Discretion in Hiring

Mitchell Hoffman
University of Toronto

Lisa B. Kahn
Yale University & NBER

Danielle Li
Harvard University

April 18, 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Please do not cite or circulate without permission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 hiring 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. This paper examines whether firms should rely on hard metrics or grant managers discretion in making hiring decisions. 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%. Further, when faced with similar applicant pools, we find that 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. In this setting, we find that firms can improve productivity by limiting managerial discretion.

*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, Arthur Campbell, Alex Frankel, Jin Li, Liz Lyons, Steve Malliaris, Mike Powell, Kathryn Shaw, Steve Tadelis, and numerous seminar participants. All errors are our own.

1 Introduction

Firms face both an information and an agency problem when making hiring decisions. 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 do not perfectly align with firm objectives.[1]

In recent years, however, the increasing adoption of "workforce analytics" and job testing has provided firms with new tools for hiring.[2] Job testing has the potential to both improve information about the quality of candidates and to reduce agency problems between firms and their hiring 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.

How should firms use this information, if at all? In the absence of agency problems, firms should allow managers 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.

This paper examines this question empirically. Using a unique personnel dataset on hiring managers, job applicants, and hired workers across sample firms who adopt job testing, we present two key findings. First, job testing substantially improves the match quality of workers that a firm is able to hire: workers hired with job testing have 15% longer tenures than those hired without testing. Second, managers who make more exceptions to test recommendations are more likely to hire workers with lower match quality. In our setting, this is implies that firms can further improve match quality by limiting managerial discretion and placing more weight on the job test.

[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, Rivera (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/.

We begin with a model in which firms rely on potentially biased hiring managers who observe private signals of worker quality. This model illustrates the central concern that firms face in deciding whether to grant managers discretion: because the information that a manager learns from subjective evaluations is unverifiable, a firm cannot tell whether a manager hires a candidate with poor test scores because he or she has evidence to the contrary or because he or she is simply mistaken or biased.

Using this model, we develop a simple empirical diagnostic for assessing whether firms can improve the match quality of their hires by limiting managerial discretion in favor of relying on on test recommendations. The intuition is as follows: 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 firms should grant discretion. 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.

To address these issues, we obtain 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 such as those engaged in data entry or call center work. Prior to the introduction of testing, our sample firms employed hiring managers who conducted interviews and made hiring decisions. After the introduction of testing, hiring managers were also given access to a test recommendation for each worker: green (hire), yellow (possibly hire), red (do not hire).[3] 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 workers that a firm hires. 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

[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 a rating: green (high potential candidate), yellow (moderate potential candidate), or red (lowest rating).

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). Using our model as a guide, we relate a manager's likelihood of making exceptions to job tests to the eventual outcomes of his or her hires. Across a variety of specifications, we find that exercise of discretion is strongly correlated with worse outcomes. Even when faced with applicant pools that are identical in terms of test results, managers that make more exceptions systematically hire worse workers. Our model shows that this negative correlation implies that firms could improve 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 matches. 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.[4] 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 adop-

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

tion of on board computers in the trucking industry. It is also related to papers on bias, discretion, and rule-making in other settings.[5]

We also contribute to a small, but growing literature on the impact of screening technologies on the quality of hires.[6] 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.

Finally, our work is also relevant to a broader literature on hiring and employer learning.[7] Finally, 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.[8]

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 6 concludes.

[5]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.

[6]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)).

[7]A central literature in labor economics emphasizes that imperfect information generates substantial problems for allocative efficiency in the labor market. The canonical work of Jovanovic (1979) points out that match quality is an experience good that is likely difficult to discern before a worker has started the job. The employer learning literature, beginning with empirical work by Farber and Gibbons (1996) and Altonji and Pierret (2001), provides evidence that firms likely learn about a worker's general productivity as a worker gains experience in the labor market, and Kahn and Lange (2014) show that this learning continues to be important even late into a worker's lifecycle because their productivity constantly evolves. This literature suggests imperfect information is a substantial problem facing those making hiring decisions.

[8]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.

2 Setting and Data

Over the past decades, 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 primary product offered by the data firm is a test designed to predict performance for a particular job in the low-skilled service sector. Our data agreement prevents us from revealing the exact nature of the job, but it is conducted in a non-retail environment and is similar to data entry work or telemarketing. Applicants are administered an online questionnaire consisting of a large battery of personality, cognitive skills, and job scenario questions. This 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 single *green, yellow, red* job test score.

Our data firm provides its services to many clients (hereafter, "client firms"). We have personnel records for 15 such client firms, with workers dispersed across 130 locations in the U.S. 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.[9] Each client firm shared these records with our data firm, once a partnership was established.[10] From this point, our data firm keeps records of all applicants,

[9]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.

[10]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 workers employed when data collection began. We infer the date when data collection began as the first recorded termination date. Workers hired before this date may have only been a selected sample of the full set of hires, those that survived up to that date. We drop these workers from our sample, but have experimented with including them along with flexible controls for being "stock sampled" in our regressions.

their test scores, and the unique ID of the hiring 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. After testing, firms made scores available to managers and encouraged them to factor scores into hiring decisions, but authority over hiring decisions was still delegated to managers.

In the first part of this paper, we examine the impact of testing technology on worker productivity. 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.[11] 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 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.[12] 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 hiring manager responsible for them (only after testing), we have 585 such hiring managers (or "recruiters") in our data. Post-testing, when we have information on applicants as well as hires, we have nearly 94,000 hires and 690,000 applicants who were not hired.

[11] 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.

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

We will find it useful to define an "applicant pool" as a group of workers under consideration for a job at the same location, month and with the same manager. 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.[13] From Table 1, we have 4,529 such applicant pools in our data consisting of, on average 268 applicants. On average, 14% of workers in a given pool are hired.

Table 1 also shows the distribution of test scores and the associated hire probabilities. Roughly 40% of all applicants 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. Greens and yellows are hired at a rate of roughly 20%, while only 8% of reds are hired. Still it is very likely for the test to be "over-ruled". In three-quarters of the pools a yellow is hired when a green also applied. In nearly a third of the pools, a red was hired when greens or yellows were available. These summary statistics foreshadow substantial variation in testing adoption that we exploit later.

3 Model

We formalize a model in which a firm makes hiring decisions with the help of a hiring manager. There are two sources of information about the quality of job candidates. First, interviews generate unverifiable information about a candidate's quality which is privately observed by the hiring 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

[13] 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.

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, which is unobserved to all parties. The distribution among applicants is known to be $a \sim N(0, \sigma_0^2)$. The firm's objective is to hire a proportion, W, of workers that maximizes expected match quality.[14]

The firm employs hiring managers whose interests are imperfectly aligned with that of the firm. In particular, a manager's expected payoff of 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 weight k, 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.[15]

[14] 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).

[15] 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 manager privately observes information about a_i and b_i. First, the manager observes a noisy signal s_i, of match quality:

$$s_i = a_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ is independent of a_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 $\sigma_\epsilon^2 = \infty$.

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 $k = 1$. We assume for simplicity that b_i is perfectly observed by the hiring manager, while the firm observes neither b nor k.

Let M denote the set of managers in a firm. For a given manager, $m \in M$, his or her type may be 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.

Partway through our sample, a public signal of match quality (a) becomes available in the form of a test score, t. Let $t \in \{G, Y\}$; a share of workers p_G are type G, and a share $p_Y = 1 - p_G$ are type Y.[16] We assume that the distribution of applicants does not change with the introduction of testing. We furthermore assume that the posterior distribution of a conditional on t is given by: $a_i \sim N(\mu_a^t, \sigma_a^2)$ with $\mu_a^G > \mu_a^Y$. Type G workers have a higher expected match quality but the same variance as type Y workers.[17] Naturally, the

[16] We assume that $W < p_G$, i.e., there are always enough Type G applicants to meet the hire rate requirement. This matches our data well where the average hire rate is 14% and the share of applicants with a G score is 40%. We furthermore assume that $t \perp \epsilon_i$.

[17] This distribution falls out of a Bayesian updating on a based on the prior and a continuous, normally distributed signal, t. The posterior mean conditional on the signal is a weighted average of the prior and the new signal. Under the normality assumptions, the posterior variance takes the same value, regardless of the signal.

posterior variance, σ_a^2 is less than the variance in the population, σ_0^2, since we condition on new information.[18]

Managers form a posterior expectation of worker quality given both their private signal and the test signal. Managers 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 is at least some threshold $\bar{u}$. 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 m under this policy as $E[a|Hire, m]$.

3.2 Model Predictions

The introduction of testing raises the question of how much firms should rely on their managers, versus relying on hard test information. We generate a number of predictions related to hiring and worker quality, given the setup described above, where managers are able to choose how to use job test recommendations. We then discuss how these predictions can be used to assess whether firms would benefit from the policy of eliminating discretion by relying only on test recommendations.

Eliminating discretion need not 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 *Holding managerial type $(k, 1/\sigma_\epsilon^2)$ constant, the match quality of hires is weakly greater with the test signal than without.*

[18]For simplicity, we assume the test signal is 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.

Intuitively, the test improves the quality of hires because it adds new information about match quality. This improves overall match quality for a fixed proportion of hires, assuming no impact of testing on the composition of the applicant pool. If there is no new information (or when $k = 0$) then managers are free to ignore it and be at least as well off as before.

Proposition 3.2 *Across the set of managers M, $E[a|Hire, m]$ is decreasing in managerial bias, k, and weakly increasing in the precision of the manager's private information, $1/\sigma_\epsilon^2$.*

Here, when bias is smaller (e.g. k is lower), the manager focuses more on maximizing a, and therefore hires higher quality workers than a manager who focuses more on maximizing b. Similarly, as long as the manager cares about a ($k \neq 0$), more precise private information (higher $1/\sigma_\epsilon^2$) improves the match quality of hires.

Proposition 3.2 demonstrates that firms face a tradeoff when granting discretion: managers may bring valuable private information about the match quality of workers, but exploiting this information leaves the firm prone to managerial bias. While discretion may be the optimal regime when interviews were the only signals of quality available to firms, the introduction of testing may change the marginal costs and benefits that firms face in continuing to grant discretion.

To determine whether firms could benefit from limiting managerial discretion post-testing, 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 if the firm made decisions based on test recommendations alone: that is, every 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. Recall that a manager's

type is defined by two parameters: the degree of bias k and the precision of the manager's private information, $1/\sigma_\epsilon^2$. The following two propositions show how exceptions can be used to infer whether discretion benefits firms after testing.

Proposition 3.3 *Across the set of managers M, R_m is increasing in both managerial bias, k, and the precision of the manager's private information, $1/\sigma_\epsilon^2$.*

Proposition 3.3 shows that increases in exceptions can be driven by both more information and more bias. 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.

As such it is difficult to discern whether granting discretion is beneficial to firms simply by examining how often managers make exceptions. Instead, Propositions 3.2 and 3.3 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, R_m, are increasing in both managerial bias and the value of the manager's private information, match quality, $E[a|Hire, m]$, is only increasing in information. If across managers, $E[a|Hire, m]$ 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 3.4 *If the quality of hired workers is decreasing in the exception rate, $\frac{\partial E[a|Hire,m]}{\partial R_m} < 0$ across M, then firms can improve outcomes by eliminating 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 Predictions

To summarize, our model makes the following empirically testable predictions:

1. The introduction of testing improves the match quality of hired workers.

2. If the quality of workers is decreasing in the probability of an exception then no discretion improves match quality relative to discretion

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 thus estimate these regressions at the location-time (month by year) level, the level of variation underlying our key explanatory variable.[19] 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 sometimes augment this specification by including controls for client-firm by year fixed effects. These control for any changes to HR or other policies that our client firms may be making in conjunction with the implementation of job testing. For example, our client firms may be implementing job testing precisely because they are facing problems with retention, or because they are attempting to improve their workplace. Our estimates of the effect of job testing, then, may reflect the net effect of all these changes. Including client-year fixed effects means that we compare locations in the same firm in the same year, some of which receive job testing sooner than others. The identifying assumption in this specification is that, within a firm, locations that receive testing sooner vs. later were on parallel trends before testing.

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, some of our specifications include controls

[19]This aggregation affords substantial savings on computation time, and, given we weight by number of hires, results will be idetical to those from a worker-level regression.

for location-specific time trends. These also account for broad trends that may impact worker retention, including, for instance, smooth changes in local labor market conditions. Finally, some of our specifications also include detailed 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, respectively. We use these controls for proxy for any time-varying, location specific changes to the desirability of a location that are not captured by the location-specific time trends. 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. Many canonical models of job search, e.g. Jovanovic 1979, predict a positive correlation between match quality and job duration. Moreover, our client firms employ low-skill service sector workers, a segment of the labor market that experiences notoriously high turnover. Figure 1 shows a histogram of job tenure for completed spells (75% of the spells in our data). 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. As a result, hiring and training make up a substantial part of the labor costs in our client firms. 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 the majority of 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 271,393 hired workers that we observe in our sample, 75%, or

202,977 workers have completed spells, with an average spell lasting 202 days. The key explanatory variable is whether or not the median hire at this location-date was tested.

In the baseline specification (Column 1) we find that employees hired with the assistance of job testing stay, on average, 0.272 log points longer, or 31%, significant at the 5% level. Column 2 introduces controls for client firm by year fixed effects to control for implementation of new strategies and HR policies at the firm level.[20] Doing so reduces the magnitude of our estimated coefficient by roughly a third, and we lose statistical significance. Column 3 introduces location-specific time trends to account for the possibility that locations may be in general improving or worsening, or local labor market conditions may be trending in a particular direction. 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 0.134 log points or 15%, significant at the 5%-level.

Finally, in Column 4, we augment our preferred specification with controls for the quality of the applicant pool. 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: once candidates 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 testing improves the ability of managers to identify productive workers, rather than by exclusively altering the quality of the applicant pool. Overall, the range of estimates in Table 2 are in line with previous estimates found in Autor and Scarborough (2008).

Table 3 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

[20] In interviews, our data firm has indicated that it was not aware of any other client-specific policy changes.

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), from Table 2.

We also explore a range of other duration-related outcomes to ensure that the estimated positive impact of testing is not due to functional form. 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.[21] We aggregate this variable to measure the proportion of hires in a location-cohort that meet each duration milestone. We also report results for mean completed spells in days (rather than the log).

Table 4 reports our estimates of the impact of testing on these variables, using both our baseline set of controls as well as our full set (Columns 1 and 4 of Table 2). We find that testing increases the probability that a worker reaches specific tenure milestones, with the largest impacts on the 6-to-12 month survival margin. Testing increases the proportion of applicants persisting longer than 6 months by 5.7 percentage points or about 12 percent (Column 6). Column 8 shows that testing increases persistence past a year by 8.1 percentage points, or 18 percent.

Figure 2 provides further evidence that the introduction of testing is not correlated with location-specific characteristics that could also impact worker duration. Here we plot 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

[21] 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.

treatment effect of testing appears to grow over time, suggesting either that hiring managers and other participants might take some time to learn how to use the test effectively.

Event studies for our 3, 6, and 12 month persistence outcomes look quite similar. Though estimates in this graph are noisy, the fact that observed location quality and location-trends in quality appear to be uncorrelated with timing of testing should help alleviate any concerns that the timing of testing introduction across locations was systematically related to unobservables at the location-level.

To summarize, we find that the introduction of job testing leads to a 15% improvement in the duration of hires. This effect comes primarily from increasing retention beyond 6 months. This increase in productivity may be driven by several mechanisms. Most intuitively, the adoption of job testing may improve the productivity of hires at a location by providing managers with new information about the quality of job candidates, or by making this information more salient.

Controls for client-by-year effects rule out that discrete changes at the client firm-level, for example policies adopted concurrent with testing that were also designed to improve outcomes of hires, drive our results. Controls for location fixed effects and location-specific time trends rule out the possibility that our results are driven by unobserved location characteristics that are either fixed or that vary smoothly with time. Controls for applicant pool characteristics show that job testing improves the quality of hires by providing information about candidates, rather than by altering the applicant pool.

Finally, we provide a complementary test of the information content of testing: we look for evidence that color score is indeed predictive of performance, by comparing job durations for green, yellow, and red hires. A simple model of hiring would predict that a firm equates productivity of the *marginal* green, yellow, and red hire. Thus we should see identical outcomes for the marginal hires from each color score. However, it should still be the case (if the test yields valuable information) that *on average* greens outperform yellows, who outperform reds.

Across a range of specifications, Table 5 shows that both green and yellow workers perform better relative to red workers. Greens, on average, stay roughly 19% longer than reds, while yellows stay 13% longer. These differences are statistically significant at the 1% level, as are the differences between greens and yellows. These results are based on the selected sample of hired workers; if managers use other signals to make hiring decisions, then a red or yellow worker would only be hired if her unobserved quality were particularly high. As such, we might expect the unconditional performance difference between green, yellow, and red applicants to be even larger.

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? Should firms grant managers discretion in how to use the test, or should they use the test to create hiring rules that limit managerial discretion?

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 1) have other information suggesting that the applicant will be successful or 2) have 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 3.4, 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.

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 is likely to vary according to other characteristics: some locations may be inherently more desirable, applicant pools may vary in size, quality of applicants, and number of hires required, and these may separately impact the probability of making an exception. Thus, in order to apply our theory to the data, we need to isolate variation in exceptions that comes from managerial information and preferences.

We do this in two ways: first, we define an empirical "exception rate" that accounts for characteristics of the applicant pool that may also impact exceptions. Second, we include additional controls in our regressions to address remaining factors other than managerial type that may also drive exceptions. We discuss these below.

5.1 Defining Exceptions

Our data allow us to see test scores of all 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 both on the manager's choices and on the type of applicants and the number of slots the manager needs to hire. For example, if all workers in a pool are green then, mechanically, this pool cannot have any exceptions. 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.

The client firms in our sample report telling 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. Despite this fairly uniform advice, we see substantial variation in the extent to which managers actually follow test recommendations when making hiring decisions. 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 sets of figures, the median exception rate is about 20% of the maximal number of possible exceptions. While there are many pools for which we see very few or even zero exceptions, these appear to be driven by either small pools or idiosyncratic factors at the pool level. When we weight by number of applicants and aggregate to the manager or location level, we see that almost all managers and locations make some exceptions on average.

5.2 Empirical Specifications

The most direct implementation of Proposition 3.4 examines the correlation between the exception rate and the realized match quality of hires in the post-testing period:

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

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 details controls for the quality and number of applicants in an application pool. This is important because our observed exception rates may be driven by factors other than managerial type. For example, some locations may be inherently less desirable than others. These locations would attract lower quality managers and lower quality applicants: managers may make more exceptions because they are biased, and hired workers may have lower tenures, but both facts are driven by unobserved location characteristics. Our inference in this case would be complicated by the fact that the desirability of the location generates a correlation between the quality managers and the quality of applicants.

A downside of this approach, however, 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 match quality of workers may be lower. In this case, a manager is 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 counter-

factual wrong because exceptions are correlated with variation in unobserved quality within color.

These two sources of confounding variation—unobserved heterogeneity and small sample idiosyncrasies—are naturally at odds with the spirit of Proposition 3.4. Our ideal experiment would randomly assign managers to a large number of applicant pools and then ask whether managers who systematically make more exceptions across all these pools do worse. The randomization would eliminate concerns about unobserved hetereogeneity while the large numbers would limit the extent to which our identifying variation is driven by small sample variation in the applicant pool. In practice, we perform two additional analyses to address these concerns.

First, rather than comparing outcomes across applicant pools with different exception rates, we take advantage of pre-testing data to estimate whether the *impact* of testing 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) studies how the impact of testing, as described in Section 4, 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, for example, that locations a positive correlation between quality of manager and unobserved quality of applicants. By including observations from a location prior to the introduction of testing,

we expand the sample on which we can estimate location-specific fixed effects. This means that our remaining variation in exception rates is less likely to be driven purely by small sample variation, compared with Equation (3).

Finally, we also estimate versions of Equations (3) and (4) where we aggregate the exception rate to the manager-level, rather than the pool-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—in the average exception rate for a given manager, across managers—is more likely to represent exceptions made because of managerial type (bias and information). As a final robsutness, we aggregate to the location level, rather than the manager level, to take into account any systematic assignment of managers to applicants within a location that might be correlated with exception rates and applicant quality. This last approach also helps us rule out any measurement error generated by the matching of applicants to hiring 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. In Section 5.4, we describe additional robustness checks.

5.3 Results

Figure 4 plots the raw correlation between the exception rate in a pool and the log completed duration of workers hired from that pool. In the first panel, each applicant pool has the same weight; in the second, panels 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 hires, as measured by the mean log of their completed durations.

Table 6 presents the regression analogue to figure 4, using two measures of the exception rate: a standardized exception rate with mean 0 and standard deviation 1 (Columns 1 and 2), or an indicator variable for whether that applicant pool had above or below median

exceptions (Columns 3 and 4). Columns 1 and 3 report specifications with our baseline set of controls—location and time fixed effects only—while Columns 2 and 4 report results with our full set of controls.

Column 1 indicates that a one standard deviation increase in the exception rate of a pool is associated with a 3.1% lower completed tenure for that group. This figure is insignificant, but including additional location-year trends and client-year controls, as reported in Column 2, increases our estimate to 3.9%, significant at the 5% level. We find a similar pattern using an indicator for high exception rates; we find that workers hired from pools with above-median exception rates have 3.6% lower tenure upon completion of the job spell.

Table 7 examines how the impact of testing varies by the extent to which managers make exceptions. The explanatory variable in Columns 1 and 2 is the interaction between an applicant pool's post-testing standardized exception rate and an indicator for being in the post-testing period. The coefficient on the main effect of testing represents the impact of testing for a location with the median number of exceptions, and reports estimates very similar to those in Table 2. In Column 1, we find that locations with median exception rates post testing experience a 0.277 log point increase in duration (32%) as a result of the implementation of testing. Each standard deviation increase in exception rates lowers the impact of testing by 0.10, or about a third of the effect at the median. Column 2 includes additional controls and this modulates our effect: we find that the median location experiences a 0.22 log point increase in duration of hires as a result of testing, but that each standard deviation increase in exceptions offsets this effect by 0.048. Using indicators for whether a pool had an above-median exception rate (Columns 3 and 4) yields qualitatively similar results.

We also point out that Columns 2 and 4 of Tables 6 and 7 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 identical 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 the workers they hire perform systematically worse. In both Tables 6 and 7, we see that this is not the case: workers hired from cohorts with more exceptions do worse.

To better illustrate the variation underlying the results in Table 7, 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.

Finally, we report results where exception rates have been aggregated to the manager- or location-level. These specificaitons help to alleviate concerns that idiosyncratic factors at the pool-level, unrelated to managerial type, drive our results. Appendix Tables A1 and A2 report analogues to Tables 6 and 7 but with this aggregation. As can be seen we obtain very similar results.

We therefore find that for both applicant pools and managers with higher exception rates quality of hires is lower. Based on Proposition 3.4, 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 centralizing decision-making.

5.4 Additional Robustness Checks

In this section, we address several specific concerns about our identification strategy.

Our results in Table 7 show that the impact of testing is smaller when managers have higher exception rates. Recall, Proposition 3.4 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. As such, while measuring the impact of testing allows for a number of important control variables, as explained above, it also hinders our ability to interpret a negative correlation between outcomes and exceptions as evidence that firms should limit discretion. We would expect a smaller impact of testing on quality of hires if managers were biased since bias leads to both more exceptions and lower quality of hires.

However, exceptions may also be driven by better private information. At the extreme, unbiased managers with perfect information would make many exceptions and receive no benefit from testing.

To infer that firms would be better off limiting discretion, it must be that exception rates are largely not driven by better private information. We test this by looking at the correlation between exception rates and the pre-testing quality of hires. Figure 6 plots this relationship at the location-level, since we cannot match hires to specific managers before testing. For every duration outcome measured, we see that there is no significant relationship between a location's exception rate and the quality of its hires. This result suggests that exceptions are not driven by managers that always had better private information. Furthermore, we can rule out that locations with more exceptions are otherwise undesirable, which might then generate a negative correlation between exceptions and job durations.

As an additional test of whether exceptions are driven by better information on the part of managers, we can also look directly at applicants hired as an exception. If our results were driven by the fact that better informed managers make more exceptions, then it should be the case that exceptions are correct: a yellow hired as an exception should perform better than a green who is not hired. In practice, it is not possible to compare the performance of workers hired as exceptions against those who were passed up, because we cannot generally observe performance for this latter group. However, we approximate this by comparing 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 8 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. 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 8 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 of the first panel indicates, comparing workers from the same applicant pool, that passed over greens stay about 8% longer than the yellows hired before them in the same pool. Column 3 of the second panel indicate that exception reds are even worse: passed over greens stay almost 19% longer and exception yellows stay almost 12% longer, than the reds they were passed over for.

The results in Table 8 mean that it is quite unlikely that exceptions are driven by better information. When workers with better test scores appear to be at first passed over, and then later hired, they still perform better than the workers chosen first. However, one alternative explanation is that applicants with higher test scores were not initially passed up, but instead were 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 be unwilling to pass up the chance to work if employment is offered.

Futhermore, Appendix Table A3 provides some evidence on this matter, by testing the relationship between the length of time until hired and duration outcome. Workers with longer gaps between application date and hire date (which we treat as temporarily passed over applicants) would likely have better outcomes if they initially had better outside options and that drove their delay in taking the job. Instead we find no significant difference in outcomes for workers hired immediately (the omitted category), compared to those who waited one, two, or three months before starting, holding constant test score. 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 initially been passed over.

Finally, we examine the hypothesis that the usefulness of the test varies across locations in a way that might be correlated with overall location quality. For very undesirable locations, green applicants might have better outside options and be more difficult to retain.

A manager attempting to maximize job retention may optimally decide to make exceptions in order to hire lower ranked workers with lower outside options, in order to avoid costly retraining of new workers. In this case, a negative correlation between exceptions and performance would not necessarily imply that firms could improve productivity by relying more on tests. We examine this by comparing the relative performance of green, yellow, and red workers by pre-testing durations. If in locations that always had lower durations, green applicants leave quickly, then higher exception rates would be justified, especially if hired reds and yellows stayed for longer. Table 9 shows that this is unlikely to be the case: we find that hired green workers have longer durations than hired yellow or red (the omitted category) workers across both above-median and below-median pre-testing average durations. Appendix Table A4 repeats this exercise, splitting the sample by high and low exception rate in the post period. Again, there is no difference in the relative performance of hired greens, yellows, and reds across high and low exception locations.

6 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 locations tend to hire appliance with the best test scores while other locations make many more exceptions. Across a range of specifications, we show that the exercise of discretion is associated with worse outcomes.

If managers make exceptions when they correctly believe that applicants will be better, we would not expect there to be a systematic, negative correlation between the extent to which managers make exceptions and the quality of their hires. Our results instead suggest that managers underweight the job test relative to what the firm may prefer. It also suggests that firms may want to take advantage of the verifiability of job test scores to impose hiring rules that centralize hiring authority.

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

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

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

[16] 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.

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

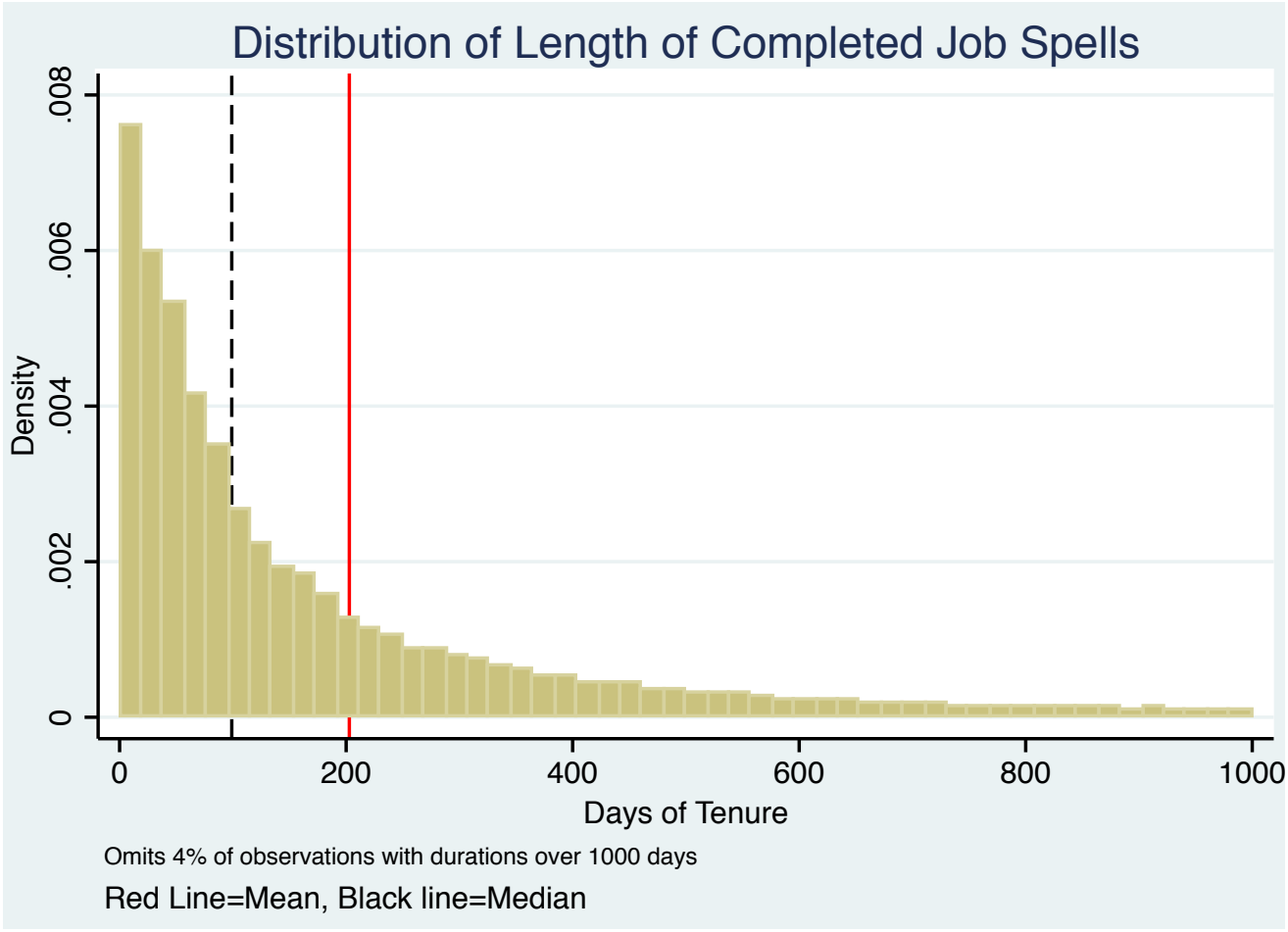
[18] Paravisini, D. and A. Schoar (2013) "The Incentive Effect of IT: Randomized Evidence from Credit Committees" NBER Working Paper #19303.

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

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

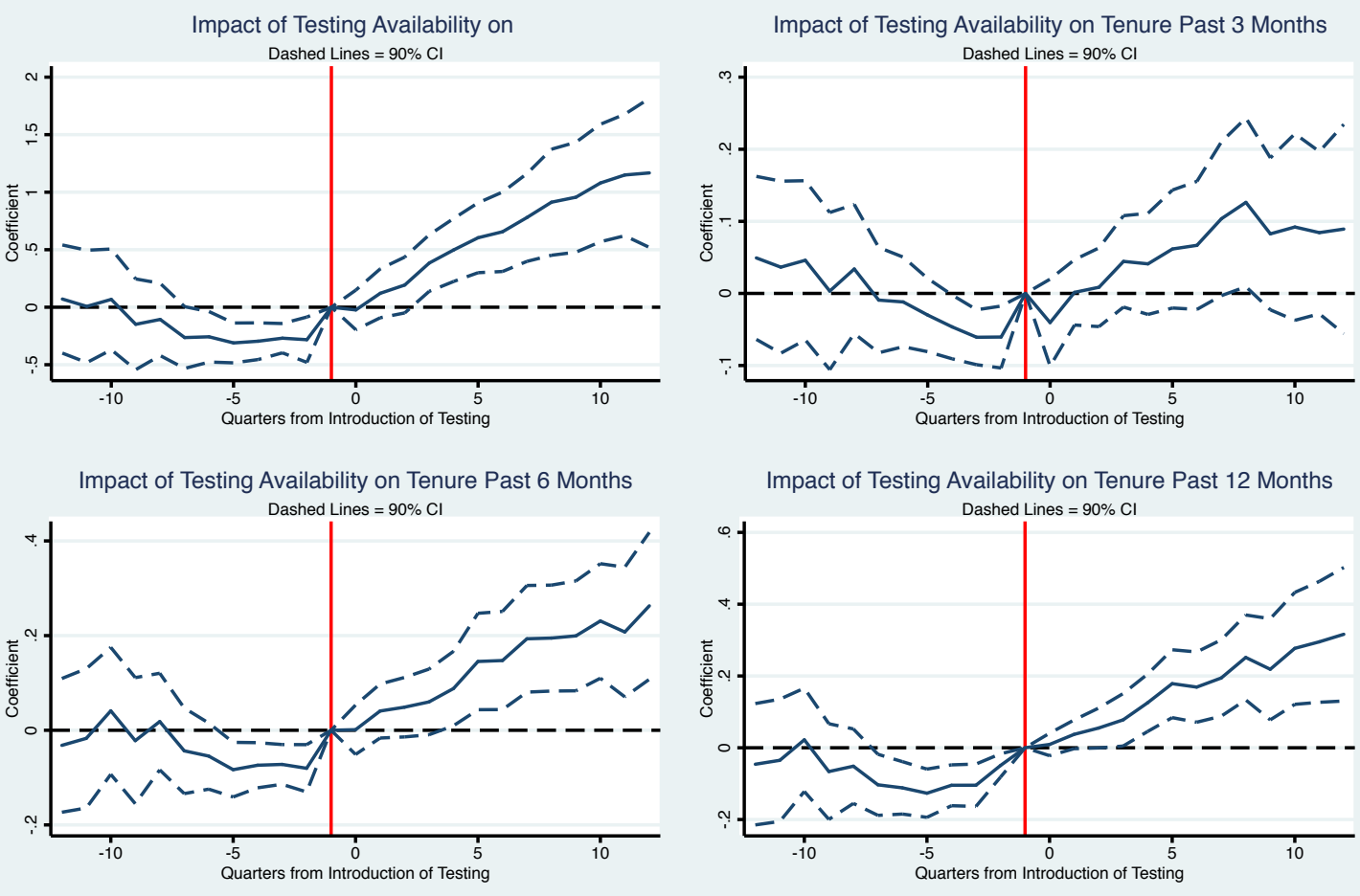
[21] Stanton, C. and C. Thomas (2014), "Landing The First Job: The Value of Intermediaries in Online Hiring," mimeo London School of Economics.

FIGURE 1: DISTRIBUTION OF COMPLETED SPELLS



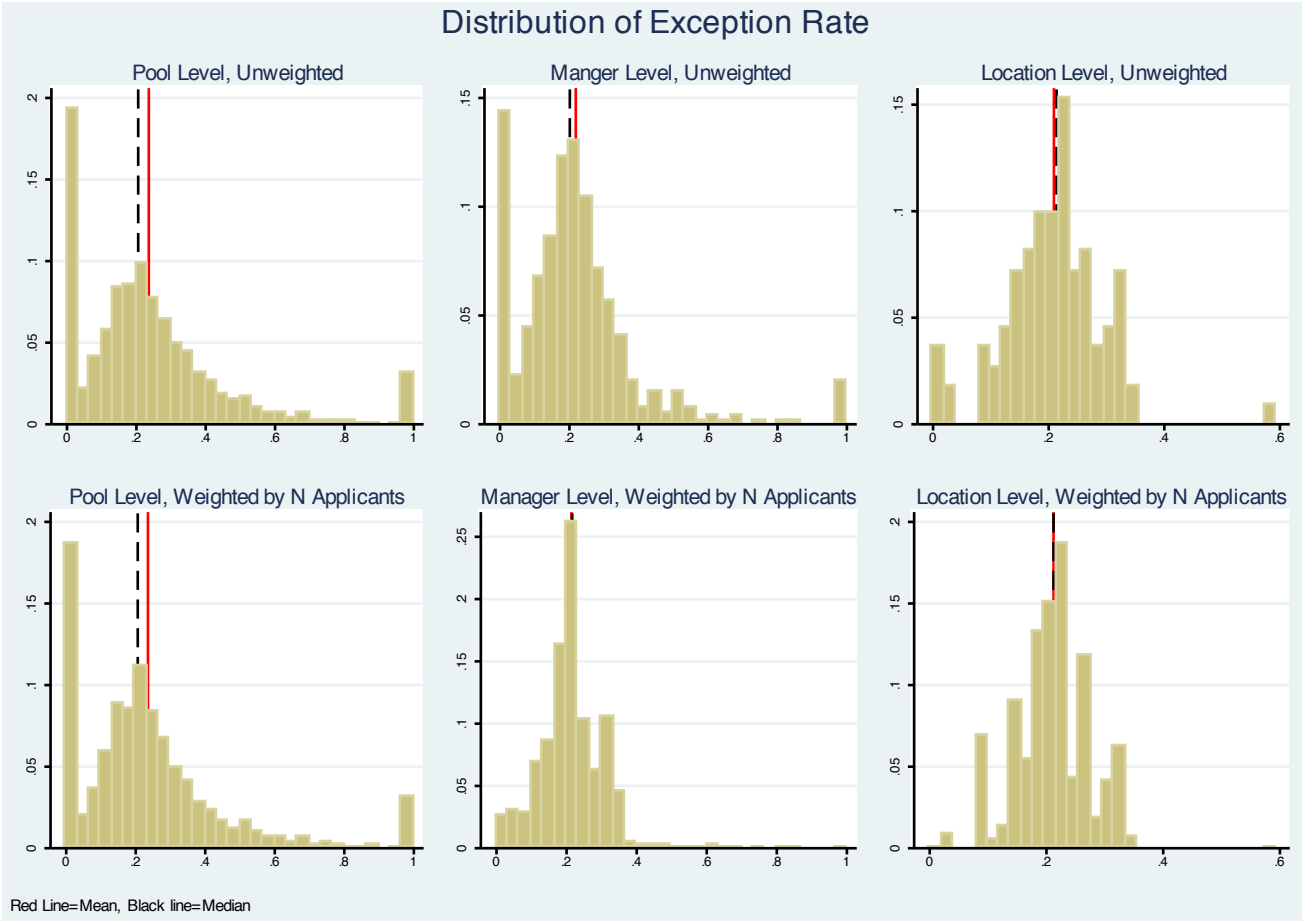
NOTES: Figure 1 plots the distribution of completed job spells at the individual level.

Figure 2: Event Study of Duration Outcomes



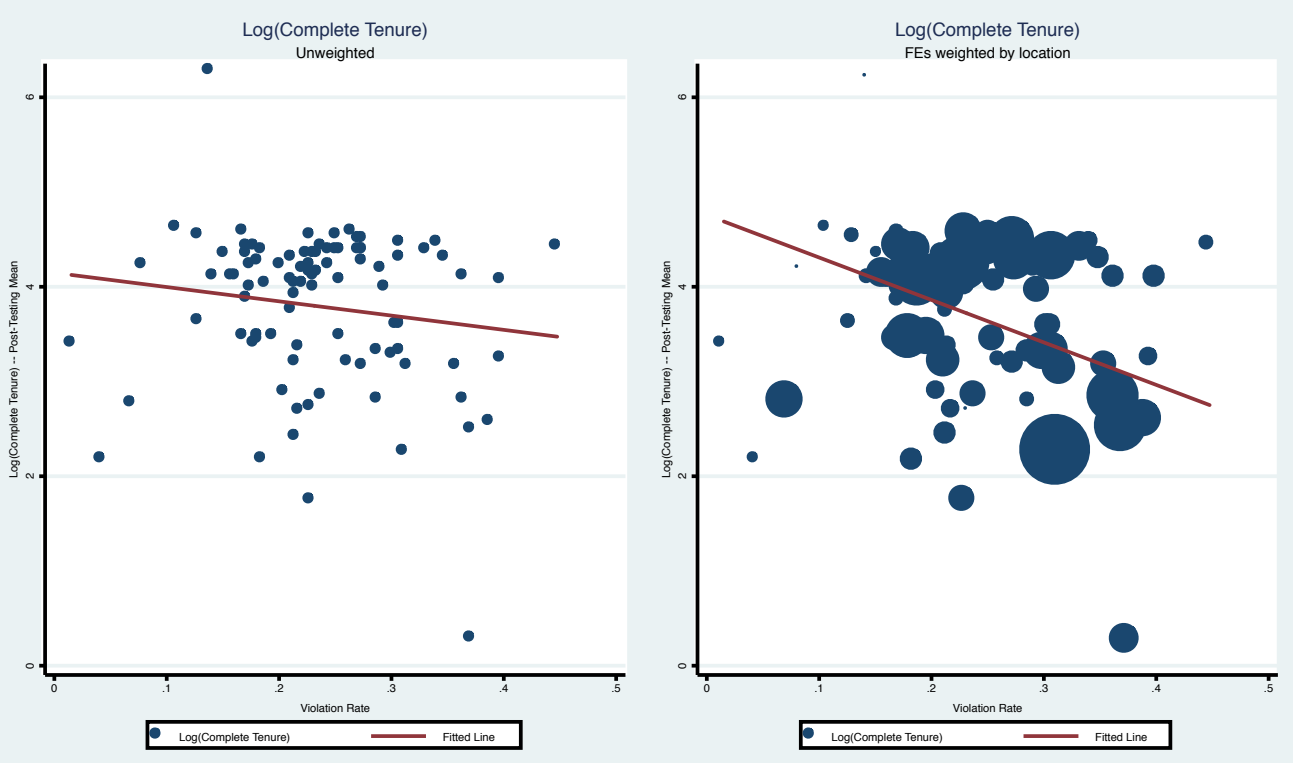
Notes: These figures plot the coefficient on the impact testing by quarter. The underlying estimating equation is given by $\text{Log(Duration)}_{lt} = \alpha_0 + \alpha_1 \text{Testing}_{lt} + \delta_l + \gamma_t + \epsilon_{lt}$. This regression does not control for location-specific time trends; if those are present, they would be visible in the figure.

Figure 3: Variation in Application Pool Exception Rate



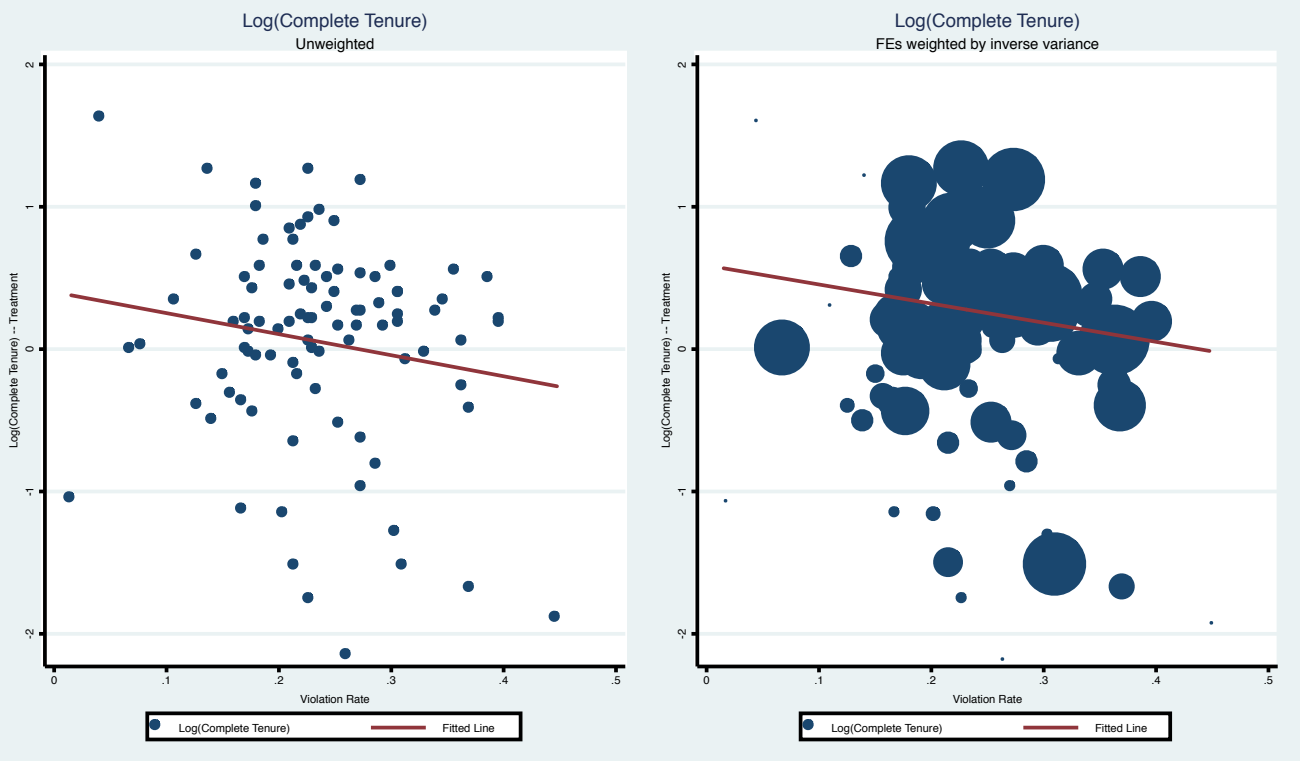
Notes: These figures plot the distribution of the exception rate, as defined by Equation (2) in Section 5. The leftmost panel present results at the applicant pool level (defined to be a manager–location–month). The middle panel aggregates these data to the manager level and the rightmost panel aggregates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

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

	All	*Location Pre-Testing*	*Location Post-Testing*
Sample Coverage			
# Locations	131	116	111
# Hired Workers	270,086	176,390	93,696
# Applicants			1,132,757
# Recruiters			555
# Applicant Pools			4,209
Worker Performance			
Mean Completed Spell (Days)	198.1	233.6	116.8
(Std. Dev.)	(267.9)	(300.8)	(139.7)
% > 3 Months	0.62	0.66	0.53
	(0.49)	(0.47)	(0.50)
% > 6 Months	0.46	0.50	0.35
	(0.50)	(0.50)	(0.48)
% > 12 Months	0.31	0.35	0.19
	(0.46)	(0.48)	(0.39)
Applicant Pool Charactertisics (unweighted averages)			
# Applicants			268
# Hired			22.2
Share Green Applicants			0.47
Share Yellow Applicants			0.32
Share of Greens Hired			0.22
Share of Yellows Hired			0.18
Share of Reds Hired			0.08

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 hiring 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)
	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)*
# Location-Time Observati	4,401	4,401	4,401	4,401
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: 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.

Table 3: The impact of job testing for completed job spells
Individual is tested

	(1)	(2)	(3)	(4)
		Log(Completed Job Spell)		
Individual Applicant is Tested	0.195* *(0.115)*	0.139 *(0.124)*	0.141** *(0.0637)*	0.228** *(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: Observations are at the individual level. Testing is defined as whether or not an individual worker has a score.

Table 4: 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**	47.00***	0.0404***	0.0292	0.0906***	0.0565**	0.107***	0.0806***
	(35.91)	*(16.00)*	*(0.00818)*	*(0.0234)*	*(0.00912)*	*(0.0267)*	*(0.00976)*	*(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.

Table 5: The information content of test scores

	Log(Completed Job Spell)		
	(1)	(2)	(3)
Green	0.176***	0.177***	0.175***
	(0.0304)	*(0.0302)*	*(0.0294)*
Yellow	0.122***	0.125***	0.121***
	(0.0296)	*(0.0292)*	*(0.0287)*
N	70,631	70,631	70,631
Year-Month FEs	X	X	X
Location FEs	X	X	X
Client Firm X Year-Month FEs		X	X
Location X Year-Month FEs			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.

Table 6: Pool Level Exception Rates and Post-Testing Duration

	(1)	(2)	(3)	(4)
	Location-Cohort Mean Log Duration of Completed Spells			
Standardized Exception Rate Post Testing	-0.0310	-0.0385**		
	(0.0213)	*(0.0192)*		
> Median Exception Rate Post Testing			-0.0230	-0.0353*
			(0.0225)	*(0.0200)*
N	3,926	3,926	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 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 7: Pool Level Exception Rates and the Impact of Job Testing

	(1)	(2)	(3)	(4)
	Location-Cohort Mean Duration of Completed Spells (Days)			
Testing Used for Median Worker	*0.277***	0.217**	0.329**	0.251***
	(0.112)	*(0.0876)*	*(0.133)*	*(0.0891)*
Standardized Exception Rate Post Testing	-0.101**	-0.0477***		
	(0.0446)	*(0.0182)*		
> Median Exception Rate Post Testing			-0.0800	-0.0545**
			(0.0594)	*(0.0226)*
N	6,830	6,830	6,956	6,956
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 of Applicant Pool		X		X

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

Notes: See notes to Table 2. 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. It is 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 8: Match Quality of Exceptions vs. Passed Over Applicants

	Log(Completed Job Spell)		
	(1)	(2)	(3)
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
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 Hire Month 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 is the passed over greens. The second panel includes red exceptions and passed over greens and yellows only. Red exceptions are the omitted category.

Table 9: Information Content of the Test, by Pre-period 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-Month FEs			X	X	X	X
Location X Year-Month 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.

A Proofs

A.1 Proof of Proposition 3.1

Holding managerial type $(k, 1/\sigma_\epsilon^2)$ constant, the match quality of hires is weakly greater with the test signal than without.

Proof We must show that the expected match quality of hired workers conditional on s_i, b, and t (denoted $E[a|Hire, s_i, b, t]$) is greater than the expectation of hires conditional on s_i and b (denoted $E[a|Hire, s_i, b]$).

Let u^1 and u^2 be the hiring thresholds before and after testing is adopted, respectively, that fix the hire rate at W. Let ϕ_a^1 be the pdf of $E[a|s_i, b]$ and ϕ_a^2 be the pdf of $E[a|s_i, b, t]$. Because the latter is an expectation conditional on more variables, the distribution is weakly more variant. In order to keep the hire rate fixed at W, the threshold $\underline{u}$ must be increasing in this variance – under the normality assumption, there is more mass above the cut point when the distribution has a higher variance.

Let ϕ_b be the pdf of b. Then we can write the probability of being hired, as a function of ϕ_a^1 or ϕ_a^2, as follows, and we know that the thresholds are chosen so that the hire probability is always W.

$$\int_{-\infty}^{\infty}\int_{\frac{u^1-ky}{1-k}}^{\infty}\phi_a^1(x)\phi_b(y)dxdy = \int_{-\infty}^{\infty}\int_{\frac{u^2-ky}{1-k}}^{\infty}\phi_a^2(x)\phi_b(y)dxdy = W \tag{5}$$

Since $\frac{u^2-ky}{1-k} > \frac{u^1-ky}{1-k}$ we must have

$$\int_{-\infty}^{\infty}\int_{\frac{u^2-ky}{1-k}}^{\infty}x\phi_a^2(x)\phi_b(y)dxdy > \int_{-\infty}^{\infty}\int_{\frac{u^1-ky}{1-k}}^{\infty}x\phi_a^1(x)\phi_b(y)dxdy$$

Finally, we point out that the expectation of hires under testing must be weakly greater than that without testing, regardless of the distributional assumptions. An optimizing manager would simply ignore the test if it contained no information, so hiring decisions must be at least as good as without the test.

A.2 Proof of Proposition 3.2

Across M, $E[a|Hire, m]$ is decreasing in managerial bias, k, and weakly increasing in the precision of the manager's private information, $1/\sigma_\epsilon^2$

Proof First, consider $k' > k''$. We can write the following, where u' is the hiring threshold implied when $k = k'$ that keeps the hiring rate at W, and similar for u''.

$$E[a|Hire, k'] - E[a|Hire, k''] = \frac{1}{W} \int_{-\infty}^{\infty} \int_{\frac{u'-(1-k')x}{k'}}^{\frac{u''-(1-k'')x}{k''}} x\phi_b(y)\phi_a(x)dydx$$

It is easy to show that for all $x > \hat{x}$, $\frac{u''-(1-k'')x}{k''} < \frac{u'-(1-k')x}{k'}$, and for all $x < \hat{x}$, $\frac{u''-(1-k'')x}{k''} > \frac{u'-(1-k')x}{k'}$. Also, since the hire rate is fixed, we must have:

$$\int_{-\infty}^{\infty} \int_{\frac{u'-(1-k')x}{k'}}^{\infty} \phi_b(y)\phi_a(x)dydx = \int_{-\infty}^{\infty} \int_{\frac{u''-(1-k'')x}{k''}}^{\infty} \phi_b(y)\phi_a(x)dydx$$

It is easy to show the following

$$\int_{-\infty}^{\hat{x}} \int_{\frac{u'-(1-k')x}{k'}}^{\frac{u''-(1-k'')x}{k''}} \phi_b(y)\phi_a(x)dydx = \int_{\hat{x}}^{\infty} \int_{\frac{u''-(1-k'')x}{k''}}^{\frac{u'-(1-k')x}{k'}} \phi_b(y)\phi_a(x)dydx$$

But then it is immediate that

$$\int_{-\infty}^{\hat{x}} \int_{\frac{u'-(1-k')x}{k'}}^{\frac{u''-(1-k'')x}{k''}} x\phi_b(y)\phi_a(x)dydx < \int_{\hat{x}}^{\infty} \int_{\frac{u''-(1-k'')x}{k''}}^{\frac{u'-(1-k')x}{k'}} x\phi_b(y)\phi_a(x)dydx$$

Since we can rewrite $E[a|Hire, k'] - E[a|Hire, k'']$ as

$$\int_{-\infty}^{\hat{x}} \int_{\frac{u'-(1-k')x}{k'}}^{\frac{u''-(1-k'')x}{k''}} x\phi_b(y)\phi_a(x)dydx - \int_{\hat{x}}^{\infty} \int_{\frac{u''-(1-k'')x}{k''}}^{\frac{u'-(1-k')x}{k'}} x\phi_b(y)\phi_a(x)dydx$$

we are done.

Second, consider $v' < v''$. The former reflects a less precise private interview signal. Let u' be the hiring threshold when $1/\sigma_\epsilon^2 = v'$ and similar for u''. It is easy to show that $u'' > u'$. When the signal is more precise, the total variance of expected quality increases, and therefore the hiring threshold must increase (as noted above). Because the hire rate is fixed at W, we must then have the following, where ϕ'_a is the pdf of $E[a|s_i, b, t]$ when $1/\sigma_\epsilon^2 = v'$ and ϕ''_a is the pdf for v''.

$$\int_{-\infty}^{\infty} \int_{\frac{u'-ky}{1-k}}^{\infty} \phi'_a(x)\phi_b(y)dxdy = \int_{-\infty}^{\infty} \int_{\frac{u''-ky}{1-k}}^{\infty} \phi''_a(x)\phi_b(y)dxdy$$

Since $\frac{u'-ky}{1-k} < \frac{u''-ky}{1-k}$ we must then have:

$$\int_{-\infty}^{\infty}\int_{\frac{u'-ky}{1-k}}^{\infty} x\phi'_a(x)\phi_b(y)dxdy < \int_{-\infty}^{\infty}\int_{\frac{u''-ky}{1-k}}^{\infty} x\phi''_a(x)\phi_b(y)dxdy$$

Thus $E[a|Hire, v''] > E[a|Hire, v']$.

A.3 Proof of Proposition 3.3

Aross M, 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.

The hiring threshold for hiring type Y must be decreasing in both k and $1/\sigma_\epsilon^2$. When the manager is more biased or when the private signal is more valuable, the manager places less weight on the test, and be willing to take more chances on Y candidates. When the hiring threshold decreases, it is more likely that any given Y will be hired.

A.4 Proof of Proposition 3.4

If the quality of hired workers is decreasing in the exception rate, $\frac{\partial E[a|Hire,m]}{\partial R_m} < 0$ across M, then firms can improve outcomes by eliminating 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 $\frac{\partial E[a|Hire,m]}{\partial R_m} < 0$, 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.

Appendix Table A1: The impact of job testing on completed job spells Location-Month Level

	(1)	(2)	(3)	(4)
	Location-Cohort Mean Log Duration of Completed Spells			
Standardized Exception Rate Post Testing	-0.0881**	-0.0432		
	(0.0376)	*(0.0435)*		
> Median Exception Rate Post Testing			-0.0394	-0.0729
			(0.0255)	*(0.0551)*
N	1,507	1,507	1,516	1,516
Year-Month FEs	X	X	X	X
Location FEs	X	X	X	X
Client Firm X Year FEs		X		X
Size and Composition of Applicant Pool		X		X

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

Notes: See notes to Table 6. The method is identical except for that variables have been aggregated to their location-month means to reduce the impact of pool to pool variation in unobserved applicant quality.

Appendix Table A2: The impact of job testing on completed job spells Location-Month Level

	(1)	(2)	(3)	(4)
	Location-Cohort Mean Duration of Completed Spells (Days)			
Testing Used for Median Worker	*0.261***	0.143	0.351**	0.199**
	(0.111)	*(0.102)*	*(0.145)*	*(0.0994)*
Standardized Exception Rate Post Testing	-0.173**	-0.0773**		
	(0.0773)	*(0.0319)*		
> Median Exception Rate Post Testing			-0.140	-0.105***
			(0.0921)	*(0.0389)*
N	4,392	4,392	4,401	4,401
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: See notes to Table 7. The method is identical except for that variables have been aggregated to their location-month means to reduce the impact of pool to pool variation in unobserved applicant quality.

Appendix Table A3: 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-Month FEs		X	X
Location X Year-Month 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.

Appendix Table A3: Information Content of the Test, by Exception Rate

	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-Month FEs			X	X	X	X
Location X Year-Month 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.