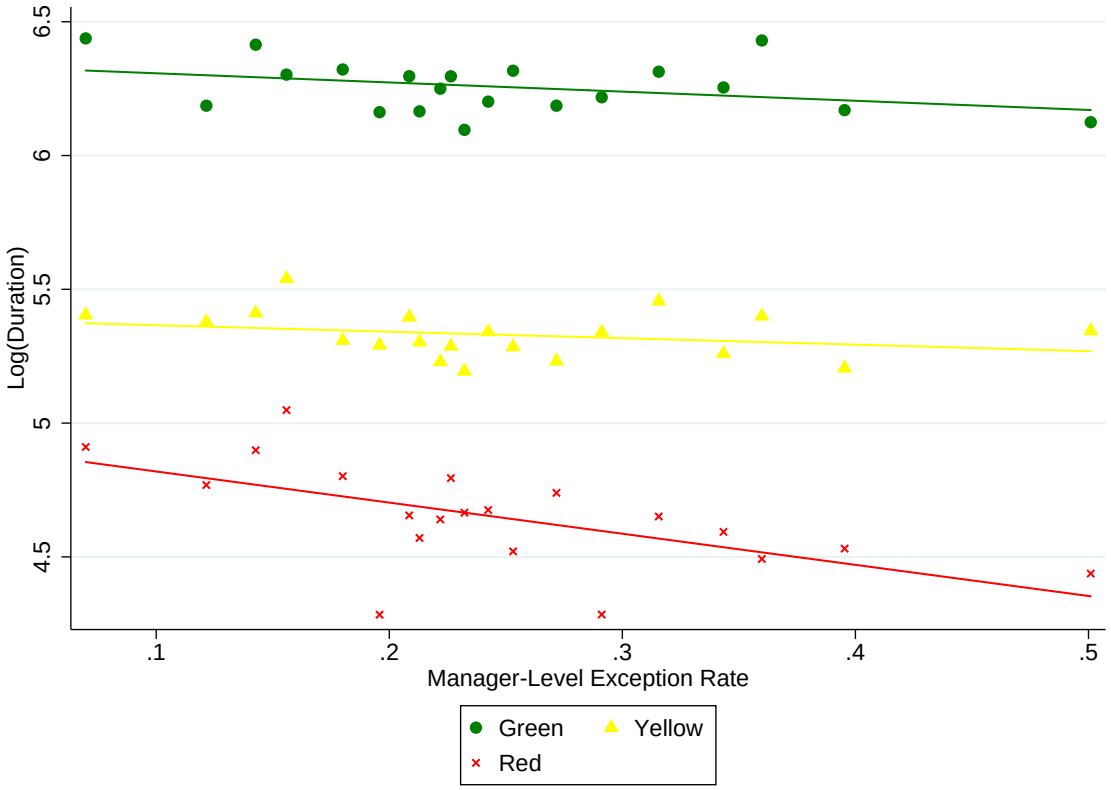
NOTES: We plot average durations (post-testing) and exceptions rates within 20 equally sized bins, weighted by number of hires, based on the average manager-level exception rate. The x-axis represents the average exception rate within each bin. The y-axis is the mean duration outcome in the specified bin. We control for location, hire month, and position fixed effects. Means for the top left panel are 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 end of our data.

FIGURE 5: MANAGER-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 exception rate, on the average manager-level exception rate in each bin. Estimates include our full set of controls: location, hire month, and position fixed effects, client-by-year effects, local unemployment rates, and location time trends. Both the scatter plot and the best linear fit line are weighted by the inverse variance of the estimated coefficients. 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 the end of our data.

FIGURE 6: 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, separately by color score. We plot the coefficients on the exception rate bins as well as the line of best fit.

TABLE 1: 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		0.46	0.33	0.21
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.

TABLE 2: 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<0.1, ** 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 3: Exception Rates and Post-Testing Duration

	Dependent Variable: Log(Duration)				
	(1)	(2)	(3)	(4)	(5)
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<0.1, ** p<0.05, * p<0.1

Notes: This table reports censored normal regressions (see Table 2 for details) of 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 possible in that applicant pool exceptions. 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 4: THE IMPACT OF TESTING BY EXCEPTION RATE

	Dependent Variable: Log(Duration)			
	(1)	(2)	(3)	(4)
Post-Testing	0.385***	0.259**	0.266***	0.251***
	(0.122)	*(0.113)*	*(0.0734)*	*(0.0596)*
Exception Rate Post-Testing*	-0.105**	-0.114***	-0.117***	-0.101***
	(0.0517)	*(0.0373)*	*(0.0355)*	*(0.0288)*
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<0.1, ** p<0.05, * p<0.1

NOTES: This table reports censored normal regressions of the differential impact of testing-adoption, by exception rate. We use the same sample as defined by the notes to Tables 2. Exception rates are defined as in the notes to Table 3.

TABLE 5: 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<0.1, ** 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 6: TESTING, EXCEPTION RATES, AND OUTPUT PER HOUR

Dependent Variable: Output per Hour				
	(1)	(2)	(3)	(4)
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.416	0.179
			(0.370)	*(0.315)*
*Exception Rate*Post-Testing*	-0.0659	-0.111	0.00223	-0.160
	(0.134)	*(0.0953)*	*(0.125)*	*(0.125)*
N	28,858	28,858	62,421	62,421
Base Controls	X	X	X	X
Full Controls		X		X

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

NOTES: See notes to Tables 3 and 4. The dependent variable in this case is output per hour and regressions are estimated with OLS. 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. For the post-testing sample regressions (columns 1 and 2), we also include applicant pool controls.

A Data Appendix

The 15 client firms in our sample have each purchased testing services from our data provider. In this section, we describe the introduction of testing across locations within these client firms. We first provide some details about the test itself. 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. We also describe sample coverage over time across client firms and show robustness to using more balanced panels. We then explore the characteristics of locations that adopt testing early vs. later. We further provide a discussion of heterogeneity in test accuracy across locations. Finally, we provide details on sample restrictions and additional details about the data set.

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; if this were a standardized test grading job, it may present questions about various student answers; if this were a call center job, it may present some call scenarios. 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; if this were test grading the applicant may be given an answer key and asked to grade a sample exam; if this were a call center job, the applicant may listen to a taped conversation before recommending a customer response.

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.[41]

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.[42] This supports our defining test-adoption as the first date 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 2 and 4 for the introduction of testing (top panel) and the differential impact of testing across exception rates (bottom panel).[43] 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.

[41] Recruiters also 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 test score. 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).

[42] 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).

[43] Results from Table 3 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.[44]

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 2 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

[44] 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 regarding 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 important for managers to follow test recommendations rather than make exceptions at these firms.

Given that some of the firms in our sample had some other form of pre-hire testing, our empirical results should thus be interpreted as suggesting (under our identifying assumptions) that our sample firms can improve outcomes by following the recommendations of this particular test.

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.[45] Among client firms with more than one location (11 out of 15 locations, accounting for 94% of hires in our data), 80% adopt testing across all their locations in under 2 years, 50% in under one year. 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 3 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

[45] 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.

possible drivers of testing adoption, including the availability/"bandwith" of managers to oversee the adoption of testing, geographic considerations, 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 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.[46] 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 choose the state-level unemployment rate for the education group most representative of workers in our sample (a high school diploma), but the graph looks similar for unemployment rates for other groups.[47]

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 discernible pattern for any of 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

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

[47]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 we cannot pinpoint finer geography.

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.

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.

Figure 6 already speaks to this concern by showing that color quality is roughly equally predictive of job durations across managers with different exception rates. Appendix Figure A6 provides more information along these lines. Here 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. Specifically, we divide locations into 20 equally sized bins (based on number of hires post-testing). We then estimate censored normal regressions of job duration on an exhaustive set of 20 indicators for bin, controlling for hire month and position type fixed effects, separately by color score.[48] All observations are restricted to the post-testing period when we observe color score.

For each characteristic reported in Appendix Figure A6, we find that color score is strongly predictive of job durations, regardless of which bin the location falls in. 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. This is indicated by the fact that the lines do not intersect and, for each bin, average job durations are generally stacked in order by color score.

We reach a similar conclusion regardless of which location characteristic we examine: we cannot reject that color score is equally predictive of worker duration across all the location characteristics we examine.

A.6 Sample Restrictions

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

[48] Location fixed effects are not included as they are collinear with the location characteristics.

1. We drop one third of applicants because they have a missing identifier for their HR manager.[49]

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 3 on the impact of testing (which do not use exception rates) are similar without the restrictions. We use all observations in the pre-testing period regardless of whether they are associated with locations that meet the post-testing criteria. These locations help identify cohort, client, and position controls. However, results are nearly identical if we drop them. Finally, for all analyses, we drop the four locations (reflecting 0.04% of remaining hires) with less than 50 hires over the sample period.

A.7 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

[49] 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.

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.[50] 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.[51]

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 firm that provided pre-testing data on new hires going back to Jan. 2010. For this firm, we would observe the full set of workers hired at each date after Jan. 2010, but for those hired before, we would only observe the subset 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.

Productivity. In addition to hire and termination dates, which we use to calculate duration, some client firms provide data on output per hour. This is available for about a quarter of hired workers in our sample, and is mentioned by our data firm in its advertising, alongside duration. We trim instances where average transaction time in a given day is less than 60 seconds.[52]

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.[53]

Roughly one-quarter of applicants have one Red/Yellow/Green score, roughly half have two scores, and roughly one quarter have more than two scores. 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 coefficient of 0.88 (for Red=0, Yel-

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

[51] Specifically, we estimated base specifications of Tables 2,3, and 4 excluding each firm one by one.

[52] This is about one percent of transactions. Our results are stronger if we do not trim. 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).

[53] For 1 of the 15 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.

low=1, Green=2).[54] Our focus on the maximum of a scores is thus without much loss of generality.[55]

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

Other managers may take part in hiring decisions as well. One firm said that its HR managers will often endorse candidates to another manager (e.g., a manager in operations one rank above the frontline supervisor) who will 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, ethnographic work suggests that HR managers play an active role in hiring; for example, in a study 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.[56] Most managers are primarily associated with one location, but some are at 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 race, suggesting that changes in worker demographics such as race are not the mechanism by which job testing improves durations.

[54] This is the correlation in the raw data before any data restrictions.

[55] Applicants may be considered for multiple positions so it would be difficult to discern which is the most relevant score for a given applicant.

[56] 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.

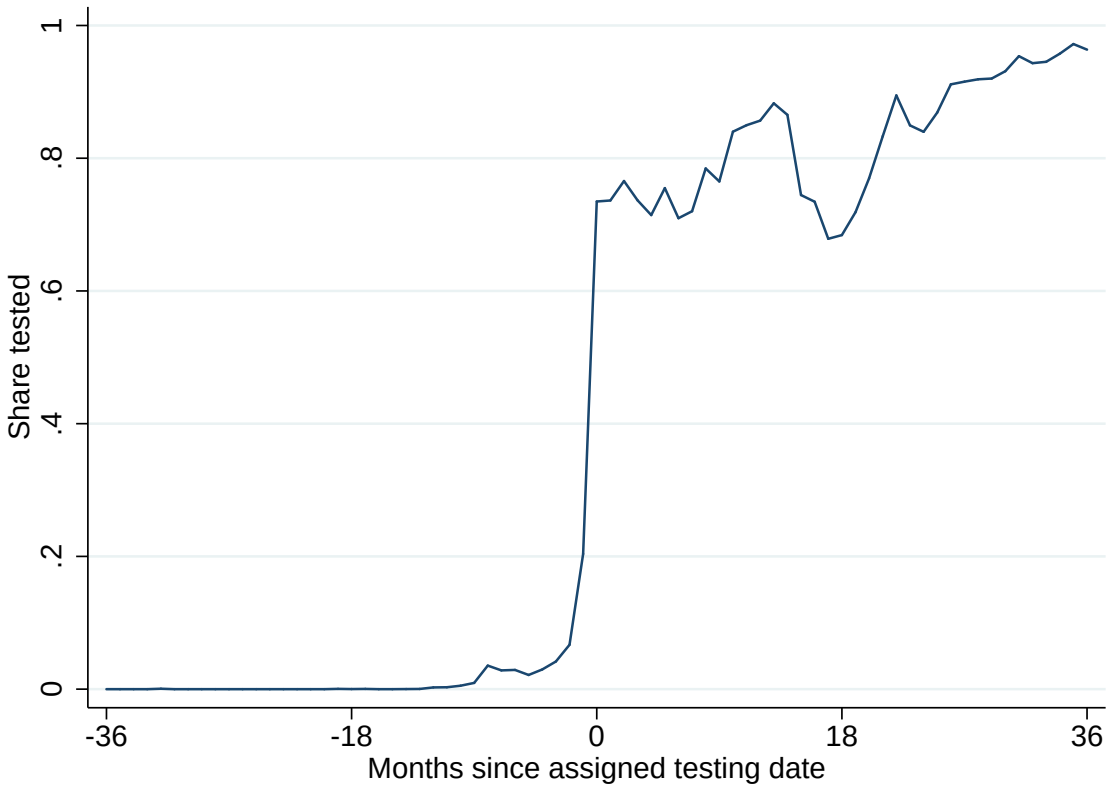
Location Identifiers. In our dataset, we do not have a common identifier for workplace location for workers hired in the pre-testing period and applicants applying post-testing. Consequently, we develop a crosswalk between anonymized location names (used for workers in the pre-testing period) and the location IDs in the post-testing period. We drop workers from our sample where the merge did not yield a clean location variable.[57]

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 in 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 would not be of use for this project.

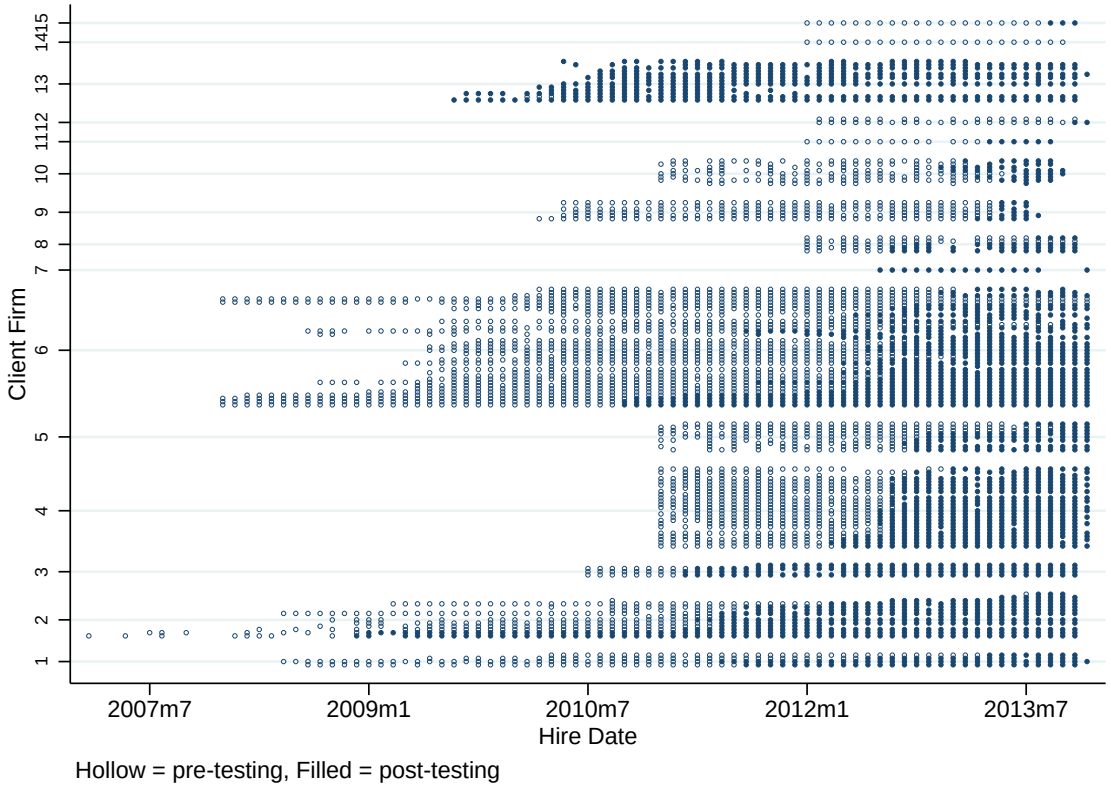
[57] This includes some locations in the pre-testing data where testing is never later introduced.

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



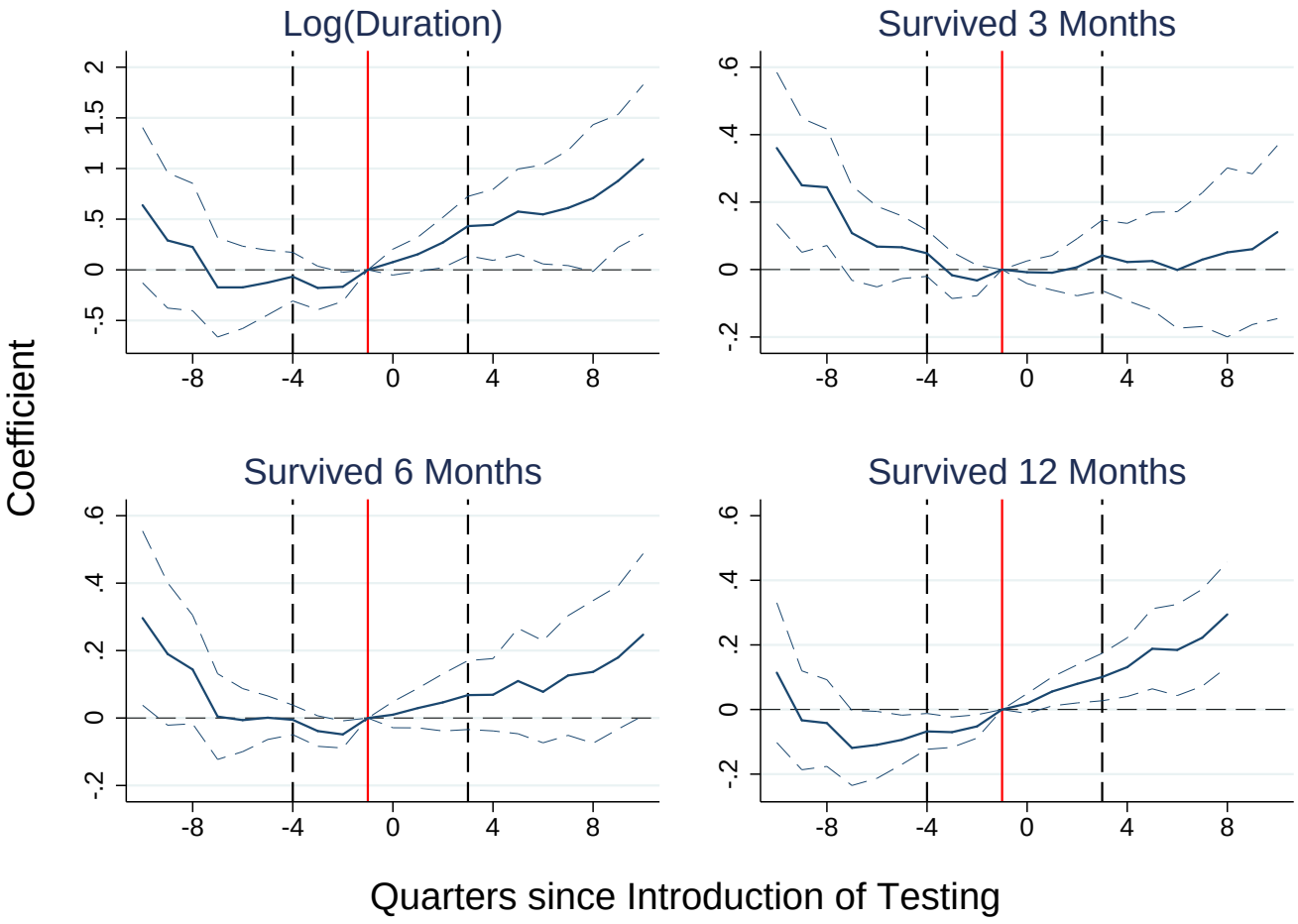
Notes: Figure A1 plots the share of hired workers with a test score as a function of time since the location-specific assigned testing date, averaged across locations. The testing date is defined at the location-month level as the first month in which the modal hire is tested. This graph is restricted 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



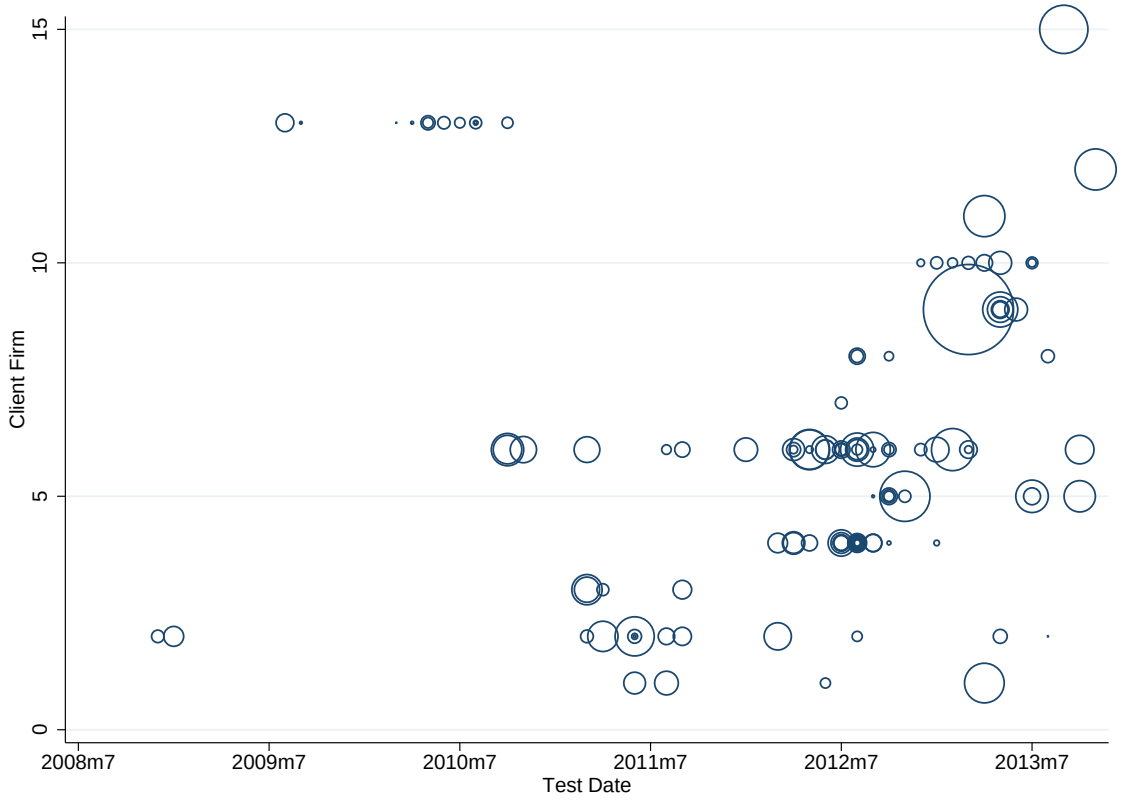
Notes: Locations are lined up on the y-axis, grouped by client firm. Dots indicate that the location hired in a given month, while a gap means no hires were made that month. Filled circles refer to periods after testing is adopted, using our definition (the modal hire was tested), while hollow circles 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



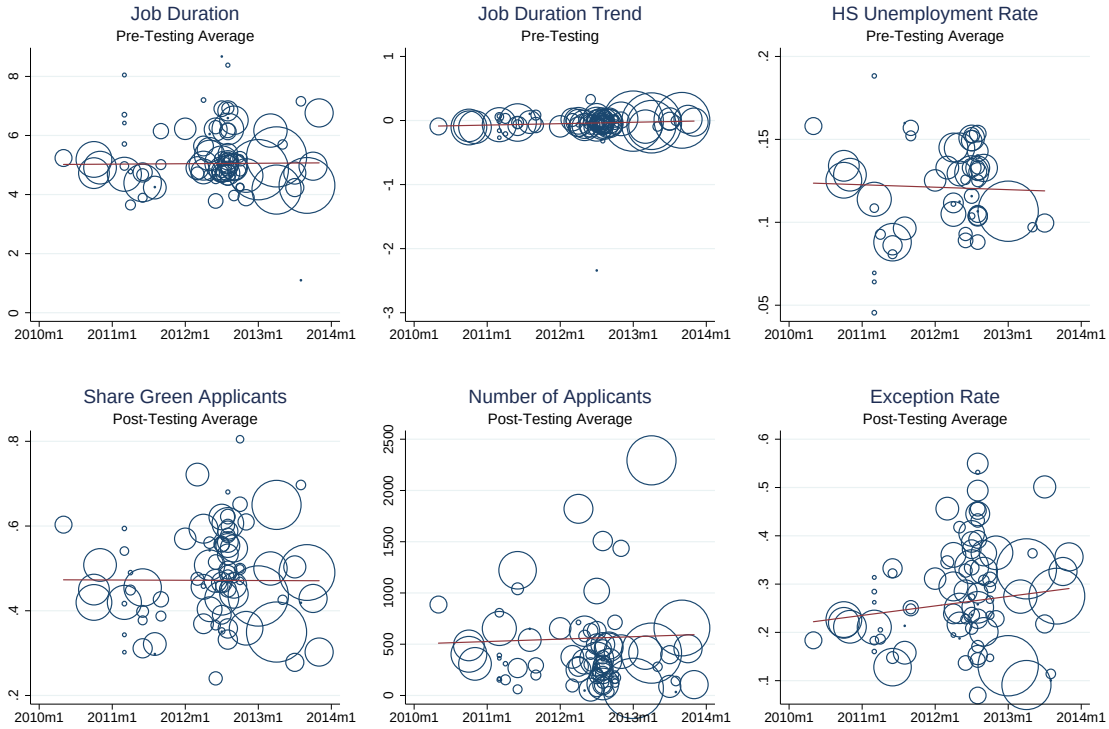
Notes: See notes to Figure 2 of the main text. The sample is restricted to locations with observations in each quarter from 4 lags before testing to 4 leads after (indicated with vertical dashed lines). The graph window is restricted to 10 quarters before and after testing. Dashed lines indicate the 95% confidence interval.

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



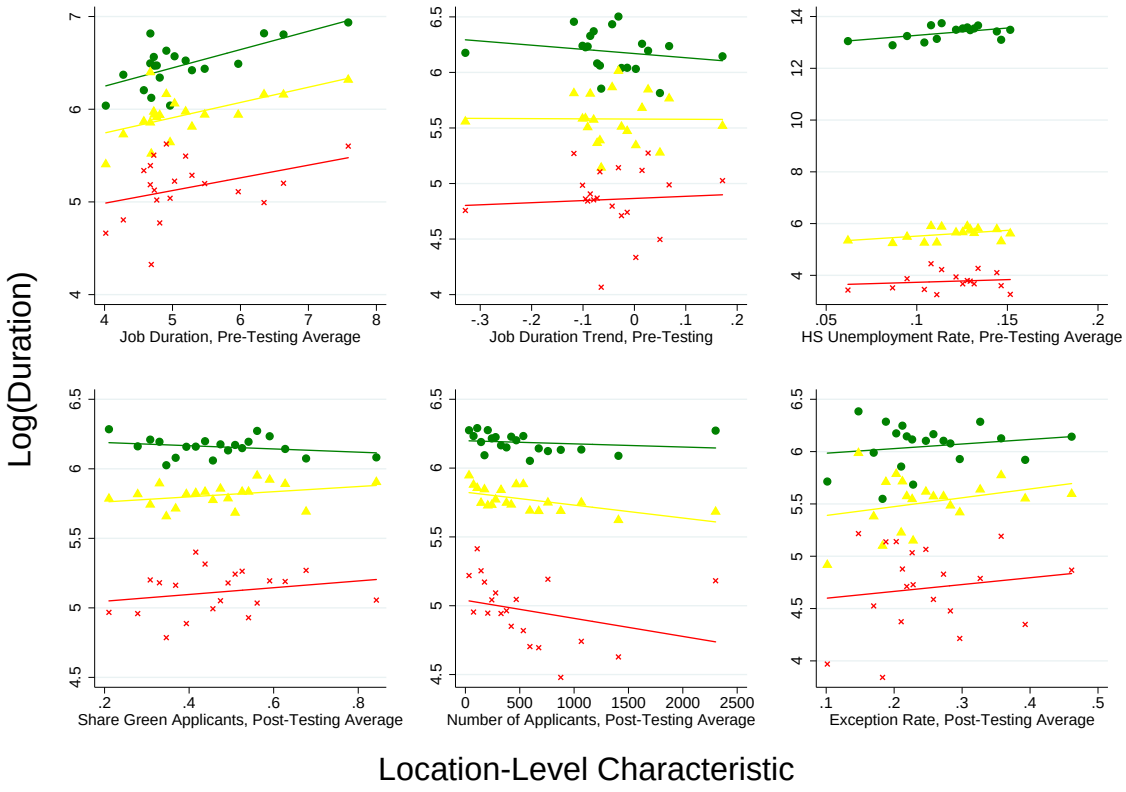
Notes: Figure A4 plots location-specific assigned testing dates on the x-axis, organized by client firm on the y-axis. Circles are weighted by location size, as defined by the number of workers currently employed on July, 2013. As noted in Figure A2, Firm 14 does not appear on the graph because it does 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. 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. The remaining variables are raw averages at the location-level either pre- (top right) or post- (bottom panels) testing.

FIGURE A6: LOCATION OBSERVABLES AND THE COLOR SCORE-JOB DURATION RELATIONSHIP



NOTES: See notes to Figure 6 of the main text. 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, 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	0.385***	0.342***	0.341***	0.323**	0.305***	0.572***
	(0.122)	*(0.124)*	*(0.122)*	*(0.156)*	*(0.111)*	*(0.206)*
*Exception Rate*Post-Testing*	-0.105**	-0.0981*	-0.0955*	-0.128**	-0.156***	-0.435**
	(0.0517)	*(0.0547)*	*(0.0543)*	*(0.0547)*	*(0.0413)*	*(0.215)*
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<0.1, ** p<0.05, * p<0.1

Notes: This table reports censored normal regressions with standard errors clustered at the location level. Column 1 reproduces baseline specifications from Tables 2 (top panel) and 4 (bottom panel). 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 any form of testing before partnering with our data firm. Base controls include location, hire month, and position fixed effects.

B Proofs

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

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

Thus $E[a|s,t]$ is normally distributed with known parameters. Also, since $s|t$ is normally distributed and the assessment of a conditional on s and t is normally distributed, the assessment of a unconditional on s (but still conditional on t) is also normally distributed with a mean μ_t and variance $\sigma = \frac{(\sigma_a^2)^2}{\sigma_\epsilon^2 + \sigma_a^2}$. Finally, define U_t as the manager's utility for a given applicant, conditional on t. The distribution of U_t unconditional on the signals and b, follows a normal distribution with mean $(1-k)\mu_t$ and variance $(1-k)^2\sigma + k^2\sigma_b^2$.

Thus, the probability of being hired is as follows, where $\tilde{z_t} = \frac{\underline{u} - (1-k)\mu_t}{\sqrt{(1-k)^2\sigma + k^2\sigma_b^2}}$.

$$W = p_G(1 - \Phi(\tilde{z_G})) + (1 - p_G)(1 - \Phi(\tilde{z_Y})) \tag{5}$$

The firm is interested in expected quality conditional on being hired under Discretion. This can be expressed as follows, where $\lambda(.)$ is the inverse Mills ratio of the standard normal and $z_t(b_i) = \frac{\frac{u - kb_i}{1-k} - \mu_t}{\sigma}$, i.e., the standard-normalized cutpoint for expected quality, above which, all applicants with b_i will be hired.

$$E[a|Hire] = E_b[p_G(\mu_G + \lambda(z_G(b_i))\sigma) + (1 - p_G)(\mu_Y + \lambda(z_Y(b_i))\sigma)] \tag{6}$$

Inside the expectation, $E_b[]$, we have the expected value of a among all workers hired for a given b_i. We then take expectations over b.

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.

Proposition B.1 *The following results 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 $k > k'$.*

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 value 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, worker quality is higher under Discretion than No Discretion.*

Proposition B.1 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 pushes 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 when the precision of a manager's private information becomes sufficiently small, firms cannot benefit from granting discretion, even if the manager has a low level of bias. 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.

B.1.1 Proof of Proposition B.1

For this proof we make use of the following lemma:

Lemma B.2 *The expected quality of hires for a given manager, $E[a|Hire]$, is decreasing in managerial bias, k.*

Proof A manager will hire all workers for whom $(1-k)E[a|s_i, t_i] + kb_i > \underline{u}$, i.e., if $b_i > \frac{\underline{u} - (1-k)E[a|s_i,t_i]}{k}$. Managers trade off b for a with slope $-\frac{1-k}{k}$. Consider two managers, Manager 1 and Manager 2, where $k_1 > k_2$, i.e., Manager 1 is more biased than Manager 2. Manager 2 will have a steeper (more negative) slope ($\frac{1-k_2}{k_2} > \frac{1-k_1}{k_1}$) than Manager 1. There will thus be some cutoff $\hat{a}$ such that for $E[a|s_i, t_i] > \hat{a}$ Manager 2 has a lower cutoff for b and for $E[a|s_i, t_i] < \hat{a}$, Manager 1 has a lower cutoff for b.

That is, some candidates will be hired by both managers, but for $E[a|s_i, t_i] > \hat{a}$, Manager 2 (less bias) will hire some candidates that Manager 1 would not, and for $E[a|s_i, t_i] < \hat{a}$ Manager 1 (more bias) will hire some candidates that Manager 2 would not. The candidates that Manager 2 would hire when Manager 1 would not, have high expected values of a, while

the candidates that Manager 1 would hire where Manager 2 would not have low expected values of a. Therefore the average a value for workers hired by Manager 2, the less biased manager, must be higher than that for those hired by Manager 1. $E[a|Hire]$ is decreasing in k.

We next prove each item of Proposition B.1

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 $k = 0$, the manager hires only applicants whose expected quality, a, is above the threshold. In this case, the firm will at least weakly prefer Discretion. Since the manager's preferences are perfectly aligned, he or she will always do at least as well as hiring only type G.

Thus, Discretion is better than No Discretion for $k = 0$ and the opposite is true for $k = 1$. Lemma B.2 shows that the firm's payoff is decreasing in k. There must therefore be a single cutpoint, k', where, below that point, the firm's payoff for Discretion is large 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 When $1/\sigma_\epsilon^2 = 0$, i.e., the manager has no information, and $k = 0$, he or she will hire based on the test, resulting in an equal payoff to the firm as No Discretion. For all $k > 0$, the payoff to the firm will be worse than No Discretion, thanks to lemma B.2. Thus when the manager has no information the firm prefers No Discretion to Discretion.

We also point out that the firm's payoff under Discretion, expressed above in equation (6), is clearly continuous in σ (which is continuous in $1/\sigma_\epsilon^2 = 0$).

Thus, when the manager has 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 value 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, worker quality is higher under Discretion than No Discretion.*

Proof First, we point out that when $k = 0$, the firm's payoff under Discretion is increasing in $1/\sigma_\epsilon^2$. An unbiased manager will always do better (from the firm's perspective) with more information than less. Second, we have already shown that for $k = 0$, Discretion is always preferable to No Discretion, regardless of the manager's information, and when σ_ϵ^2 approached ∞, there is no difference between Discretion and No Discretion from the firm's perspective.

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). We know that $\Delta(\sigma_\epsilon^2, 0)$ is positive and decreasing in σ_ϵ^2, and approaches 0 as σ_ϵ^2 approaches ∞. Also, since the firm's payoff under discretion is continuous in both k and $1/\sigma_\epsilon^2$ (see Equation (6) above), $\Delta()$ must also be continuous in these variables.

Fix any $\overline{\rho}$ and let $\overline{\sigma_\epsilon^2} = 1/\overline{\rho}$. Let $y = \Delta(\overline{\sigma_\epsilon^2}, 0)$. We know that $\Delta(\sigma_\epsilon^2, 0) > y$ for all $\sigma_\epsilon^2 < \overline{\sigma_\epsilon^2}$.

Let $d(k) = \max_{\sigma_\epsilon^2 \in [0, \overline{\sigma_\epsilon^2}]} \Delta(\sigma_\epsilon^2, k) - \Delta(\sigma_\epsilon^2, 0)$. We know $d(k)$ exists because $\Delta()$ is continuous wrt σ_ϵ^2 and the interval over which we take the maximum is compact. We also know that $d(0) = 0$, i.e., for an unbiased manager, the return to discretion is maximized when managers have full information. Finally, $d(k)$ is continuous in k because $\Delta()$ is.

Therefore, we can find $k'' > 0$ such that $d(k) = d(k) - d(0) < y$ whenever $k < k''$. This means that $\Delta(\sigma_\epsilon^2, k) > 0$ for $\sigma_\epsilon^2 < \overline{\sigma_\epsilon^2}$. In other words, at bias k and $\rho > \underline{\rho}$, Discretion is better than No Discretion.

B.2 Proof of Proposition 4.1

The exception rate, R_m, is increasing in both managerial bias, k, and 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.

Above, we defined U_t, a manager's utility of a candidate conditional on t, and showed that it is normally distributed with mean $(1 - k)\mu_t$ and variance $\Sigma = (1 - k)^2\sigma + k^2\sigma_b^2$. 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. This rescaling of course does nothing to the cutoff $\underline{u}$, and 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 $\tilde{U_Y}$: $\Delta_U = \frac{(1-k)(\mu_G - \mu Y)}{\sqrt{\Sigma}}$.

It is easy to show (with some algebra) that $\frac{\partial \Delta U}{\partial k} = \frac{-(\mu_G - \mu_Y)\sigma_b^2}{\Sigma^{3/2}}$, which is clearly negative. 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, it is each to 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. 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.

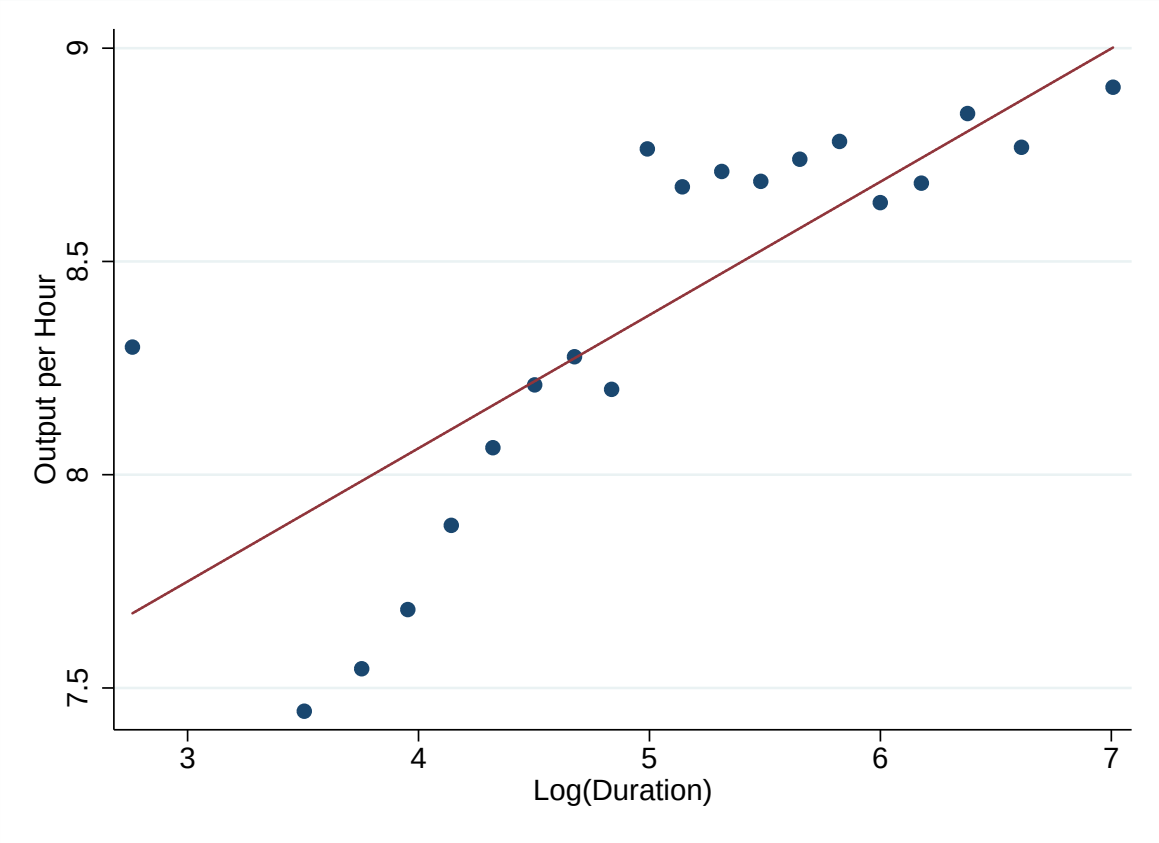
B.3 Proof of Proposition 4.2

If the quality of hired workers is decreasing in the exception rate, $\frac{\partial E[a|Hire]}{\partial R_m} < 0$, then firms can improve outcomes by eliminating discretion. If quality is increasing in the exception rate then Discretion is better than No Discretion.

Proof Consider a manager who makes no exceptions even when given discretion: Across a large number of applicants, this only occurs if this manager has no information and no bias. Thus the quality of hires by this manager is the same as that of hires under a no discretion regime, i.e., hiring decisions made solely on the basis of the test. Compare outcomes for this manager to one who makes exceptions. If $\frac{\partial E[a|Hire]}{\partial R_m} < 0$, then the quality of hired workers for the latter manager will be worse than for the former. Since the former is equivalent to hires under no discretion, it then follows that the quality of hires under discretion will be lower than under no discretion. If the opposite is true and the manager who made exceptions, thereby wielding discretion, has better outcomes, then discretion improves upon no discretion.

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.

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						
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 Exception Rate						
Post-Testing	0.0469**	0.0310	0.0955**	0.0625***	0.108***	0.0768***
	(0.0220)	*(0.0192)*	*(0.0373)*	*(0.0223)*	*(0.0370)*	*(0.0197)*
*Exception Rate*Post-Testing*	-0.0291**	-0.0291***	-0.0256*	-0.0196***	-0.0250	-0.0118
	(0.0127)	*(0.00727)*	*(0.0145)*	*(0.00611)*	*(0.0175)*	*(0.00777)*
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<0.1, ** p<0.05, * p<0.1

Notes: See notes to Tables 2, 3, and 4 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 end of data collection. 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.

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<0.1, ** 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			0.384***	0.285***
			(0.123)	*(0.0605)*
*Exception Rate*Post-Testing*	-0.112***	-0.111***	-0.155**	-0.121***
	(0.0355)	*(0.0303)*	*(0.0723)*	*(0.0297)*
N	76,425	76,425	250,754	250,754
Base Controls	X	X	X	X
Full Controls		X		X

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

Notes: See notes to Tables 3 and 4 in the main text. 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).

Appendix Table C4: 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			0.389***	0.254***
			(0.124)	*(0.0597)*
*Exception Rate*Post-Testing*	-0.0730**	-0.0635**	-0.124**	-0.0940***
	(0.0327)	*(0.0258)*	*(0.0570)*	*(0.0266)*
Exception Score Relative to Max Score				
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.377***	0.237***
			(0.123)	*(0.0654)*
*Exception Rate*Post-Testing*	-0.0237	-0.0707***	-0.0621	-0.0166
	(0.0261)	*(0.0190)*	*(0.0510)*	*(0.0253)*
Exception Score Relative to Random Score				
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.394***	0.242***
			(0.128)	*(0.0620)*
*Exception Rate*Post-Testing*	-0.0585	-0.0149	-0.230	-0.0762**
	(0.0364)	*(0.0241)*	*(0.160)*	*(0.0381)*
N	91,319	91,319	265,648	265,648
Base Controls	X	X	X	X
Full Controls		X		X

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

Notes: See Tables 3 and 4 in the main text. 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).