Discretion in Hiring

Mitchell Hoffman
University of Toronto

Lisa B. Kahn
Yale University & NBER

Danielle Li
Harvard University

July 3, 2015

PRELIMINARY & INCOMPLETE

Please do not cite or circulate without permission

Abstract

Who should make hiring decisions? Many firms rely on hiring managers to evaluate applications and make job offers. These managers may be informed about a worker's quality, but their efficacy may be undermined by biases or bad judgement. The use of quantitative metrics such as job testing enables firms to limit these concerns, but potentially at the cost of ignoring valuable information. In this paper, we evaluate the staggered introduction of a job test across 131 locations of 15 firms employing low-skill service sector workers. We show that testing improves the match quality of hired workers, as measured by their completed tenure, by about 15%. We then propose an empirical test for assessing whether firms should rely on hard metrics or grant managers discretion in making hiring decisions. When applied to our setting, we find that firms can improve worker quality by limiting managerial discretion. This is because, when faced with similar applicant pools, managers who exercise more discretion (as measured by their likelihood of overruling the test recommendations) systematically end up with worse hires. This result suggests that managers make exceptions to test recommendations because they are biased, not because they are better informed. Our test can be applied to a broad range of settings to help inform how firms allocate authority.

---

*Correspondence: Mitchell Hoffman, University of Toronto, 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, Ricardo Alonso, David Berger, Arthur Campbell, Alex Frankel, Jin Li, Liz Lyons, Steve Malliaris, Mike Powell, Kathryn Shaw, Steve Tadelis, and numerous seminar participants. 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 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 or referrals, 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 if hires and how should firms use job tests, if at all? 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 he or she has 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 provide a diagnostic to inform how firms should incorporate it into their hiring decisions. Using a unique personnel dataset on HR managers, job applicants, and hired workers across 15 firms that adopt job testing, we present two key findings. First, job testing substantially improves the match quality of hired workers: those hired with job testing have 15% longer tenures than those

[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/.

hired without testing. Second, managers who overrule test recommendations more often hire, on average, workers with lower match quality, as measured by job tenure. This second result suggests that managers are exercising discretion because they are biased or have poor judgement, not because they are better informed. This implies that firms in our setting can further improve match quality by limiting managerial discretion by placing more weight on the test.

Our paper makes the following contributions. First, we provide new evidence that managers systematically make hiring decisions that are not in the interest of the firm. In our setting, this behavior generates increased turnover in an industry in which training and hiring make up a substantial portion of labor costs. Second, we show that job testing can improve hiring outcomes not simply by providing more information, but by making information verifiable, and thereby expanding the scope for contractual solutions to agency problems within the firm. Finally, we develop a simple tractable test for assessing the value of discretion in hiring. Our test uses data likely available to any firm with job testing, and is applicable to a wide variety of settings where at least one objective correlate of productivity is available.

We begin with a model in which firms rely on potentially biased HR managers who observe both public and private signals of worker quality. Using this model, we develop a simple empirical diagnostic based on the following intuition: if managers make exceptions to test recommendations because they have superior private information about a worker's quality, then we would expect better informed managers to both be more likely to make exceptions and to hire workers who are a better fit. As such, a positive correlation between exceptions and outcomes suggests that the discretion granted was valuable. If, in contrast, managers who make more exceptions hire workers with worse outcomes, then it is likely that managers are either biased or mistaken, and firms should limit discretion.

We apply this test using data from an anonymous firm that provides online job testing services to client firms. Our sample consists of 15 client firms who employ low-skill service-sector workers. Prior to the introduction of testing, our sample firms employed HR managers who conducted interviews and made hiring decisions. After the introduction of testing, HR managers were also given access to a test score for each worker: green (high potential

candidate), yellow (moderate potential candidate), or red (lowest rating).[3] In most settings, managers were told to factor the test into their hiring decisions but were still given discretion to use other signals of quality.

First, we estimate the impact of introducing a job test on the match quality of hired workers. By examining the staggered introduction of job testing across our sample locations, we show that cohorts of workers hired with job testing have 15% longer tenures than cohorts of workers hired without testing. We provide a number of tests in the paper to ensure that our results are not driven by the endogenous adoption of testing or by other policies that firms may have concurrently implemented.

This finding suggests that job tests contain valuable information about the match quality of candidates. Next, we ask how firms should use this information, in particular, whether firms should limit discretion and follow test recommendations, or allow managers to exercise discretion and make exceptions to those recommendations. A unique feature of our data is that it allows us to measure the exercise of discretion explicitly: we observe every instance in which a manager hires a worker with a test score of yellow when an applicant with a score of green goes unhired (or similarly, when a red is hired above a yellow or a green). As explained above, the correlation between a manager's likelihood of making these exceptions and eventual outcomes of hires can inform whether the exercise of discretion is beneficial from the firm's perspective. Across a variety of specifications, we find that the exercise of discretion is strongly correlated with worse outcomes. Even when faced with applicant pools that are identical in terms of test scores, managers that make more exceptions systematically hire workers who are more likely to quit or be fired.

An alternative explanation for these findings is that managers sacrifice job tenure in search of workers who have higher quality on other dimensions. For example, interviews may be informative about an applicant's fluency in English, which will impact quality of service but may not be captured in a written test. In this case, eliminating discretion may improve worker durations, but at the expense of other quality measures. To assess whether this is a possible explanation for our findings, we examine the relationship between hiring,

[3]This test is an online assessment gathering information on a number of dimensions, including technical knowledge, personality, cognitive skills, fit for the job, and the ability to address various workplace scenarios. Our data provider then uses a proprietary algorithm to aggregate this information into the rating.

exceptions, and a measure of productivity. For a subsample of our client firms we have daily measures based on the efficiency with which customers are served by individual workers.[4] Based on this supplemental analysis, we see no evidence that firms are trading of duration for higher productivity. Taken together, our findings suggest that firms could improve both match quality and worker productivity by placing more weight on the recommendations of the job test.

As data analytics becomes 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, ours is the first paper to provide a simple tractable test for assessing the value of discretion in hiring.[5] Our findings provide direct evidence that screening technologies can help resolve agency problems by improving information symmetry, and thereby relaxing contracting constraints. 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. It is also related to papers on bias, discretion, and rule-making in other settings.[6]

We also contribute to a small, but growing literature on the impact of screening technologies on the quality of hires.[7] Our paper is most closely related to Autor and Scarborough (2008), which provides the first estimate of the impact of job testing on worker performance. 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. Our paper, by contrast, studies the implications of job testing on the allocation of authority within the firm.

[4]Confidentiality restrictions limit our ability to provide details about this productivity measure. Some examples include: the number of calls telemarketers or customer service agents are able to complete per hour, and the speed at which retail checkout clerks scan items.

[5]See for example, Dessein (2002), Alonso, Dessein, Matouschek (2008), Alonso and Matouschek (2008), and Rantakari (2008).

[6]Paravisini and Schoar (2012) finds that credit scoring technology aligns loan offer incentives and improves lending performance. Li (2012) documents an empirical tradeoff between expertise and bias among grant selection committees.

[7]Other screening technologies include labor market intermediaries (e.g., Autor (2001), Stanton and Thomas (2014), and employee referrals (e.g., Brown et al., (2014), Burks et al. (2013) and Pallais and Sands (2013)).

Our work is also relevant to a broader literature on hiring and employer learning.[8] Oyer and Schaefer (2011) note in their handbook chapter that hiring remains an important open area of research. We point out that hiring is made even more challenging because firms must often entrust these decisions to managers who may be biased or exhibit poor judgment.[9]

Lastly, 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.[10]

The remainder of this paper proceeds as follows. Section 2 describes the setting and data. Section 3 presents a model of hiring with both hard and soft signals of quality. Section 4 evaluates the impact of testing on the quality of hires, and Section 5 evaluates the role of discretion in test adoption. Section 7 concludes.

2 Setting and Data

Firms have increasingly incorporated testing into their hiring practices. One explanation for this shift is that the increasing power of data analytics has made it easier to look for regularities that predict worker performance. We obtain data from an anonymous consulting firm that follows such a model. We hereafter term this firm the "data firm." The data firm offers a test designed to predict performance for a particular job in the low-skilled service sector. To preserve the confidentiality of the data firm, we are unable to reveal the exact nature of the job, but it is conducted in a non-retail environment and is similar to jobs such as data entry work, telemarketing, or standardized test grading. 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.

[8] A central literature in labor economics emphasizes that imperfect information generates substantial problems for allocative efficiency in the labor market. This literature suggests imperfect information is a substantial problem facing those making hiring decisions. See for example Jovanovic (1979), Farber and Gibbons (1996), Altonji and Pierret (2001), and Kahn and Lange (2014).

[9] This notion stems from the canonical principal-agent problem, for instance as in Aghion and Tirole (1997). In addition, many other models of management focus on moral hazard problems generated when a manager is allocated decision rights.

[10] See Kuncel, et. al. 2013 for a meta-analysis of this literature and Kahneman 2011 for a behavioral economics perspective.

The job test consists of an online questionnaire comprising a large battery of questions, including those on personality, cognitive skills, technical skills, and various job scenarios. The data firm then matches these responses with subsequent performance in order to identify the questions or sets of questions that are the most predictive of future workplace success in this setting. These correlations are then aggregated by a proprietary algorithm to deliver a *green, yellow, red* job test score.

In its marketing materials, our data firm emphasizes the ability of its job test to reduce worker turnover, which has been a perennial challenge for firms employing low skill service sector workers. To illustrate this concern, Figure 1 shows a histogram of job tenure for completed spells (75% of the spells in our data) among employees in our sample client firms. The median worker (solid red line) stays only 99 days, or just over 3 months. Twenty percent of hired workers leave after only a month. At the same time, our client firms report spending between 3 and 5 weeks training each new hire, during which time the hire is being paid. As a result, hiring and training make up a substantial fraction of the labor costs in our client firms. Correspondingly, our analysis will also focus on job retention as the primary measure of hiring quality. For a subset of our client firms we also observe a direct measure of worker productivity: customers served per hour. Because these data exist for a much smaller set of workers (roughly a quarter of hired workers), we report these findings separately in Section 6.

Before partnering with our data firm, client firms kept records for each hired worker, consisting of start and stop dates, the reason for the exit, some information about job function, and location.[11] Each client firm shared these records with our data firm, once a partnership was established.[12] From this point, our data firm keeps records of all applicants,

[11] The information on job function is related to the type of service provided by the worker, details of which are difficult to elaborate on without revealing more about the nature of the job.

[12] One downside of the pre-testing data is that they are collected idiosyncratically across client firms. For some clients, we believe we have a stock-sampling problem: when firms began keeping track of these data, they retrospectively added in start dates for anyone currently working. This generates a survivor bias for incumbent workers, relative to new workers. For example, for a firm that began collecting data in January 2010, we would observe the full set of workers hired at each date after January 2010, but for those hired before, we would only observe the subset who survived to January 2010. We do not explicitly observe the date at which the firm began collecting data; instead, we use the date of the first recorded termination as a conservative proxy for when data collection began. 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

their test scores, and the ID of the HR manager responsible for a given applicant, in addition to the personnel records (exactly as described above) for hired workers.

Prior to testing, our client firms gave their managers discretion to make hiring decisions based on interviews and resumes.[13] After testing, firms made scores available to managers and encouraged them to factor scores into hiring decisions, but authority over hiring decisions was still typically delegated to managers.[14]

In the first part of this paper, we examine the impact of testing technology on worker match quality, as measured by tenure. For any given client firm, testing was rolled out gradually at roughly the location level. However, because of practical considerations in the adoption process, not all workers in a given location and time period share the same testing status. That is, in a given month some applicants in a location may have test scores, while others do not.[15] We therefore impute a location-specific date of testing adoption. Our preferred metric for the date of testing adoption is the first date in which at least 50% of the workers hired in that month and location have a test score. Once testing is adopted at a location, based on our definition, we impose that testing is thereafter always available.[16] We also report specifications in which testing adoption is defined as the first month in which any hire has a test score, as well as individual-level specifications in which our explanatory variable is an indicator for whether a specific individual receives testing.

Table 1 provides sample characteristics. Across our whole sample period we have nearly 300,000 hires; two-thirds of these were observed before testing was introduced and one-third were observed after, based on our preferred imputed definition of testing. Once we link applicants to the HR manager responsible for them (only after testing), we have 555 such managers in the data. These managers are primarily responsible for hiring, and are unlikely

our primary sample, but have experimented with including them along with flexible controls for being stock sampled in our regressions.

[13] In addition, the data firm informed us that several client firms had some other form of testing before the introduction of the data firm's test.

[14] We do not directly observe authority relations in our data. However, in surveys that the data firm conducted with a number of the client firms, the client firms reported that managers were not required to hire strictly by the test.

[15] We are told by the data firm, however, that the intention of clients was to bring testing into a location at the same time for everyone in that location.

[16] This fits patterns in the data, for example, that most locations weakly increase the share of applicants that are tested throughout our sample period.

to manage day-to-day production.[17] Post-testing, when we have information on applicants as well as hires, we have nearly 94,000 hires and a total of 690,000 applicants.

We will find it useful to define an "applicant pool" as a group of applicants being considered by the same manager for a job at the same location in the same month. We restrict to months in which at least one worker was hired. We allow non-hired applicants to be under consideration for up to 4 months, from their application date.[18] From Table 1, we have 4,209 such applicant pools in our data consisting of, on average 268 applicants. On average, 19% of workers in a given pool are hired.

Table 1 also shows the distribution of test scores and the associated hire probabilities within an applicant pool. Roughly 40% of all applicants in a given pool receive a "green", while "yellow" and "red" candidates make up roughly 30%, each. The test score is predictive of whether or not an applicant is hired. In the average pool, greens and yellows are hired at a rate of roughly 20%, while only 9% of reds are hired. Still, managers very frequently make exceptions to test recommendations: though not show, we find that in three-quarters of applicant pools, a yellow applicant is hired even though green applicants are not. In nearly a third of the pools, a red was hired when greens or yellows were available. This foreshadows substantial variation in how much managers exercise discretion, which we explore later.

Finally, Table 1 reports worker performance pre- and pos-testing, and by color score. On average, greens stay 12 days (11%) longer than yellows, who stay 17 days (18%) longer than reds. These differences are statistically significant and hold up to the full range of controls described below. On our productivity measure, customers served per hour, which averages roughly 12, greens outperform yellows and reds as well. This provides some evidence that test scores are indeed informative about worker performance. Even among the selected sample of hired workers, better test scores predict better outcomes. We might expect these differences to be even larger in the overall applicant population if managers hire red and yellow applicants only when unobserved quality is particularly high.

.

[17] In some cases, committees or other managers are also involved in hiring decisions.

[18] We observe in our post-testing data that over 90% of hired workers are hired within 4 months of the date they first submitted an application.

3 Model

We formalize a model in which a firm makes hiring decisions with the help of an HR manager. There are two sources of information about the quality of job candidates. First, interviews generate unverifiable information about a candidate's quality that is privately observed by the HR manager. Second, the introduction of job testing generates verifiable information about quality that is observed by both the manager and the firm. Managers then make hiring decisions with the aid of both sources of information.

In this setting, job testing can improve hiring in two ways. First, it can help managers make more informed decisions by providing an additional signal of worker quality. Second, because test information is verifiable, it enables the firm to limit the influence of potentially biased managers by relying more on the test signal. Granting managers discretion enables the firm to take advantage of both interview and test signals, but may also leave it vulnerable to managerial biases. Limiting discretion and relying on the test removes scope for bias, but at the cost of ignoring information.

The following model formalizes this tradeoff and outlines an empirical test of whether firms can improve worker quality by eliminating discretion.

3.1 Setup

A mass one of applicants apply for job openings within a firm. The firm's payoff of hiring worker i is given by the worker's match quality, 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 match quality distribution enables us to naturally incorporate the discrete test score into the hiring environment. We do so by assuming that before testing, both a and t are unobserved. Once testing is introduced, it publicly reveals t.[19] We further assume that the distribution of applicants does not change with the introduction of testing.

[19]For simplicity, we assume the test signal and t are binary, even though in our data the signal can take three possible values. This is without loss of generality for the mechanics of the model.

The firm's objective is to hire a proportion, W, of workers that maximizes expected match quality.[20] For simplicity, we also assume $W < p_G$.[21] To hire workers, the firm must employ HR managers whose interests are imperfectly aligned with that of the firm. In particular, a manager's payoff for hiring worker i is given by:

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

In addition to valuing match quality, managers also receive an idiosyncratic payoff b_i, which they value with a weight k that is assumed to fall between 0 and 1. We assume that $a \perp b$.

The additional quality, b, can be thought of in two ways. First, it may capture idiosyncratic preferences of the manager for workers in certain demographic groups or with similar backgrounds (same alma mater, for example). Second, b can represent beliefs that the manager has about worker quality that are untrue. For example, the manager may genuinely have the same preferences as the firm but draw incorrect inferences from his or her interview.[22]

The manager privately observes information about a_i and b_i. First, for simplicity, we assume that b_i is perfectly observed by the HR manager, and is distributed in the population by $N(0, \sigma_b^2)$ with $\sigma_b^2 \in (0, \infty)$. Second, the manager observes a noisy signal of match quality, s_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 level of 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 $\sigma_\epsilon^2 = \infty$.

[20] A profit maximizing firm will hire all workers whose expected match quality is greater than their cost (wage). As we point out below, the firm cannot contract on the expected value of a_i. One rationale for imposing a fixed share of hires, W is that it is contractible. A firm with rational expectations will know the typical share of applicants that are worth hiring and can impose this share as a rule on managers. Assuming a fixed hiring share is also consistent with the previous literature, for example, Autor and Scarborough (2008).

[21] This implies that a manager could always fill a hired cohort with type G applicants. In our data, 0.43 of applicants are green and 0.6 of the green or yellow applicants are green, while the hire rate is 19%, so this will be true for the typical pool.

[22] If a manager's mistakes were random noise, 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, which fits our assumed form for managerial utility.

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. It captures the extent to which the variability of hires is determined by idiosyncratic factors rather than match quality. An unbiased manager has $k = 0$, while a manager who makes decisions entirely based on idiosyncratic factors corresponds to $k = 1$.

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, nor do they observe s_i or b_i.

Managers form a posterior expectation of worker quality given both their private signal and the test signal. They then maximize their own utility by hiring a worker if and only if the expected value of U_i conditional on s_i, b_i, and t_i is at least some threshold. Managers thus wield "discretion" because they choose how to weigh the 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).

3.2 Model Predictions

Under discretion, testing will always improve the utility of managers because they retain the option to ignore the test. It can also improve outcomes from the firm's perspective. Because a and b are independent, and testing does not impact a manager's relative preference for a over b (parameterized by k), better information about a is likely to translate into better match quality of hires. However, the precise impact of testing will depend on specific distributional assumptions. We thus leave as an empirical question whether testing improves match quality.[23]

Our model instead focuses on the question of how much firms should rely on their managers, versus relying on hard test information. Firms can follow the set up described above, allowing their managers to weigh both signals and make ultimate hiring decisions

[23] With the distributional assumptions made in this section, we cannot obtain a closed form solution for the change in $E[a|Hire]$ before and after testing. However, based on simulations, we believe testing will improve outcomes. See the Appendix for more details.

(we call this the “Discretion" regime). Alternatively, firms may eliminate discretion and rely solely on test recommendations (“No Discretion"). In the remainder of this section we generate a diagnostic for when one policy will dominate the other.

Neither retaining nor fully eliminating discretion 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 his or her preferred candidates, and these may generate more benefits. Rather than solving for the optimal hiring policy, we focus on the extreme of eliminating discretion entirely. This is because we can provide a tractable test for whether this counterfactual policy would make our client firms better off, relative to their current practice. All proofs are in the Appendix.

Proposition 3.1 *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 $k < k'$ match 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, match 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, match quality is higher under Discretion than No Discretion.*

Proposition 3.1 illustrates the fundamental tradeoff firms face when allocating authority: managers have private information, but they are also biased. In general, larger 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, firms benefit from Discretion when a manager has very precise information, but only if the manager is not too biased.

To understand whether No Discretion improves upon Discretion, employers would ideally like to directly observe a manager's type (bias and information). In practice, this is not possible. Instead, it is easier to observe 1) the choice set of applicants available to managers when they made hiring decisions and 2) the performance outcomes of workers hired from those applicant pools. These are also two pieces of information that we observe in our data.

Specifically, we observe cases in which managers exercise discretion to explicitly contradict test recommendations. 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 is available but not hired.

Denote the probability of an exception for a given manager, $m \in M$, as R_m. Given the assumptions made above, $R_m = E_m[Pr(Hire|Y)]$.[24] That is, 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 3.2 *Across the set of managers M, 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 then place less weight on the test relative to their own signal of a. More biased managers also make more exceptions because they place more weight on maximizing other qualities, b. Thus, increases in exceptions can be driven by both more information and more bias.

It is therefore difficult to discern whether granting discretion is beneficial to firms simply by examining how often managers make exceptions. Instead, Propositions 3.1 and 3.2 suggest that it is instructive to examine the relationship between how often managers make exceptions and the subsequent match quality of their workers. Specifically, while exceptions

[24] In particular, the assumption that $W < p_G$ helps simplify the probability of an exception to this expression. However, all results hold for the more complicated expression of the probability that a Y is hired AND a G is not hired, that arises when we relax this assumption.

(R_m) are increasing in both managerial bias and the value of the manager's private information, match quality ($E[a|Hire]$) is decreasing in bias. If across managers, $E[a|Hire]$ 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 match 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. The following proposition formalizes this intuition.

Proposition 3.3 *If the quality of hired workers is decreasing in the exception rate,* $\frac{\partial E[a|Hire]}{\partial R_m} < 0$ *across* M*, then firms can improve outcomes by eliminating discretion. If quality is increasing in the exception rate then discretion is better than no discretion.*

The intuition behind the proof is as follows. Consider two managers, one who never makes exceptions, and one who does. If a manager never makes exceptions, it must be that he or she has no additional information and no bias. As such, the match quality of this manager's hires is equivalent to match quality of workers that would be hired if the firm eliminated discretion by relying only on test information. If increasing the probability of exceptions increases the match quality of hires, then granting discretion improves outcomes relative to no discretion. If match quality declines in the probability that managers make exceptions, then firms can improve outcomes by moving to a regime with no exceptions—that is, by eliminating discretion and using only the test.

3.3 Empirical Implications

Based on this model, we ask two empirical questions:

1. Does the introduction of testing improve the match quality of hired workers?
2. Is the match quality of workers increasing or decreasing in the probability of an exception?

If the latter is decreasing then No Discretion improves match quality relative to Discretion, and if increasing then Discretion improves upon No Discretion. We will use job tenure

variables as our primarily measure of match quality. In Section 6, we will examine an alternative measure, customers served per hour, which is available for a subset of our sample.

In practice, managers may differ not only in their information or bias parameters, but also in the applicant pools that they face as well as in the unobserved quality of the locations they work at. If these factors vary systematically across managers in a way that is correlated with either the introduction of testing or the rate at which they make exceptions, then our results may be biased. Sections 4 and 5 address how we deal with these issues empirically.

4 The Impact of Testing

4.1 Empirical Strategy

We first analyze the impact of testing on worker productivity, exploiting the gradual roll-out in testing across locations and over time. We examine the impact of job testing using difference-in-difference regressions of the form:

$$\text{Outcome}_{lt} = \alpha_0 + \alpha_1 \text{Testing}_{lt} + \delta_l + \gamma_t + \text{Controls} + \epsilon_{lt} \tag{1}$$

Equation (1) compares outcomes for workers hired with and without job testing. We regress a productivity outcome (Outcome_{lt}) for workers hired to a location l, at time t, on an indicator for whether testing was available at that location at that time (Testing_{lt}) and controls. In practice, we define testing availability as whether the median hire at that location-date was tested, though we discuss robustness to other measures. As mentioned above, the location-time-specific measure of testing availability is preferred to using an indicator for whether an individual was tested (though we also report results with this metric) because of concerns that an applicant's testing status is correlated with his or her perceived quality. We estimate these regressions at the location-time (month-by-year) level, the level of variation underlying our key explanatory variable, and weight by number of hires in a

location-date.[25] The outcome measure is the average outcome for workers hired to the same location at the same time.

All regressions include a complete set of location (δ_l) and month by year of hire (γ_t) fixed effects. They control for time-invariant differences across locations within our client firms, as well as for cohort and macroeconomic effects that may impact job duration. We also experiment with a number of additional control variables, described in our results section, below. In all specifications, standard errors are clustered at the location level to account for correlated observations within a location over time.

Our primary outcome measure, Outcome_{lt}, is the log of the length of completed job spells, averaged across workers hired to firm-location l, at time t. We focus on this, and other outcomes related to the length of job spells, for several reasons. The length of a job spell is a measure that both theory and the firms in our study agree is important. Canonical models of job search (e.g., Jovanovic 1979), predict a positive correlation between match quality and job duration. Moreover, as discussed in Section 2, our client firms employ low-skill service sector workers and face high turnover and training costs: up to 5 weeks of paid training in a setting where the median worker stays only 99 days (see Figure 1.) Job duration is also a measure that has been used previously in the literature, for example by Autor and Scarborough (2008), who also focus on a low-skill service sector setting (retail). Finally, job duration can be measured reliably for all workers in our sample.

4.2 Results

Table 2 reports regression results for the log duration of completed job spells. We later report results for several duration-related outcomes that do not restrict the sample to completed spells. Of the 270,086 hired workers that we observe in our sample, 75%, or 202,728 workers have completed spells (4,401 location-month cohorts), with an average spell lasting 203 days and a median spell of 99 days. The key explanatory variable is whether or not the median hire at this location-date was tested.

[25] This aggregation affords substantial savings on computation time, and, will produce identical results to those from a worker-level regression, given the regression weights.

In the baseline specification (Panel 1, Column 1 of Table 2) we find that employees hired with the assistance of job testing stay, on average, 0.272 log points, or 31% longer, significant at the 5% level.

Panel 1 Column 2 introduces 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.[26] In this specification, we compare locations in the same firm in the same year, some of which receive job testing sooner than others. The identifying assumption is that, within a firm, locations that receive testing sooner vs. later were on parallel trends before testing. Here our estimated coefficient falls by roughly a third in magnitude, and we lose statistical significance.

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 the locations that were on an upward (or downward) trajectory, Column 3 introduces location-specific time trends. These trends also account for broad trends that may impact worker retention, for instance, smooth changes in local labor market conditions. Adding these controls reduces the magnitude of our estimate but also greatly reduces the standard errors. We thus estimate an increased completed job duration of 0.137 log points or 15%, significant at the 5%-level.

Finally, in Column 4, we add controls for the composition of the applicant pool at a location after testing is implemented: fixed effects for the number of green, yellow, and red applicants. Because these variables are defined only after testing, these controls should be thought of as interactions between composition and the post-testing indicator. With these controls, the coefficient α_1 on Testing_{lt} is the impact of the introduction of testing, for locations that end up receiving similarly qualified applicants. However, these variables also absorb any impact of testing on the quality of applicants that a location receives. For instance, the introduction of testing may have a screening effect: as candidates gradually learn about testing, the least qualified may be deterred from applying. Our point estimate remains unchanged with the inclusion of this set of controls, but the standard errors do increase substantially. This suggests that match quality improves because testing aids managers in identifying productive workers, rather than by exclusively altering the quality of the

[26] Our data firm has indicated that it was not aware of any other client-specific policy changes.

applicant pool. Overall, the range of estimates in Table 2 are similar to previous estimates found in Autor and Scarborough (2008).

Panel 2 of Table 2 examines robustness to defining testing at the individual level. For these specifications we regress an individual's job duration (conditional on completion) on whether or not the individual was tested. Because these specifications are at the individual level, our sample size increases from 4,401 location-months to 202,728 individual hiring events. Using these same controls, we find numerically similar estimates. The one exception is Column 4, which is now significant and larger: a 26% increase. From now on, we continue with our preferred metric of testing adoption (whether the median worker was tested).

Figure 2 shows event studies where we estimate the treatment impact of testing by quarter, from 12 quarters before testing to 12 quarters after testing, using our baseline set of controls. The top left panel shows the event study using log length of completed tenure spells as the outcome measure. The figure shows that locations that will obtain testing within the next few months look very similar to those that will not (because they either have already received testing or will receive it later). After testing is introduced, however, we begin to see large differences. The treatment effect of testing appears to grow over time, suggesting either that HR managers and other participants might take some time to learn how to use the test effectively. This alleviates any concerns that any systematic differences across locations drive the timing of testing adoption.

We also explore a range of other duration-related outcomes to examine whether the impact of testing is concentrated at any point in the duration distribution. For each hired worker, we measure whether they stay at least three, six, or twelve months, for the set of workers who are not right-censored.[27] We aggregate this variable to measure the proportion of hires in a location-cohort that meet each duration milestone. Regression results (analogous to those reported in panel 1 of Table 2 are reported in Appendix table A1, while event studies are shown in the remaining panels of Figure 2. For each of these measures, we again see that testing improves job durations, and we see no evidence of any pre-trends.

[27] That is, a worker will be included in this metric if his or her hire date was at least three, six, or twelve months, respectively, before the end of data collection.

5 Managerial Discretion

The evidence presented in Section 4 demonstrates that the introduction of testing improves the quality of hires. How should firms use this information, given that its managers often have their own unverifiable observations about a candidate? Is it better to implemented a rules-based hiring process based on test recommendations, or to allow managers discretion to incorporate the test results as they wish?

We assess the value of granting managers discretion by examining what happens to match quality when managers choose to exercise discretion. In our model, we showed that managers make exceptions to test recommendations when they have 1) private information suggesting that the applicant will be successful or 2) preferences or beliefs that are not fully aligned with that of the firm. If exceptions are driven by better information, then managers who make more exceptions will have higher quality hires, and firms should prefer to grant managers discretion. If exceptions are instead driven by bias, then managers with more exceptions will make lower quality hires, and firms should prefer to limit discretion. As formalized in Proposition 3.3, the key to distinguishing between these two cases is to consider the relationship between exceptions and match quality: a negative relationship indicates that firms can improve the match quality of hires by eliminating discretion, while a positive relationship means that, on average, discretion is valuable relative to no discretion.

In order to implement this test, we construct an empirical measure of the extent to which a manager makes exceptions. This presents a challenge because exceptions in our theoretical framework are a function of managerial type (bias and information) only. Empirically, however, a manger's likelihood of making exceptions will vary according to other characteristics. In order to apply our theory to the data, we need to isolate variation in exceptions that comes only from managerial information and preferences.

In particular, there are two potential sources of confounding variation in exceptions. The first comes from systematic differences in how managers are assigned to locations, or clients, or applicant pools. The second stems from idiosyncratic variation at the applicant pool level, implying that period-to-period variation in the exception rate is driven by factors

other than a manager's type. We discuss how we limit both types of confounding variation in the following two subsections.

5.1 Defining Exceptions

Our data provides us with the test scores of applicants post-testing. As a result, we are able to 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.

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

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 and green that goes unhired.

The number of exceptions 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 an exception. Similarly, if the manager hires 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 isolate the portion of variation in exceptions that are driven by managerial decisions, we normalize the number of order violations by the maximum number of violations that could occur, given the applicant pool that the recruiter faces and the number of hires. Importantly, although propositions in Section 3 are derived for the probability of an exception, their proofs hold equally for this definition of an exception rate.[28]

[28]Results reported below are qualitatively robust to a variety of different assumptions on functional form for the exception rate.

Despite the fact that firms in our sample had fairly uniform policies regarding the test, we see substantial variation in the extent to which managers actually follow test recommendations when making hiring decisions.[29] Figure 3 shows histograms of the exception rate, at the application pool level, as well as aggregated to the manager and location levels. The top panels show unweighted distributions, while the bottom panels show distributions weighted by the number of applicants.

In all figures, the median exception rate is about 20% of the maximal number of possible exceptions. At the pool level, the standard deviation is also about 20 percentage points; at the manager and location levels, it is about 11 percentage points. This means that managers very frequently make exceptions and that some managers and locations consistently make more exceptions than others.

5.2 Empirical Specifications

The most direct implementation of Proposition 3.3 examines the correlation between the exception rate and the realized match quality of hires in the post-testing period:

$$\text{Duration}_{mlt} = a_0 + a_1 \text{Exception Rate}_{mlt} + X_{mlt}\gamma + \delta_l + \delta_t + \epsilon_{mlt} \tag{3}$$

The coefficient of interest is a_1. A negative coefficient, $a_1 < 0$, indicates that the match quality of hires is decreasing in the exception rate, meaning that firms can improve the match quality of hires by eliminating discretion and relying solely on job test information.

In addition to normalizing exception rates to account for differences in applicant pool composition, we estimate multiple version of Equation (3) that include location and time fixed effects, client-year fixed effects, location-specific linear time trends, and detailed controls for the quality and number of applicants in an application pool.

These controls are important because observed exception rates may be driven by factors other than a manager's type (bias and information parameters). For example, some locations may be inherently less desirable than others, attracting both lower quality managers and

[29] The client firms in our sample often told their managers that job test recommendations were informative and should be used in making hiring decisions. Following that, many firms gave managers discretion over how to use the test, though some locations strongly discouraged managers from hiring red candidates.

lower quality applicants. In this case, lower quality managers may make more exceptions because they are biased. At the same time, lower quality workers may be more likely to quit or be fired. Both facts would be driven by unobserved location characteristics. Another potential concern is that undesirable locations may have difficulty hiring green workers, even conditional on them having applied. In our data, we cannot distinguish a green worker who refuses a job offer from one who was never offered the job. As long as these characteristics are fixed or vary smoothly at the location-level, our controls absorb this variation.

A downside of including many fixed effects in Equation (3) is that it increases the extent to which our identifying variation is driven by pool-to-pool variation in the idiosyncratic quality of applicants. To see why this is problematic, imagine an applicant pool with a particularly weak draw of green candidates. In this case, we may expect a manager to make more exceptions. Yet, because the green workers in this pool are weak, it may also separately be the case that the pool as a whole is weak. In this case, a manager could be using his or her discretion to improve match quality, but exceptions will still be correlated with poor outcomes. That is, when we identify off of pool-to-pool variation in exception rates, we may get the counterfactual wrong because exceptions are correlated with variation in unobserved quality within color.

To deal with the concern that Equation (3) relies too much on pool-to-pool variation in exception rates, we can aggregate exception rates to the manager- or location-level. Aggregating across multiple pools removes the portion of exception rates that are driven by idiosyncractic differences in the quality of workers in a given pool. The remaining variation—differences in the average exception rate across managers or locations—is more likely to represent exceptions made because of managerial type (bias and information). Doing so, however, reduces the amount of within-location variation left in our explanatory variable, making controlling for location fixed effects difficult or impossible.

To accommodate aggregate exception rates, we expand our data to include pre-testing worker observations. Specifically, we estimate whether the *impact* of testing, as described in

Section 4, varies with exception rates:

$$\begin{aligned} \text{Duration}_{mlt} &= b_0 + b_1 \text{Testing}_{lt} \times \text{Exception Rate}_{mlt} + b_2 \text{Testing}_{lt} \\ &\quad + X_{mlt}\gamma + \delta_l + \delta_t + \epsilon_{mlt} \end{aligned} \tag{4}$$

Equation (4) estimates how the impact of testing differs when managers make exceptions. The coefficient of interest is b_1. Finding $b_1 < 0$ indicates that making more exceptions decreases the improvement that locations see from the implementation of testing, relative to their pre-testing baseline. Because exception rates are not defined in the pre-testing period (there are no test scores in the pre-period), there is no main effect of exceptions in the pre-testing period, beyond that which is absorbed by the location fixed effects δ_l.

This specification allows us to use the pre-testing period to control for location-specific factors that might drive correlations between exception rates and outcomes. It also expands the sample on which we estimate location-specific time trends. This allows us to use exception rates that are aggregated to the manager- or location-level, avoiding small sample variation.[30] Aggregating exception rates to the location level also helps remove variation generated by any systematic assignment of managers to applicants within a location that might be correlated with exception rates and applicant quality.[31]

To summarize, we test Proposition 3.3 with two approaches. First, we estimate the correlation between pool-level exception rates and quality of hires across applicant pools. This is the most literal analogue to our model. Second, we estimate the differential impact of testing across pools with different exception rates of hires, where exception rates can be at the manager-, or location-level. In Section 5.4, we describe additional robustness checks.

5.3 Results

To gain a sense of the correlation between exception rates and outcome of hires, we first summarize the raw data by plotting both variables at the location level. Figure 4 shows a

[30] We define a time-invariant exception rate for managers (locations) that equals the average exception rate across all pools the manager (location) hired in (weighted by the number of applicants).

[31] It also helps us rule out any measurement error generated by the matching of applicants to HR managers. This would be a problem if in some cases hiring decisions are made more collectively, or with scrutiny from multiple managers, and these cases were correlated with applicant quality.

scatter plot of the location-level average exception rate on the x-axis and the location-level average tenure (log of completed duration) for workers hired post-testing on the y-axis. In the first panel, each location has the same weight; in the second, locations are weighted by the inverse variance of their pre-period mean, which takes into account their size and the confidence of our estimates. In both cases, we see a negative correlation between the extent to which managers exercise discretion by hiring exceptions, and the match quality of those hired.

Table 3 presents the correlation between exception rates and worker tenure. We use two measures of the exception rate: a standardized exception rate with mean 0 and standard deviation 1 (Columns 1-2), and an indicator variable for whether that applicant pool had above–median exceptions (Columns 3-4). In this table, exception rates are defined at the pool level (based on the set of applicants and hires a manager makes at a particular location in a given month). Columns 1 and 3 contain our basic specification (Equation (3)) while Columns 2 and 4 add our full set of controls: location-specific time trends, client-year effects, and applicant pool controls.

The coefficient on the exception rate is negative and similar in magnitude regardless of the controls. For example, Column 2 indicates that a one standard deviation increase in the exception rate of a pool is associated with a 3.9% lower completed tenure for that group, significant at the 5% level. We thus find that hires from applicant pools where managers exercised more discretion perform worse than hires from pools where managers exercised less discretion.

Table 4 examines how the impact of testing varies by the extent to which managers make exceptions. Our main explanatory variable is the interaction between the introduction of testing and a post-testing exception rate. Here, we report results with the full set of controls, though our other specifications look similar.

In Column 1, we continue to use pool-level exception rates. The coefficient on the main effect of testing represents the impact of testing at the mean exception rate (since the exception rate has been standardized), and will thus be very similar to those reported in Table 2. We find that locations with the mean exception rate experience a 0.22 log point increase in duration as a result of the implementation of testing, but that this effect is offset

by a quarter (0.05) for each standard deviation increase in the exception rate, significant at the 1% level.

In Columns 2 and 3, we aggregate exception rates to the manager- and location-level, respectively.[32] Results are quite consistent, using these aggregations, and the differential effects are even larger in magnitude. Managers and locations that tend to exercise discretion benefit much less from the introduction of testing. A one standard deviation increase in the exception rate reduces the impact of testing by roughly half. Columns 4-6 use an indicator for high- and low-exception rates for each level of aggregation, and yields similar results.

To better illustrate the variation underlying the results in Table 4, we plot location-specific treatment effects of testing on the location's average exception rate. Figure 5 plots these for both an unweighted and weighted sample, as described above. The relationship is clearly negative, and does not look to be driven by any particular location.

We therefore find that the match quality of hires is lower for applicant pools, managers, and locations with higher exception rates. It is worth emphasizing that all our estimates include detailed controls for both the size and the quality of the applicant pool. With these controls, our identification comes from comparing outcomes of hires across managers who make different numbers of exceptions when facing similar applicant pools. Given this, differences in exception rates should be driven by a manager's own weighting of his or her private preferences and private information. If managers were making these decisions optimally from the firm's perspective, we should not expect to see (as we do in Tables 3 and 4) that the workers they hire perform systematically worse. Based on Proposition 3.3, we can infer then that exceptions are largely driven by managerial bias, rather than private information, and these firms could improve outcomes of hires by limiting discretion.

5.4 Additional Robustness Checks

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

[32] We have 555 managers who are observed in an average of 18 pools each (average taken over all managers, unweighted). We have 111 locations with on average 87 pools each (average taken over all locations, unweighted).

5.4.1 Quality of Individual Exceptions

There are several scenarios under which we might find a negative correlation between overall exceptions and outcomes without biased managers. For example, managers may make more exceptions when green applicants in an 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 match quality. Similarly, our results in Table 4 show that locations with more exceptions see fewer benefits from the introduction of testing. An alternative explanation for this finding is that high exception locations are ones in which managers have always had better information about applicants: these locations see fewer benefits 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. Whereas, if, as we argue, exceptions are primarily driven by bias, then yellow candidates who beat out greens will still perform worse. While we cannot observe the performance of non-hired greens, we can proxy for this comparison by exploiting the timing of hire. We can compare the performance of yellow workers hired as exceptions against green workers from the same applicant pool who are not hired that month, but who subsequently begin working in a later month. If it is the case that managers are making exceptions to increase the match quality of workers, then the exception yellows should have longer completed tenures than the passed over greens.

Table 5 shows that is not the case. The first panel estimates our typical duration regression, restricting the sample to workers who are either exception yellows, or greens who are initially passed over but then subsequently hired, and including an indicator for being in the latter group. Because these workers are hired at different times, all regressions control for hire year-month fixed effects to account for mechanical differences in duration. For the last column, which includes applicant pool fixed effects, the coefficient on being a passed over green compares this group to the specific yellow applicants who were hired before them.

The second panel of Table 5 repeats this exercise, comparing red workers hired as exceptions (the omitted group), against passed over yellows and passed over greens.

In both panels, we find that workers hired as exceptions have shorter tenures. Column 3 is our preferred specification because it adds controls for applicant pool fixed effects. This means we compare the green (and yellow) applicants who were passed over one month but eventually hired, to the actual yellow (red) applicants hired first. We find that passed over greens stay about 8% longer than the yellows hired before them in the same pool (top panel column 3) and greens and yellows stay almost 19% and 12% longer, respectively, compared to the reds they were passed over for.

The results in Table 5 mean 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. However, an alternative explanation is that the applicants with higher test scores were not initially passed up, but were instead initially unavailable because of better outside options. Unfortunately, in our data, we cannot distinguish the hire date from the start date. However, given the general undesirability of the job, and the fact that hire rates are low for all types of workers (from Table 1, only one-fifth of greens are hired), we believe that most applicants would not delay the job opportunity.

Table 6 provides additional evidence that workers with longer gaps between application date and hire date (which we treat as temporarily passed over applicants) are not simply ones who were delayed because of better outside options. If this were the case, we would expect these workers to have better outcomes once they do begin work. In Table 6, we compare match quality for workers hired immediately (the omitted category), compared to those who waited one, two, or three months before starting, holding constant test score. Because these workers are hired at different times, all regressions again control for hire year-month fixed effects. Across all specifications, we find no significant differences between these groups. If anything we find for greens and yellows that were hired with longer delays have shorter job spells than immediate hires. We thus feel more comfortable interpreting the workers with longer delays as having been initially passed over.

Table 6 also provides insights 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. Table 6 shows that this does not appear to be the case. We estimate only small and insignificant differences in tenure, within color, across start dates. That is, within color, workers who appear to be a manager's first choice do not perform better than workers who appear to be a manager's last choice. This again suggests the value of managerial discretion is small, relative to the test.

5.4.2 Validity of test recommendations

Another possible concern is that the usefulness of the test varies across locations. 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. Here, a negative correlation between exceptions and performance would not necessarily imply that firms could improve productivity by relying more on testing.

Our results on individual exceptions already suggest that this is not the case. In addition, we can compare differences in tenure of green, yellow, and red workers, by various location characteristics. This will help determine, whether the test is more or less meaningful in some locations compared to others. However, we see no evidence that the "return" to test score in terms of tenure varies across locations. For example, when we split locations by pre-testing worker durations (Appendix Table A2) or by exception rates post-testing (Appendix Table A3) we see no systematic differences in the impact of color on job duration.

5.4.3 Differences between high and low exception rate locations

Our results in Table 4 show that the impact of testing is smaller when managers have higher exception rates. Recall, Proposition 3.3 can be used to infer the benefits of discretion by examining the average quality of hires as a function of exceptions, not the impact of testing as a function of exceptions. A location might have a smaller impact of testing AND a higher exception rate, either because managers are more biased, or because it had better information pre-testing so the test reveals relatively less. That is, high exception locations could be the ones in which managers have always had better information about applicants:

these locations see fewer benefits from testing because they simply do not need the test. As such, while measuring the impact of testing allows us to use pre-testing outcomes to control more effectively for location fixed effects, without being sensitive to pool-to-pool variation in exceptions, it also has an alternative interpretation.

To better understand what drives heterogeneity in exception rates, we look at the correlation between exception rates and pre-testing outcomes at the location level. If exceptions are driven by information, rather than bias, we should see that higher exception rate locations were more productive pre-testing. Figure 6 plots the relationship between a location's eventual exception rates and the match quality of its hires prior to the introduction of testing. We do this at the location level because we cannot match hires to specific managers before testing, and present both unweighted and weighted (as described above) graphs. We see that there is no systematic relationship between a location's exception rate and the quality of its hires in the pre-testing period.[33] This result suggests that our result on the impact of testing is not driven by a correlation between exceptions and better informed managers. Furthermore, this finding suggests that locations with more exceptions are not otherwise undesirable, which might then generate a spurious negative correlation between exceptions and job durations.

6 Robustness to Productivity

Our results show that firms can improve the match quality of their workers, as measured by duration, by relying more on job test recommendations. Firms may not want to pursue this strategy, however, if their HR managers exercise discretion in order to improve worker quality on other metrics. For example, managers 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. Specifically, for a subset of 62,508 workers (one-quarter of all hires) in 6 client firms, we observe a direct

[33] The small positive slope in the unweighted graph is not statistically significant, and is driven by an outlier in the top-right quadrant that is estimated quite imprecisely (see the same dot in the weighted graph).

measure of worker productivity: number of customers served per hour.[34] We are unable to reveal the exact nature of this measure but some examples may include: the number of calls telemarketers or customer service agents are able to complete per hour, the number of tests an examiner can grade per hour, and the speed at which retail checkout clerks scan items. In all of these examples, customers served per hour is an important measure of efficiency and worker productivity. Our particular measure has an average of roughly 12 clients served per hour with a noisy standard deviation of roughly 60.

Table 7 repeats our main findings, using customers served per hour instead of job duration as the dependent variable. Column 1 examines the impact of the introduction of testing, Columns 2 documents the post-testing correlation between pool-level exceptions and customers served per hour, and Columns 3-5 examines how the impact of testing varies by exception rates. We provide estimates only using our base specification (controlling for date and location fixed effects) because the smaller sample and number of clients makes estimation of the other controls difficult.

In all cases, we find no evidence that managerial exceptions improve productivity, as defined by customers served per hour. Instead, we find noisy estimates indicating that worker quality appears to be lower on this dimension as well. For example, Column 1 of Table 7 shows that the introduction of testing leads to a statistically insignificant increase of 2 customers served per hour, or a roughly 17% increase. This finding is consistent with the improvement in job duration associated with testing.

Column 2, we find that the correlation between the pool-level exception rate and customers served per hour is negative and sizeable in magnitude, though noisy and insignificant. In Columns 3-5, we similarly find that the positive impact of testing on calls per hour is substantially offset in pools, managers, and locations with higher exception rates. For example, the positive impact of testing seen for the average hire completely disappears for workers hired by a manager with a one standard deviation higher exception rate.

These estimates are noisy and often not statistically distinguishable from zero. They are also not as stable across various sample selection criteria as our main tenure estimates.

[34]However, we have repeated our main analyses on the subsample of observations that have this performance measure and obtained similar results.

However, across a wide range of specifications, we can always rule out that exceptions positively impact productivity at any meaningful magnitude. For example, from Column 2, we can rule out a positive correlation between pool-level exceptions and post-testing productivity above 0.09 (less than 1%) with 90% confidence. From Columns 3-5, we can rule out that exceptions improve the impact of testing by any more than 1-8% with 90% confidence.

Taken together, the results in Table 7 provide no evidence that exceptions are positively correlated with productivity. This refutes the hypothesis that, when making exceptions, managers optimally sacrifice job tenure in favor of workers who perform better on other quality dimensions.

7 Conclusion

We evaluate the introduction of a hiring test across a number of firms and locations in a low-skill service sector industry. Exploiting variation in the timing of adoption across locations within firms, we show that testing increases the durations of hired workers by 15%. We then document substantial variation in how managers use job test recommendations. Some managers tend to hire applicants with the best test scores while others make many more exceptions. Across a range of specifications, we show that the exercise of discretion (hiring against the test recommendation) is associated with worse outcomes.

This finding suggests that managers underweight the job test relative to what the firm would prefer. In the setting, firms may want to then take advantage of the verifiability of job test scores to impost hiring rules that centralize hiring authority. Our paper thus contributes a new methodology for evaluating the value of discretion in firms. Our test is simple, tractable, and requires only data that would readily be available for firms using workforce analytics.

These findings highlight the role new technologies can play in solving agency problems in the workplace through contractual solutions. 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 with the firm. This paper makes an important step towards understanding and quantifying these issues.

References

[1] Aghion, P. and J. Tirole (1997), "Formal and Real Authority in Organizations," *The Journal of Political Economy*, 105(1).

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

[3] Alonso, Dessein, Matouschek (2008)

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

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

[6] Baker, G. and T. 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] Brown, M., E. Setren, and G. Topa (2014), "Do Informal Referrals Lead to Better Matches? Evidence from a Firm's Employee Referral System," mimeo New York Federal Reserve Bank.

[8] Burks, S., B. Cowgill, M. Hoffman, and M. Housman (2013), "The Facts About Referrals: Toward an Understanding of Employee Referral Networks," mimeo University of Toronto.

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

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

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

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

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

[14] Koenders, K. and R. Rogerson (2005), "Organizational Dynamics Over the Business Cycle: A View on Jobless Recoveries," Federal Reserve Bank of St. Louis Review.

[15] 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.

[16] Lazear, Edward (2000), "Performance Pay and Productivity," *American Economic Review*, 90(5), p. 1346-1461.

[17] Li, D. (2012), "Expertise and Bias in Evaluation: Evidence from the NIH" mimeo Harvard University.

[18] Oyer, P. and S. Schaefer (2011), "Personnel Economics: Hiring and Incentives," in the *Handbook of Labor Economics*, 4B, eds. David Card and Orley Ashenfelter, pp. 1769-1823.

[19] Pallais, A. and E. Sands, "Why the Referential Treatment? Evidence from Field Experiments on Referrals," mimeo Harvard University.

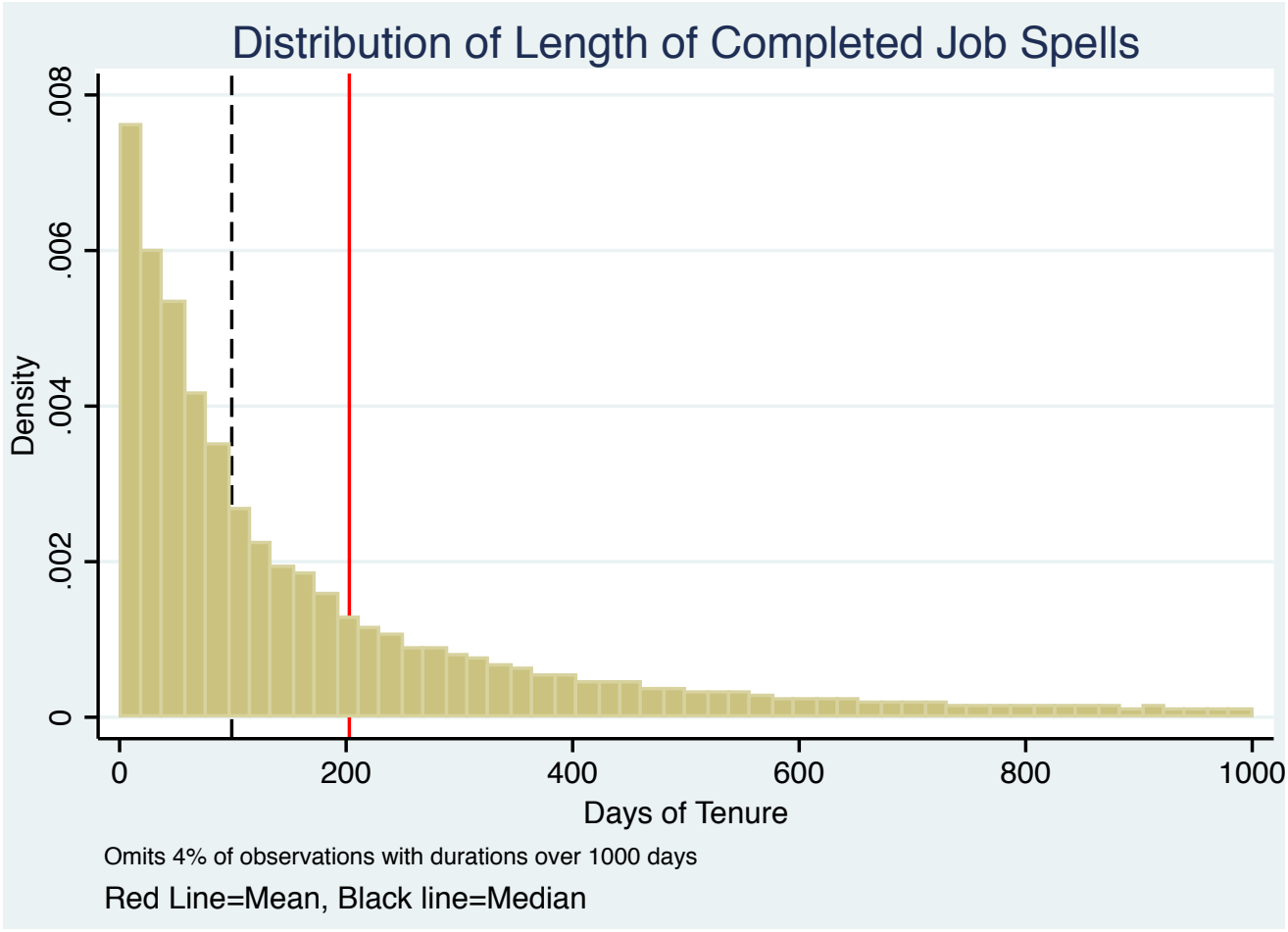
[20] Paravisini, D. and A. Schoar (2013) "The Incentive Effect of IT: Randomized Evidence from Credit Committees" NBER Working Paper #19303.

[21] Rantakari, H. (2008) "Governing Adaptation," *Review of Economic Studies.* 75, pp. 1257?1285

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

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

FIGURE 1: DISTRIBUTION OF COMPLETED SPELLS

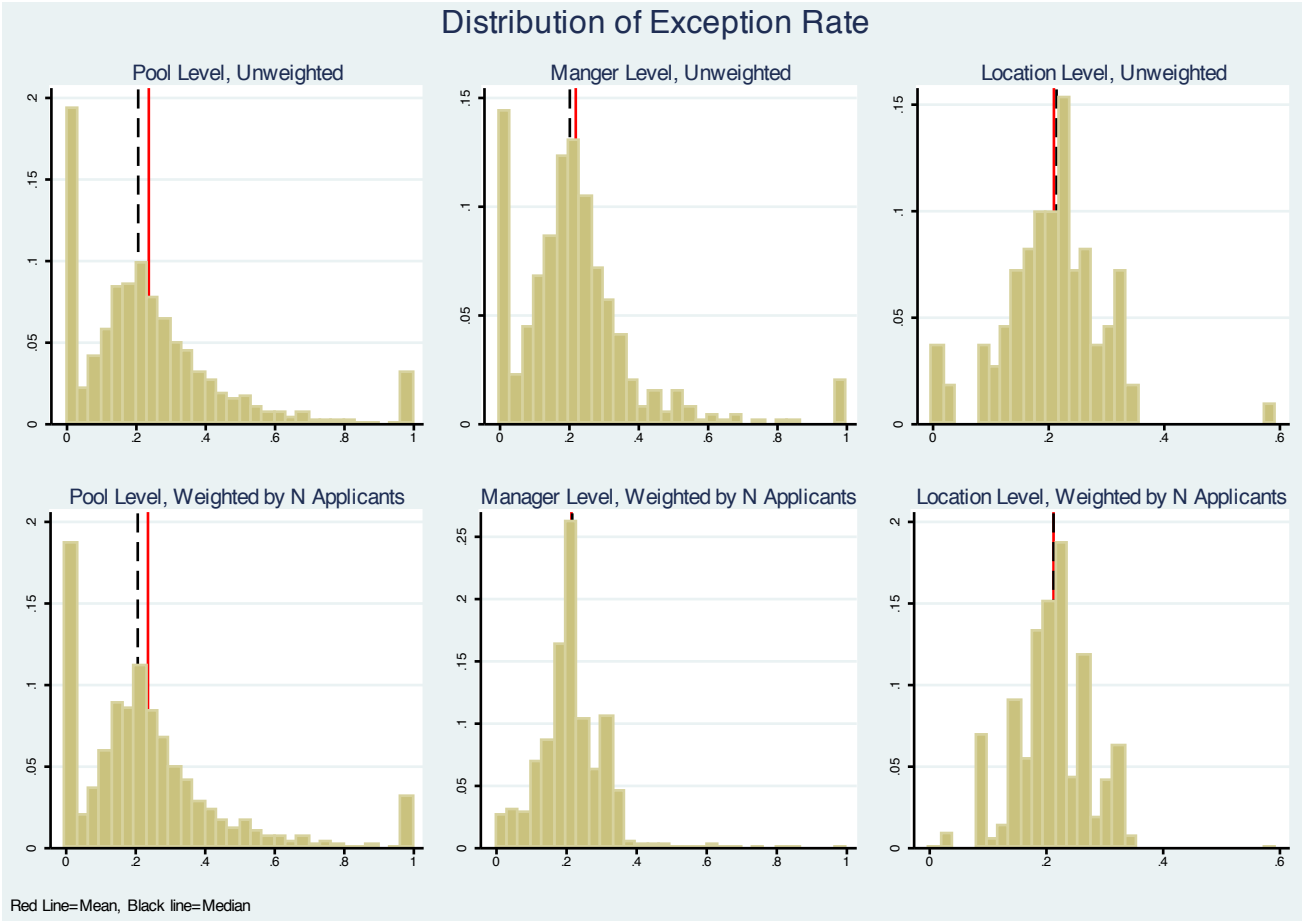


NOTES: Figure 1 plots the distribution of completed job spells at the individual level.

Figure 2: Event Study of Duration Outcomes

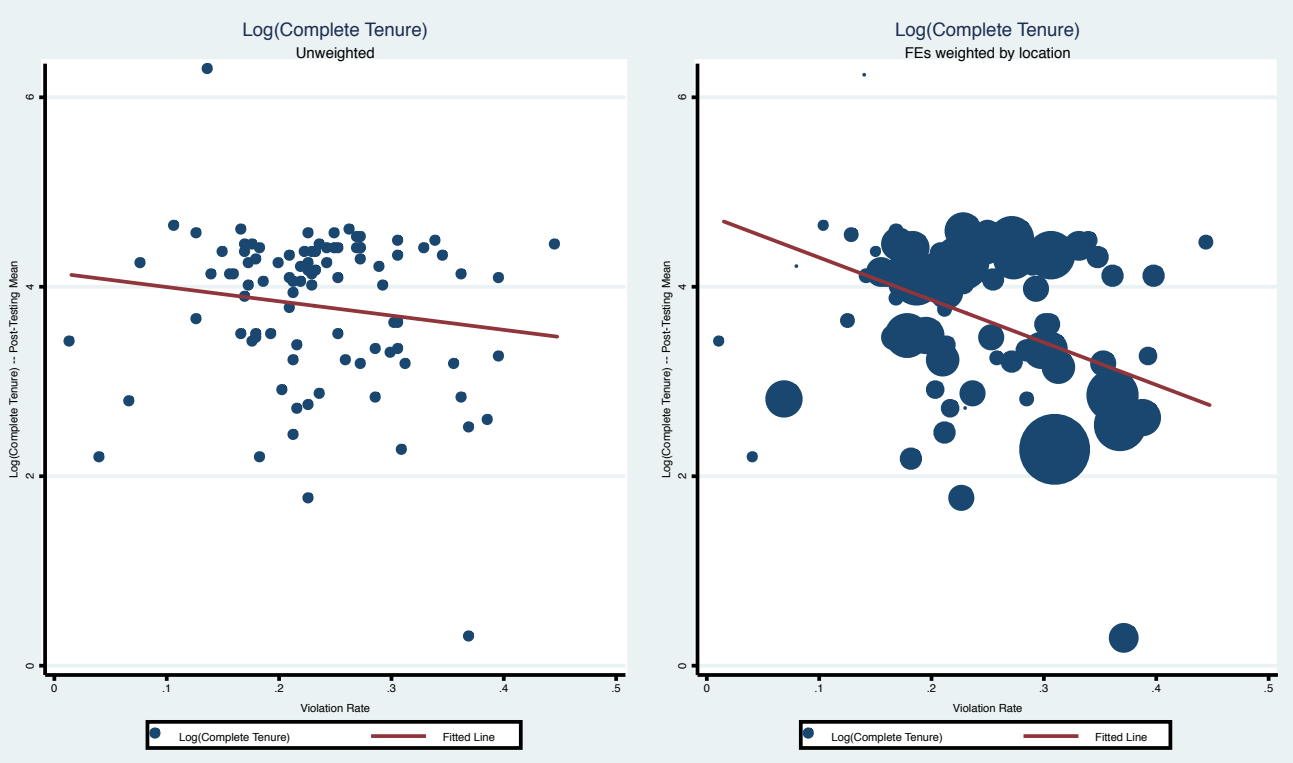
Notes: These figures plot the average duration outcome by time (in quarters) until or time after testing is adopted. The underlying estimating equation is given by $\text{Log(Duration)}_{lt} = \alpha_0 + I_{lt}^{\text{time since testing}}\alpha_1 + \delta_l + \gamma_t + \epsilon_{lt}$, where $I^{\text{time since testing}}$ is a vector of dummies indicating how many quarters until or after testing is adopted, with one quarter before as the omitted category. This regression does not control for location-specific time trends; if those are present, they would be visible in the figure.

FIGURE 3: VARIATION IN APPLICATION POOL EXCEPTION RATE



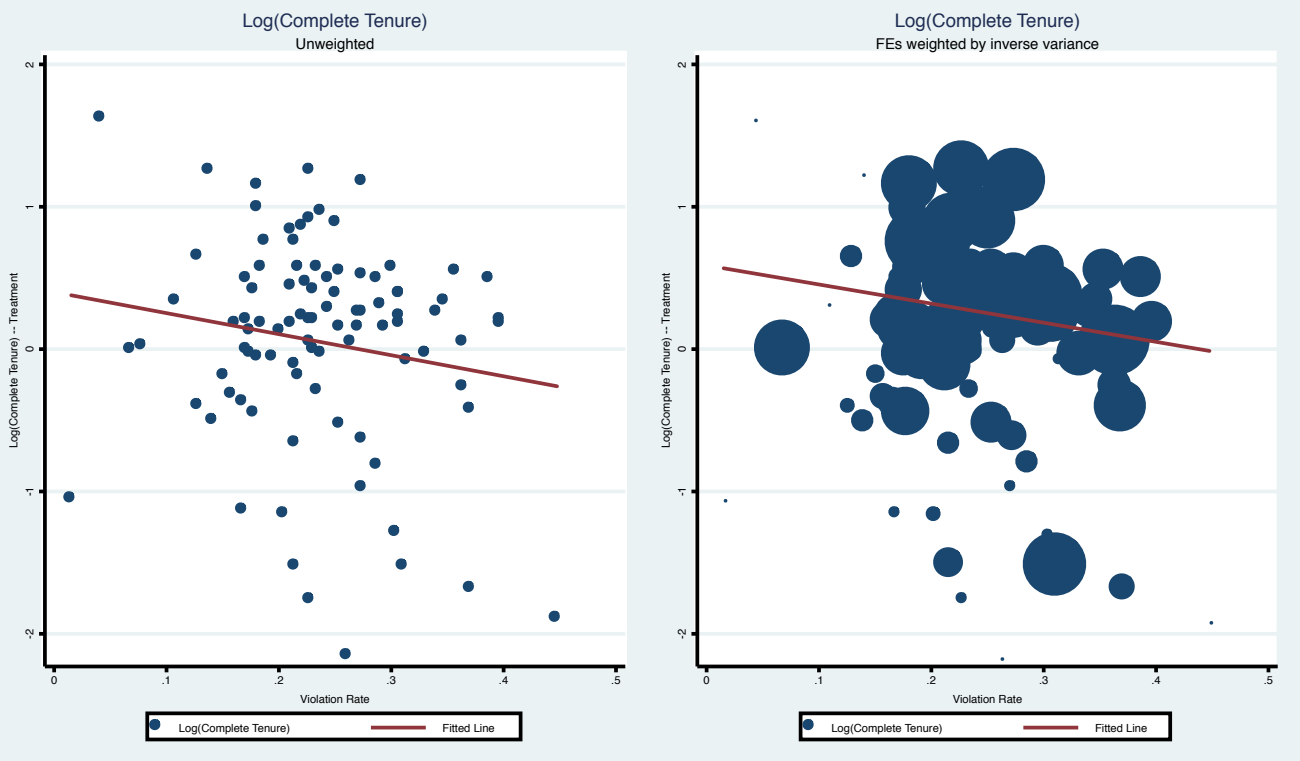
NOTES: These figures plot the distribution of the exception rate, as defined by Equation (2) in Section 5. The leftmost panel present results at the applicant pool level (defined to be a manager–location–month). The middle panel aggregates these data to the manager level and the rightmost panel aggregates further to the location level.

Figure 4: Exception Rate vs. Post-testing Match Quality



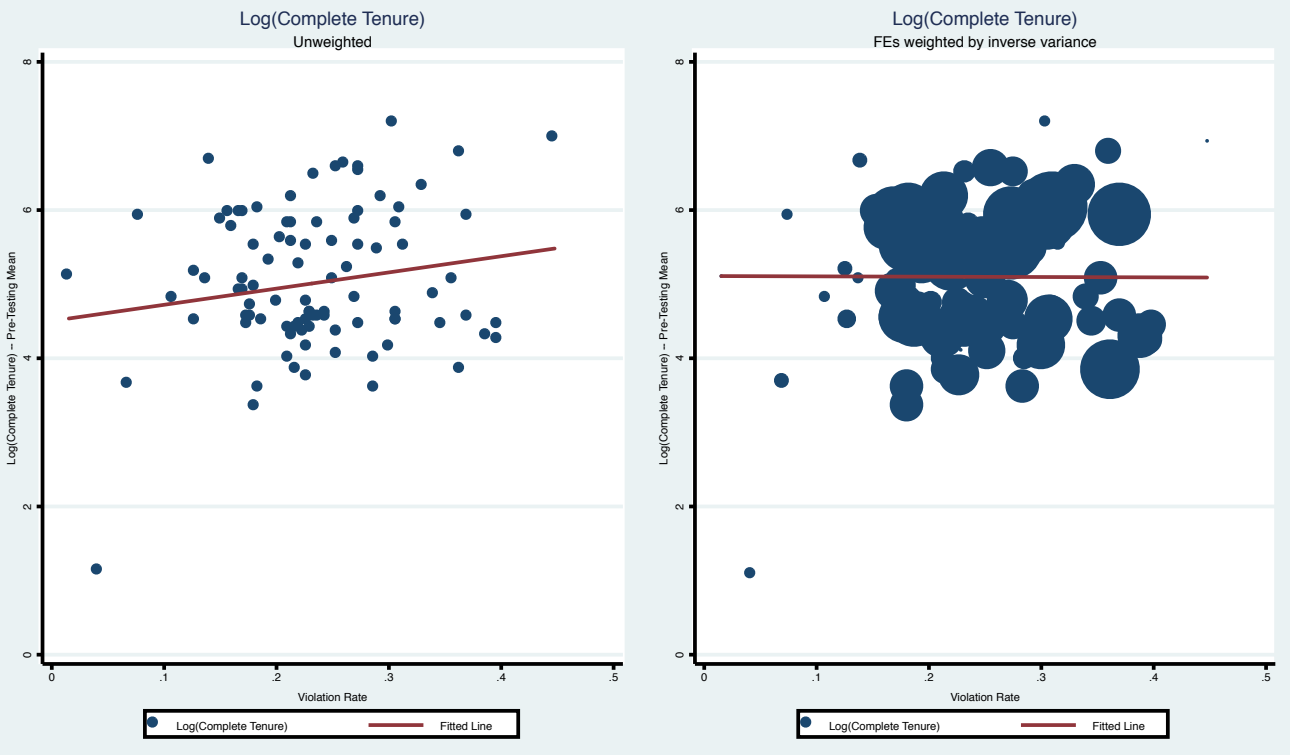
Notes: Each dot represents a given location. The y-axis is the mean log completed tenure at a given location after the introduction of testing; the x-axis is the average exception rate across all application pools associated with a location. The first panel presents unweighed correlations; the second panel weights by the number of applicants to a location.

Figure 5: Exception Rate vs. Impact of Testing



Notes: Each dot represents a given location. The y-axis is the coefficient on the location-specific estimate of the introduction of testing; the x-axis is the average exception rate across all application pools associated with a location. The first panel presents unweighed correlations; the second panel weights by the inverse variance of the error associated with estimating that location's treatment effect.

Figure 6: Exception Rate vs. Pre-testing Job Durations



Notes: Each dot represents a given location. The y-axis reports mean duration variables at a given location prior to the introduction of testing; the x-axis is the average exception rate across all application pools associated with a location. The first panel presents unweighed correlations; the second panel weights by the inverse variance of the error associated with estimating that location's treatment effect, to remain consistent with Figure 5

TABLE 1: SUMMARY STATISTICS

	Sample Coverage		
	All	Pre-testing	Post-testing
Sample Coverage			
# Locations	131	116	111
# Hired Workers	270,086	176,390	93,696
# Applicants			691,352
# HR Managers			555
# Pools			4,209
# Applicants/Pool			268

	Applicant Pool Characteristics			
	Post-testing	Green	Yellow	Red
Share Applicants		0.43	0.29	0.28
Share Hired	0.19	0.22	0.18	0.09

		Worker Performance	*mean (st dev)*		
	Pre-testing	Post-testing	Green	Yellow	Red
Duration of Completed Spell (Days) (N=202,728)	247 *(314)*	116 *(116)*	122 *(143)*	110 *(131)*	93 *(122)*
# Customers Served/Hr (N=62,676)	12.9 *(64.5)*	11.9 *(55.2)*	12.4 *(69.1)*	11.0 *(27.3)*	12.2 *(25.7)*

NOTES: The sample includes all non stock-sampled workers. 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 HR manager-location-month level and includes all applicants that had applied within four months of the current month and not yet hired.

TABLE 2: THE IMPACT OF JOB TESTING ON COMPLETED JOB SPELLS

	(1)	(2)	(3)	(4)
Panel 1: Location-Cohort Mean Log Duration of Completed Spells				
Testing Used for Median Worker	0.272**	0.178	0.137**	0.142
	(0.113)	*(0.113)*	*(0.0685)*	*(0.101)*
N	4,401	4,401	4,401	4,401
Panel 2: Individual-Level Log Duration of Completed Spells				
Individual Applicant is Tested	0.195*	0.139	0.141**	0.228**
	(0.115)	*(0.124)*	*(0.0637)*	*(0.0940)*
N	202,728	202,728	202,728	202,728
Year-Month FEs	X	X	X	X
Location FEs	X	X	X	X
Client Firm X Year FEs		X	X	X
Location Time Trends			X	X
Size and Composition of Applicant Pool				X

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

NOTES: In Panel 1, an observation in this regression is a location-month. The dependent variable is average duration, conditional on completion, for the cohort hired in that month. 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. Regressions are weighted by the number of applicants. Standard errors in parentheses are clustered at the location level. In Panel 2, observations are at the individual level. Testing is defined as whether or not an individual worker has a score. regressions are unweighted.

Table 3: Exception Rates and Post-Testing Duration

	(1)	(2)	(3)	(4)
	Log Duration of Completed Spells			
Standardized Exception Rate Post Testing	-0.0491** *(0.0223)*	-0.0385** *(0.0192)*		
> Median Exception Rate Post Testing			-0.0211 *(0.0183)*	-0.0353* *(0.0200)*
N	3,839	3,839	3,926	3,926
Year-Month FEs	X	X	X	X
Location FEs	X	X	X	X
Client Firm X Year FEs		X		X
Location Time Trends		X		X
Size and Composition of Applicant Pool		X		X

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

Notes: Each observation is a manager-location-month, for the post-testing sample only. The exception rate is the number of times a yellow is hired above a green or a red is hired above a yellow or green in a given applicant pool, divided by the maximum number of such violations. It is standardized to be mean zero and standard deviation one.

Table 4: Exception Rates and the Impact of Testing

	(1)	(2)	(3)	(4)	(5)	(6)
	Log Duration of Completed Spells					
	Level of Aggregation for Exception Rate					
	Pool	*Manager*	*Location*	*Pool*	*Manager*	*Location*
Testing Used for Median Worker	0.217**	0.243***	0.252***	0.251***	0.267***	0.336***
	(0.0876)	*(0.0869)*	*(0.0880)*	*(0.0891)*	*(0.0848)*	*(0.126)*
Standardized Exception Rate Post Testing	-0.0477***	-0.131***	-0.170**			
	(0.0182)	*(0.0335)*	*(0.0744)*			
> Median Exception Rate Post Testing				-0.0545**	-0.0783**	-0.188
				(0.0226)	*(0.0331)*	*(0.161)*
N	3,839	3,912	3,926	3,926	3,926	3,926
Year-Month FEs	X	X	X	X	X	X
Location FEs	X	X	X	X	X	X
Client Firm X Year FEs	X	X	X	X	X	X
Location Time Trends	X	X	X	X	X	X
Size and Composition of Applicant Pool	X	X	X	X	X	X

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

Notes: See notes to Table 3. Each observation is a manager-location-month, for the entire sample period. The exception rate is the number of times a yellow is hired above a green or a red is hired above a yellow or green in a given applicant pool. This baseline exception rate is the pool level exception rate. It is then aggregated to either the manager or location level to reduce the impact of pool to pool variation in unobserved applicant quality. All exception rates are standardized to be mean zero and standard deviation one. Exception rates are only defined post testing and are set to 0 pre testing. See text for additional details.

Table 5: Match Quality of Exceptions vs. Passed Over Applicants

	Log(Completed Job Spell)		
	(1)	(2)	(3)
Panel 1: Quality of Yellow Exceptions vs. Passed over Greens			
Passed Over Greens	0.0436***	0.0436***	0.0778***
	(0.0140)	*(0.0140)*	*(0.0242)*
N	59,462	59,462	59,462
Panel 2: Quality of Red Exceptions vs. Passed over Greens and Yellows			
Passed Over Greens	0.131***	0.131***	0.171***
	(0.0267)	*(0.0267)*	*(0.0342)*
Passed Over Yellows	0.0732***	0.0732***	0.112***
	(0.0265)	*(0.0265)*	*(0.0328)*
N	44,456	44,456	44,456
Hire Month FEs	X	X	X
Location FEs	X	X	X
Client Firm X Year FEs		X	X
Application Pool FEs			X

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

Notes: Each observation is an applicant-pool, at the individual level, post testing only. The top panel includes only yellow exceptions and passed over green applicants who are later hired. The omitted category are yellow exceptions. The second panel includes red exceptions and passed over greens and yellows only. Red exceptions are the omitted category.

Table 6: Job Duration of Workers, by Length of Time in Applicant Pool

	Log(Completed Job Spell)		
	(1)	(2)	(3)
Green Workers		**Green Workers**	
Waited 1 Month	-0.00908	-0.00908	0.00627
	(0.0262)	*(0.0262)*	*(0.0204)*
Waited 2 Months	-0.0822	-0.0822	-0.0446
	(0.0630)	*(0.0630)*	*(0.0385)*
Waited 3 Months	-0.000460	-0.000460	-0.0402
	(0.0652)	*(0.0652)*	*(0.0639)*
N	41,020	41,020	41,020
		Yellow Workers	
Waited 1 Month	-0.00412	-0.00412	0.00773
	(0.0199)	*(0.0199)*	*(0.0243)*
Waited 2 Months	-0.0100	-0.0100	-0.0474
	(0.0448)	*(0.0448)*	*(0.0509)*
Waited 3 Months	0.103	0.103	0.114
	(0.0767)	*(0.0767)*	*(0.0979)*
N	22,077	22,077	22,077
		Red Workers	
Waited 1 Month	0.0712	0.0712	0.0531
	(0.0520)	*(0.0520)*	*(0.0617)*
Waited 2 Months	0.0501	0.0501	0.0769
	(0.0944)	*(0.0944)*	*(0.145)*
Waited 3 Months	0.103	0.103	0.149
	(0.121)	*(0.121)*	*(0.168)*
N	4,919	4,919	4,919
Year-Month FEs	X	X	X
Location FEs	X	X	X
Client Firm X Year FEs		X	X
Application Pool FEs			X

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

Notes: Each observation is an individual hired worker. The first panel restricts to green workers only, with green workers who are hired immediately serving as the omitted group. The other panels are defined analogously for yellow and red.

TABLE 7: TESTING, EXCEPTION RATES, AND CUSTOMERS SERVED PER HOUR

	Impact of testing	**Exceptions and outcomes, post testing**	**Impact of testing, by exception rate**		
			Level of aggregation for exception rate		
			Pool	*Manager*	*Location*
	(1)	(2)	(5)	(6)	(7)
			Customers served per hour		
Testing Used for Median Worker	2.057 *(1.430)*		2.326* *(1.343)*	2.783** *(1.256)*	3.082*** *(1.114)*
Standardized Exception Rate Post Testing		-0.747 *(0.511)*	-1.361 *(0.933)*	-2.213 *(1.779)*	-3.253 *(2.646)*
N	1,824	748	1,824	1,824	1,824
Year-Month FEs		X	X	X	X
Location FEs		X	X	X	X

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

NOTES: This table replicates the baseline specifications in Tables 2, 3, and 4, using the number of customers served per hour (mean 8.38, std. dev. 3.14) as the dependent variable. Each regression is at the location-month level, weighted by number of hires.

A Proofs

A.1 The impact of testing on hires

We cannot sign the difference between expected match quality before and after testing. We instead performed a simulation exercise to understand in general what this difference looks like. Specifically, we drew parameter values for $\sigma_a^2, \sigma_\epsilon^2, \sigma_b^2$, each independently from a $U(0, 20)$, μ_Y from a $U(-30, 0)$, and k from a $U(0, 1)$. We fixed p_G and W at their empirical values (0.6 and 0.18, respectively). We fixed μ_G so that the unconditional mean of a would be 0 (to match the unconditional mean of b).

We then simulated 10,000 applicants, where an applicant consists of a $\{t, a, b, s\}$ draw. We obtained $E[a|I^{testing}]$ and $E[a|I^{notesting}]$, calculated manager utility $((1-k)E[a|I] + kb$, hired the top 1800 candidates under each information structure and compared the average a of hires.[35]

We performed 100 simulations. For every simulation we found that $E[a|Hire, I^{testing}] > E[a|Hire, I^{notesting}]$.[36]

A.2 Preliminaries

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

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 $\tilde{z_t} = \frac{\underline{u} - (1-k)\mu_t}{\sqrt{(1-k)^2\sigma + k^2\sigma_b^2}}$.

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

[35] I stands for the information available at the time of hire: $I^{notesting} = \{s, b\}$ and $I^{testing} = \{s, b, t\}$

[36] For a minority of cases, the difference was within sampling error. Sampling error is generated from the fact that we must perform a simulation for each $\{s, t\}$ pair to obtain the expected value of a. We can estimate the sampling error because we have an analytical solution for $E[a|I^{testing}]$ which we can compare to the simulated value.

The firm is interested in expected match quality conditional on being hired under Discretion. This can be expressed as follows, where $\lambda(.)$ is the inverse Mill's 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 match 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 match quality of hires equals μ_G.

A.3 Proof of Proposition 3.1

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 $k < k'$ match 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, match 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, match quality is higher under Discretion than No Discretion.*

For this proof we make use of the following lemma:

Lemma A.1 *The expected match 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 3.1

1. *For any given precision of private information, $1/\sigma_\epsilon^2 > 0$, there exists a $k' \in (0,1)$ such that if $k < k'$ match 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 match quality of hires equals μ_G). When $k = 0$, the manager hires only applicants whose expected match 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 A.1 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, match 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 A.1. 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, match 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 match 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.

A.4 Proof of Proposition 3.2

Across M, 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.

A.5 Proof of Proposition 3.3

If the quality of hired workers is decreasing in the exception rate, $\frac{\partial E[a|Hire]}{\partial R_m} < 0$ across M, 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.

Appendix Table A1: The impact of job testing for completed job spells Additional outcomes

	Mean Completed Duration (Days, Mean=211; SD=232)		>3 Months (Mean=0.62; SD=0.21)		>6 Months (Mean=0.46; SD=0.24)		>12 Months (Mean=0.32; SD=0.32)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Testing Used for Median Worker	88.89** *(35.91)*	47.00*** *(16.00)*	0.0404*** *(0.00818)*	0.0292 *(0.0234)*	0.0906*** *(0.00912)*	0.0565** *(0.0267)*	0.107*** *(0.00976)*	0.0806*** *(0.0228)*
N	4,401	4,401	4,505	4,505	4,324	4,324	3,882	3,882
Year-Month FEs	X	X	X	X	X	X	X	X
Location FEs	X	X	X	X	X	X	X	X
Client Firm X Year FEs		X		X		X		X
Location Time Trends		X		X		X		X

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

Notes: See notes to Table 2. The dependent variables are the mean length of completed job spells in days and the share of workers in a location-cohort who survive 3, 6, or 12 months, among those who are not right-censored.

Appendix Table A2: Impact of Color Score on Job Duration by Pre-testing Location Duration

	Log(Completed Job Spell)					
	High Duration	*Low Duration*	*High Duration*	*Low Duration*	*High Duration*	*Low Duration*
	(1)	(2)	(3)	(4)	(5)	(6)
Green	0.165***	0.162***	0.161***	0.172***	0.170***	0.163***
	(0.0417)	*(0.0525)*	*(0.0406)*	*(0.0541)*	*(0.0481)*	*(0.0514)*
Yellow	0.0930**	0.119**	0.0886**	0.130***	0.0990**	0.113**
	(0.0411)	*(0.0463)*	*(0.0403)*	*(0.0481)*	*(0.0467)*	*(0.0465)*
N	23,596	32,284	23,596	32,284	23,596	32,284
Year-Month FEs	X	X	X	X	X	X
Location FEs	X	X	X	X	X	X
Client Firm X Year FEs			X	X	X	X
Application Pool FEs					X	X

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

Notes: Each observation is an individual hire-month, for hired workers post testing only. The omitted category is red workers. Locations are classified as high duration if their mean duration pre-testing was above median for the pre-testing sample.

Appendix Table A3: Impact of Color Score on Job Duration by Location-Specific Exception Rates

	Log(Completed Job Spell)					
	High Exception Rate	*Low Exception Rate*	*High Exception Rate*	*Low Exception Rate*	*High Exception Rate*	*Low Exception Rate*
	(1)	(2)	(3)	(4)	(5)	(6)
Green	0.173***	0.215***	0.172***	0.205***	0.178***	0.170***
	(0.0317)	*(0.0689)*	*(0.0307)*	*(0.0711)*	*(0.0312)*	*(0.0606)*
Yellow	0.112***	0.182**	0.112***	0.174**	0.111***	0.137**
	(0.0287)	*(0.0737)*	*(0.0280)*	*(0.0760)*	*(0.0278)*	*(0.0656)*
N	36,088	31,928	36,088	31,928	36,088	31,928
Year-Month FEs	X	X	X	X	X	X
Location FEs	X	X	X	X	X	X
Client Firm X Year FEs			X	X	X	X
Application Pool FEs					X	X

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

Notes: Each observation is an individual hire-month, for hired workers post testing only. The omitted category is red workers. Locations are classified as high exception rate duration if their mean exception rate post-testing was above median for the post-testing sample.