Discretion in Hiring

Mitchell Hoffman
University of Toronto

Lisa B. Kahn
Yale University & NBER

Danielle Li
Harvard University

March 2016

Abstract

Who should make hiring decisions? We propose an empirical test for assessing whether firms should rely on hard metrics such as job test scores or grant managers discretion in making hiring decisions. We implement our test in the context of the introduction of a valuable job test across 15 firms employing low-skill service sector workers. Our results suggest that firms can improve worker quality by limiting managerial discretion. This is because, when faced with similar applicant pools, managers who exercise more discretion (as measured by their likelihood of overruling job test recommendations) systematically end up with worse hires.

JEL Classifications: M51, J24

Keywords: Hiring; rules vs. discretion; job testing

---

*Correspondence: Mitchell Hoffman, University of Toronto Rotman School of Management, 105 St. George St., Toronto, ON M5S 3E6. Email: mitchell.hoffman@rotman.utoronto.ca. Lisa Kahn, Yale School of Management, 165 Whitney Ave, PO Box 208200, New Haven, CT 06511. Email: lisa.kahn@yale.edu. Danielle Li, Harvard Business School, 211 Rock Center, Boston, MA 02163. Email: dli@hbs.edu. We are grateful to Jason Abaluck, Ricardo Alonso, David Berger, Arthur Campbell, David Deming, Alex Frankel, Harry Krashinsky, Jin Li, Liz Lyons, Steve Malliaris, Mike Powell, Kathryn Shaw, Steve Tadelis, and numerous seminar participants. We are grateful to the anonymous data provider for providing access to proprietary data. Hoffman acknowledges financial support from the Social Science and Humanities Research Council of Canada. All errors are our own.

1 Introduction

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

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

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

In this paper, we evaluate the introduction of a job test and develop a diagnostic to inform how firms should incorporate it into their hiring decisions. Using a unique personnel dataset on HR managers, job applicants, and hired workers across 15 firms employing low-skilled service sector workers, we present two key findings. First, the adoption of job testing substantially improves the quality of hired workers, as measured by job tenure: those hired

[1] For example, a manager could have preferences over demographics or family background that do not maximize productivity. In a case study of elite professional services firms, Riviera (2012) shows that one of the most important determinants of hiring is the presence of shared leisure activities.

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

with job testing have about 15% longer tenures than those hired without testing. In our setting, job tenure is a key measure of quality because turnover is costly and workers already spend a substantial fraction of their tenure in paid training. Second, managers who overrule test recommendations hire workers with shorter eventual job tenures. This second result suggests that managers exercise discretion because they are biased or have poor judgement, not because they are better informed. This implies that firms in our setting can further improve worker quality by limiting managerial discretion and placing more weight on the test.

Our paper makes the following contributions. First, we provide new evidence that managers systematically make hiring decisions that are not in the interest of the firm. Second, we show that job testing can improve hiring outcomes not simply by providing more information, but by making information verifiable, and thereby expanding the scope for contractual solutions to agency problems within the firm. Finally, we develop a simple test for assessing the value of discretion in hiring. Our test uses data likely available to many firms with job testing and is applicable to a wide variety of settings where at least one correlate of productivity is available.

We begin with a model that reflects how our sample firms make hiring decisions: potentially biased HR managers observe both public (job test) and private (interview) signals of worker quality and are permitted to use their discretion in choosing which applicants to hire. Using this model, we develop an empirical diagnostic for whether firms can improve the quality of their hires by limiting discretion and relying only on the job test. Intuitively, the value of discretion can be inferred from how effectively managers choose to overrule test recommendations. Discretion is valuable when managers make exceptions to test recommendations based on superior private information about a worker's quality. A manager with a more precise signal of worker quality is both more likely to make exceptions to test recommendations and to hire workers who are a better fit. By contrast, a manager with biases or poor judgment will also be more likely to make exceptions, but will hire workers with worse outcomes. In the latter case, firms can improve outcomes by limiting discretion.

We apply this test using data from an anonymous firm that provides online job testing services to client firms. Our sample consists of 15 client firms who employ low-skill service-

sector workers. Prior to the introduction of testing, firms employed HR managers involved in hiring new workers. After the introduction of testing, HR managers were also given access to a test score for each applicant: green (high potential candidate), yellow (moderate potential candidate), or red (lowest rating).[3] Managers were encouraged 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 quality of hired workers. By examining the staggered introduction of job testing across our sample locations, we show that cohorts of workers hired with job testing have about 15% longer tenures than cohorts of workers hired without testing. We provide a number of tests in the paper to ensure that our results are not driven by the endogenous adoption of testing or by other policies that firms may have concurrently implemented.

This finding suggests that job tests contain valuable information about the quality of candidates. Next, we ask how firms should use this information, in particular, whether firms should limit discretion by relying more on test recommendations, relative to the status quo. A unique feature of our data is that we observe applicants as well as hired workers. We can thus observe exceptions: when a manager hires a worker with a test score of yellow and a green goes unhired (or similarly, when a red is hired above a yellow or green). As explained above, the correlation between a manager's likelihood of making these exceptions and eventual outcomes of hires can inform whether allowing discretion is beneficial from the firm's perspective. Across a variety of specifications, we find that exceptions are strongly correlated with worse outcomes. Even when facing applicants that are identical in terms of test scores, managers who make more exceptions systematically hire workers who are more likely to quit or be fired.

Finally, we show that our results are unlikely to be driven by the possibility that managers sacrifice job tenure in search of workers who have higher quality on other dimensions. If this were the case, limiting discretion may improve worker durations, but at the expense of other quality measures. To assess whether this is a possible explanation for our findings, we examine the relationship between hiring, exceptions, and a direct measure of productivity, daily output per hour, which we observe for a subset of firms in our sample. Based

[3]Section 2 provides more information on the job test.

on this supplemental analysis, we see no evidence that firms are trading off duration for higher productivity. Taken together, our findings suggest that placing more weight on job test recommendations will result in better hires for the firm.

This empirical approach differs from an evaluation in which discretion is granted to some managers and not others. Rather, managers in our data have the right to overrule test recommendations, but they differ in the extent to which they choose to do so. Our approach uses variation in this willingness to exercise discretion to understand whether discretion improves hiring. Specifically, we compare outcomes of workers hired by managers who make more exceptions, to those hired by managers who follow test recommendations more closely. The consequences (in terms of worker outcomes) of managerial choices reveals whether exceptions are driven primarily by better information or by biases.

The validity of this approach relies on two key assumptions. First, exceptions must be reflective of managerial choices, and not driven by mechanical factors or lower yield rates for high-quality applicants. Second, unobserved quality must be similar across low- and high-exception cohorts. We provide extensive discussion of both assumptions in the text and conclude that, after careful empirical treatment, they are not important confounds in this setting. Our approach thus has the advantage of not requiring exogenous variation in discretion regimes, which may be hard to locate in observational data or potentially artificial to engineer in a randomized trial. Instead, we make greater use of theory to make firm policy prescriptions.

As data analytics becomes more frequently applied to human resource management decisions, it becomes increasingly important to understand how these new technologies impact the organizational structure of the firm and the efficiency of worker-firm matching. While a large theoretical literature has studied how firms should allocate authority, ours is the first paper to provide an empirical test for assessing the value of discretion in hiring.[4] Our findings provide direct evidence that screening technologies can help resolve agency problems

[4]For theoretical work, see Bolton and Dewatripont (2012) for a survey and Dessein (2002) and Alonso and Matouschek (2008) for particularly relevant instances. There is a small empirical literature on bias, discretion and rule-making in other settings. For example, Paravisini and Schoar (2012) find 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. Kuziemko (2013) shows that the exercise of discretion in parole boards is efficient, relative to fixed sentences. Wang (2014) finds that loan officers may improve lending decisions by using soft information.

by improving information symmetry, and thereby relaxing contracting constraints. In this spirit, our paper is related to the classic Baker and Hubbard (2004) analysis of the adoption of on board computers in the trucking industry.

We also contribute to a small, but growing literature on the impact of screening technologies on the quality of hires.[5] Our work is most closely related to Autor and Scarborough (2008), the first paper in economics to provide an estimate of the impact of job testing on worker performance. 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.

Our work is also relevant to a broader literature on hiring and employer learning.[6] 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.[7]

Lastly, our results are broadly aligned with findings in psychology and behavioral economics that emphasize the potential of machine-based algorithms to mitigate errors and biases in human judgement across a variety of domains.[8]

The remainder of this paper proceeds as follows. Section 2 describes the setting and data. Section 3 evaluates the impact of testing on the quality of hires. Section 4 presents a model of hiring with both hard and soft signals of quality and derives the empirical diagnostic for whether firms should limit managerial discretion by relying more on hard signals. Section 5 applies the diagnostic to our empirical setting. Section 6 concludes.

[5]Other screening technologies include labor market intermediaries (e.g., Autor (2001), Stanton and Thomas (2014), Horton (2013)), and employee referrals (e.g., Brown et al., (2015), Burks et al. (2015) and Pallais and Sands (2015)).

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

[7]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.

[8]See Kuncel et. al. (2013) for a meta-analysis of this literature, Kahneman (2011) for a behavioral economics perspective, and Kleinberg at al. (2015) for empirical evidence that machine-based algorithms outperform judges in deciding which arrestees to detain pre-trail.

2 Setting and Data

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

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

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

In its marketing materials, our data firm emphasizes the ability of its job test to reduce worker turnover, which is a perennial challenge for firms employing low skill service sector workers. To illustrate this concern, Figure 1 shows a histogram of job tenure for completed spells (75% of the spells in our data) among employees in our sample. The median worker (solid red line) stays only 99 days, or just over 3 months. Twenty percent of hired workers leave after only a month. At the same time, our client firms generally report spending the first several weeks training each new hire, during which time the hire is being paid.[9] Correspondingly, our analysis will also focus on job retention as the primary measure of hiring quality. For a subset of our client firms we also observe a direct measure of worker

[9]Each client firm in our sample provides paid training to its workforce. Reported lengths of training vary considerably, from around 1-2 weeks to around a couple months or more.

productivity: output per hour.[10] Because these data are available for a much smaller set of workers (roughly a quarter of hired workers), we report these findings separately when we discuss alternative explanations.

Prior to testing, our client firms gave their managers discretion to make hiring decisions or recommendations based on interviews and resumes.[11] After testing, firms made scores available to managers and encouraged them to factor scores into hiring recommendations, but authority over hiring decisions was still typically delegated to managers.[12]

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

In the first part of this paper, we examine the impact of testing technology on worker quality, as measured by tenure. For any given client firm, testing was rolled out gradually at roughly the location level. During the period in which the test is being introduced, not all applicants to the same location received test scores.[13] We therefore impute a location-specific date of testing adoption. Our preferred metric for the date of testing adoption is the first date in which at least 50% of the workers hired in that month and location have a test score. Once testing is adopted at a location, based on our definition, we impose that testing is thereafter always available.[14] In practice, this choice makes little difference and we are robust to a number of other definitions: whether any hire in a cohort is tested, whether the individual is tested, and instrumenting for whether an individual is tested with whether any, or the majority, of applicants in the cohort is tested.

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

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

[12] We do not directly observe authority relations in our data. However, drawing on information for several client firms (information provided by the data firm), managers were not required to hire strictly by the test.

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

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

Table 1 provides sample characteristics. Across our whole sample period we have nearly 300,000 hires; two-thirds of these were observed before testing was introduced and one-third were observed after, based on our preferred imputed definition of testing. Once we link applicants to the HR manager responsible for them (only after testing), we have 555 such managers in the data.[15] These managers primarily serve a recruiting role, and are unlikely to manage day-to-day production. Post-testing, when we have information on applicants as well as hires, we have nearly 94,000 hires and a total of 690,000 applicants.

Table 1 also reports worker performance pre- and post-testing, by job test score. On average, greens stay 12 days (11%) longer than yellows, who stay 17 days (18%) longer than reds. These differences are statistically significant and hold up to the full range of controls described below. This provides some evidence that test scores are indeed informative about worker performance. Even among the selected sample of hired workers, better test scores predict longer tenures. We might expect these differences to be even larger in the overall applicant population if managers hire red and yellow applicants only when unobserved quality is particularly high. On our productivity measure, output per hour, which averages roughly 8, performance is fairly similar across color.

3 The Impact of Testing

3.1 Empirical Strategy

Before examining whether firms should grant managers discretion over how to use job testing information, we first evaluate the impact of introducing testing information itself. To do so, we exploit the gradual roll-out in testing across locations and over time, and examine its impact on worker quality, as measured by tenure:

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

[15] The HR managers we study are referred to as recruiters by our data provider. Other managers may take part in hiring decisions as well. One firm said that its recruiters will often endorse candidates to another manager (e.g., a manager in operations one rank above the frontline supervisor) who will make a "final call."

Equation (1) compares outcomes for workers hired with and without job testing. We regress a productivity outcome (Outcome_{lt}) for workers hired to a location l, at time t, on an indicator for whether testing was available at that location at that time (Testing_{lt}) and controls. In practice, we define testing availability as whether the median hire at that location-date was tested, though we discuss robustness to other measures. As mentioned above, the location-time-specific measure of testing availability is preferred to using an indicator for whether an individual was tested (though we also report results with this metric) because of concerns that an applicant's testing status is correlated with his or her perceived quality. We estimate these regressions at the location-time (month-by-year) level, the level of variation underlying our key explanatory variable, and weight by number of hires in a location-date.[16] The outcome measure is the average outcome for workers hired to the same location at the same time.

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

Our outcome measures, Outcome_{lt}, summarize the average length of the employment relationship for a given lt cohort. We focus on tenure measures for several reasons. The length of a job spell is a measure that both theory and the firms in our study agree is important. Canonical models of job search (e.g., Jovanovic 1979), predict a positive correlation between match quality and job duration. Moreover, as discussed in Section 2, our client firms employ low-skill service sector workers and face high turnover and training costs: several weeks of paid training in a setting where the median worker stays only 99 days (see Figure 1.) Job duration is also a measure that has been used previously in the literature, for example by Autor and Scarborough (2008), who also focus on a low-skill service sector setting (retail). Finally, job duration is available for all workers in our sample.

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

3.2 Results

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

In the baseline specification (Panel 1, Column 1 of Table 2) we find that employees hired with the assistance of job testing stay, on average, 0.272 log points, or 31% longer, significant at the 5% level. In the subsequent columns we cumulatively add controls. Column 2 adds client firm-by-year fixed effects, to control for the implementation of any new strategies and HR policies that firms may have adopted along with testing.[17] Column 3 adds location-specific time trends to account for the possibility that the timing of the introduction of testing is related to trends at the location level, for example, that testing was introduced first to locations that were on an upward (or downward trajectory). Column 4 adds fixed effects for the number of green, yellow, and red applicants, which are set to zero pre-testing. The estimated impact of testing in column 4 thus holds applicant pool quality constant. These controls also absorb any impact of testing on the quality of applicants that a location receives. For instance, the introduction of testing may have a screening effect: as candidates gradually learn about testing, the least qualified may be deterred from applying. In general, coefficients decline in magnitude with the inclusion of stronger controls but are still sizable.

Panel 2 of Table 2 examines robustness to defining testing at the individual level, and we obtain similar results.[18] Because the decision to test an individual worker may be endogenous, we continue with our preferred metric of testing adoption (whether the median worker was tested). Overall, the range of estimates in Table 2 are broadly similar to previ-

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

[18] For these specifications we regress an individual's job duration (conditional on completion) on whether or not the individual was tested. Because these specifications are at the individual level, our sample size increases from 4,401 location-months to 202,728 individual hiring events.

ous estimates found in Autor and Scarborough (2008). With full controls, we find that the introduction of testing improves completed job tenures by about 15%.

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

We also explore a range of other duration-related outcomes to examine whether the impact of testing is concentrated at any point in the duration distribution. For each hired worker, we measure whether they stay at least three, six, or twelve months. For these samples, we restrict to workers hired three, six, or twelve months, respectively, before the data end date. These milestone measures thus allow us to examine the impact of testing for all workers, not just those with completed spells. We aggregate these variables to measure the proportion of hires in a location-cohort that meet each duration milestone. Regression results (analogous to those reported in Panel 1 of Table 2 are reported in Appendix Table A1, while event studies are shown in the remaining panels of Figure 2. For each of these measures, we again see that testing improves job durations, and we see no evidence of any pre-trends.

4 Model

In this section, we formalize a model in which a firm makes hiring decisions with the help of an HR manager. This model has two purposes. First, it builds intuition for the tradeoff a firm faces when deciding whether to grant managers discretion to make hiring decisions or whether to base hiring decisions solely on the job test. Granting discretion enables firms

to take advantage of a manager's private information but comes at the cost of allowing for managerial biases and mistakes. Second, this model lets us derive an empirical diagnostic for whether firms that currently allow for discretion can improve hiring outcomes by relying more on test recommendations. We then apply this test in Section 5.

4.1 Setup

A mass one of applicants apply for job openings within a firm. The firm's payoff of hiring worker i is given by a_i. We assume that a_i is drawn from a distribution which depends on a worker's type, $t_i \in \{G, Y\}$; a share of workers p_G are type G, a share $1 - p_G$ are type Y, and $a|t \sim N(\mu_t, \sigma_a^2)$ with $\mu_G > \mu_Y$ and $\sigma_a^2 \in (0, \infty)$. This worker-quality distribution enables us to naturally incorporate the discrete test score into the hiring environment. We do so by assuming that the test publicly reveals t.[19]

The firm's objective is to hire a proportion, W, of workers that maximizes expected quality, $E[a|Hire]$.[20] For simplicity, we also assume $W < p_G$.[21]

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

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

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

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

[20] In theory, firms should hire all workers whose expected value is greater than their cost (wage). In practice, we find that having access to job testing information does not impact the number of workers that a firm hires. One explanation for this is that a threshold rule such as $E[a] > \overline{a}$ is not contractable because a_i is unobservable. Nonetheless, a firm with rational expectations will know the typical share W of applicants that are worth hiring, and W itself is contractable. Assuming a fixed hiring share is also consistent with the previous literature, for example, Autor and Scarborough (2008).

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

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 manager mistakes that drive them to prefer the wrong candidates.[22]

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

$$s_i = a_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ is independent of a_i, t_i, and b_i. The parameter $\sigma_\epsilon^2 \in \mathbb{R}_+ \cup \{\infty\}$ measures the level of the manager's information. A manager with perfect information on a_i has $\sigma_\epsilon^2 = 0$, while a manager with no private information has $\sigma_\epsilon^2 = \infty$.

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

Let M denote the set of managers in a firm. For a given manager, $m \in M$, his or her type is defined by the pair $(k, 1/\sigma_\epsilon^2)$, corresponding to the bias and precision of private information, respectively. These have implied subscripts, m, which we suppress for ease of notation. We assume firms do not observe manager type, nor do they observe s_i or b_i.

Managers form a posterior expectation of worker quality given both their private signal and the test signal. They then maximize their own utility by hiring a worker if and only if the expected value of U_i conditional on s_i, b_i, and t_i is at least some threshold. Managers thus wield "discretion" because they choose how to weigh the various signals about an applicant when making hiring decisions. We denote the quality of hires for a given manager under this policy as $E[a|Hire]$ (where an m subscript is implied).

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

4.2 Model Predictions

Our model focuses on the question of whether firms should rely on their managers, versus relying on hard test information. Firms can follow the set up described above, allowing their managers to weigh both signals and make ultimate hiring decisions (we call this the "Discretion" regime). Alternatively, firms may eliminate discretion and rely solely on test recommendations ("No Discretion").[23] In this section we generate a diagnostic for when one policy will dominate the other.

Neither retaining nor fully eliminating discretion need be the optimal policy response after the introduction of testing. Firms may, for example, consider hybrid policies such as requiring managers to hire lexicographically by the test score before choosing his or her preferred candidates, and these may generate more benefits. Rather than solving for the optimal hiring policy, we focus on the extreme of eliminating discretion entirely. This is because we can provide a tractable test for whether this counterfactual policy would make our client firms better off, relative to their current practice.[24] All proofs are in the Appendix.

Proposition 4.1 *The following results formalize conditions under which the firm will prefer Discretion or No Discretion.*

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

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

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

[23] Under this policy firms would hire applicants with the best test scores, randomizing within score to break ties.

[24] We also abstract away from other policies the firm could adopt, for example, directly incentivizing managers based on the productivity of their hires or fully replacing managers with the test.

Proposition 4.1 illustrates the fundamental tradeoff firms face when allocating authority: managers have private information, but they are also biased. In general, greater bias pushes the firm to prefer No Discretion, while better information pushes it towards Discretion. Specifically, the first finding states that when bias, k, is low, firms prefer to grant discretion, and when bias is high, firms prefer No Discretion. Part 2 states that when the precision of a manager's private information becomes sufficiently small, firms cannot benefit from granting discretion, even if the manager has a low level of bias. Uninformed managers would at best follow test recommendations and, at worst deviate because they are mistaken or biased. Finally, part 3 states that for any fixed information precision threshold, there exists an accompanying bias threshold such that if managerial information is greater and bias is smaller, firms prefer to grant discretion. Put simply, firms benefit from Discretion when a manager has very precise information, but only if the manager is not too biased.

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

Specifically, we observe cases in which managers exercise discretion to explicitly contradict test recommendations. We define a hired worker as an "exception" if the worker would not have been hired under No Discretion (i.e., based on the test recommendation alone): any time a Y worker is hired when a G worker is available but not hired.

Denote the probability of an exception for a given manager, $m \in M$, as R_m. Given the assumptions made above, $R_m = E_m[Pr(Hire|Y)]$. That is, the probability of an exception is simply the probability that a Y type is hired, because this is implicitly also equal to the probability that a Y is hired *over* a G.

Proposition 4.2 *The exception rate, R_m, is increasing in both managerial bias, k, and the precision of the manager's private information, $1/\sigma_\epsilon^2$.*

Intuitively, managers with better information make more exceptions because they then place less weight on the test relative to their own signal of a. More biased managers also

make more exceptions because they place more weight on maximizing other qualities, b. Thus, increases in exceptions can be driven by both more information and more bias.

It is therefore difficult to discern whether granting discretion is beneficial to the firm simply by examining how often managers make exceptions. Instead, Propositions 4.1 and 4.2 suggest that it is instructive to examine the relationship between how often managers make exceptions and the subsequent quality of their workers. Specifically, while exceptions (R_m) are increasing in both managerial bias and the value of the manager's private information, quality ($E[a|Hire]$) is decreasing in bias. If $E[a|Hire]$ is negatively correlated with R_m, then it is likely that exceptions are being driven primarily by managerial bias (because bias increases the probability of an exception and decreases the squality 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 4.3 *If the quality of hired workers is decreasing in the exception rate,* $\frac{\partial E[a|Hire]}{\partial R_m} < 0$*, then firms can improve outcomes by eliminating discretion. If quality is increasing in the exception rate then discretion is better than no discretion.*

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 quality of this manager's hires is equivalent to that of workers under a regime where hires are made based solely on the test. If increasing the probability of exceptions increases the quality of hires, then granting discretion improves outcomes relative to no discretion. If 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.

The key insight in our model is that exceptions, R_m, are driven by a manager's bias and information parameters. Because of this, the relationship between exception rates and worker quality is informative about whether managerial exceptions are primarily driven by bias—in which case discretion should be limited— or by information—in which case discretion should be allowed. In the following section, we discuss how we bring this intuition to the data.

5 Empirical Analysis on Discretion

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

In order to implement this test, we must address two issues that are outside the scope of the model. First, we must carefully define an "exception rate" that corresponds to a manager's choice to exercise discretion. For example, we should attribute more discretion to a manager when he or she hires a yellow applicant from a pool of 100 green applicants and 1 yellow applicant, rather than if he or she hires a yellow applicant from a pool of 1 green applicant and 100 yellow applicants. Second, we must ensure find appropriate comparison groups for applicant pools in which managers exercised more discretion by making more exceptions. A concern is that applicant pools in which managers make more exceptions may have lower unobservable applicant quality overall.

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

5.1 Defining Exceptions

To construct an empirical analogue of the exception rate R, we use data on hiring and test scores of applicants in the post-testing period. First, we define an "applicant pool" as a group of applicants being considered by the same manager for a job at the same location in the same month.[25]

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

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

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

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

The numerator of Exception Rate$_{mlt}$ counts the number of exceptions (or "order violations") a manager makes when hiring, i.e., the number of times a yellow is hired for each green that goes unhired plus the number of times a red is hired for each yellow and green that goes unhired.

The number of exceptions in a pool depends on both the manager's choices and on factors related to the applicant pool, such as size and color composition. For example, if a pool has only green applicants, it is impossible to make an exception. Similarly, if the manager hires all available applicants, then there can also be no exceptions. These variations were implicitly held constant in our model, but need to be accounted for in the empirics.

To isolate variation in exceptions that is driven by manager type, rather than by other confounding factors, 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 4 are derived for the probability of an exception, their proofs hold equally for this definition of an exception rate.[26]

From Table 1, we have 4,209 applicant pools in our data consisting of, on average 268 applicants.[27] On average, 19% of workers in a given pool are hired. Roughly 40% of all applicants in a given pool receive a "green", while "yellow" and "red" candidates make up roughly 30%, each. The test score is predictive of whether or not an applicant is hired. In

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

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

the average pool, greens and yellows are hired at a rate of roughly 20%, while only 9% of reds are hired. Still, managers very frequently make exceptions to test recommendations: the average applicant is in a pool where 24% of the maximal number of possible exceptions are made.

Furthermore, we see substantial variation in the extent to which managers actually follow test recommendations when making hiring decisions.[28] Figure 3 shows histograms of the exception rate, at the application pool level, as well as aggregated to the manager and location levels. The top panels show unweighted distributions, while the bottom panels show distributions weighted by the number of applicants.

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

Importantly, because we normalize exceptions by the maximum possible exceptions, variation in the exception rate is driven by differences in managerial choices, given observably identical applicant pools and hiring needs. That is, for identical numbers of red, yellow, and green applicants, and an identical number of hired workers, an exception rate will be higher if more yellows are hired above greens or more reds are hired above yellows and greens. The exception rate will thus not be impacted by *observable* differences in applicant pool quality. We discuss in the next subsection how we deal with unobserved differences in applicant quality, for example that greens produce a lower yield rate in one pool than greens in another.

[28] According to the data firm, client firms often told their managers that job test recommendations should be used in making hiring decisions but gave managers discretion over how to use the test (though some firms strongly discouraged managers from hiring red candidates).

5.2 Empirical Specifications

Proposition 4.3 examines the correlation between the exception rate and the realized 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 quality of hires is decreasing in the exception rate, meaning that firms can improve outcomes by eliminating discretion and relying solely on job test information.

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

These controls are important because the exception rate may be driven by differences in within-color quality across locations and pools. For example, some locations may be inherently less desirable than others, attracting both lower quality managers and lower quality applicants. The lower quality managers could be more biased, and, lower quality workers would be more likely to be fired. Both facts would be driven by unobserved location characteristics, not managerial bias. Furthermore, these locations might have a more difficult time attracting green applicants, even those who applied, resulting in higher exception rates. As long as this negative correlation is fixed across locations or varies smoothly at the location-level, our controls absorb this variation.

However, these biases could vary at the pool-level as well. For example, holding constant average quality and exception rate of a location, a given pool might have particularly low-quality greens. A manager would then rightly hire more yellows and reds that perform better than the unhired greens would have, but perhaps not better than the typical green hired by that manager. That is, for a given location, differences across pools in within-color quality could drive both variation in exceptions and outcomes. In addition, a local labor market shock (unobserved by the econometrician) could drive differences across pools in the ability

to attract greens for a given location. This type of pool-to-pool variation drives spurious correlations between exception rates and outcomes and is not accounted for by our controls.

To deal with the concern that Equation (3) relies too much on pool-to-pool variation in exception rates, we can aggregate exception rates to the manager- or location-level. Aggregating across multiple pools removes the portion of exception rates that are driven by idiosyncratic differences in within-color quality across pools. The remaining variation is driven by differences in the average exception rate across managers or locations.

To accommodate aggregate exception rates (which reduces or eliminates within-location variation in exception rates post testing), we expand our data to include pre-testing worker observations. Specifically, we estimate whether the *impact* of testing, as described in Section 3, varies with exception rates:

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

We can estimate this equation for the pool-level exception rate, Exception Rate_{mlt}, or for the average exception rate for a given manager or location across t.[29]

Equation (4) estimates how the impact of testing differs when managers make exceptions. The coefficient of interest is b_1. Finding $b_1 < 0$ indicates that making more exceptions decreases the improvement that locations see from the implementation of testing, relative to their pre-testing baseline. The main effect of exceptions (i.e., the average exception rate after testing) is absorbed by the location fixed effects δ_l.

This specification allows us to use the pre-testing period to control for location-specific factors that might drive correlations between exception rates and outcomes. It also expands the sample on which we estimate location-specific time trends. This allows us to use exception rates that are aggregated to the manager- or location-level, avoiding pool-to-pool variation.[30]

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

[30] It also helps us rule out any measurement error generated by the matching of applicants to HR managers (see Appendix A for details). 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.

To summarize, we test Proposition 4.3 with two approaches. First, we estimate the correlation between pool-level exception rates and quality of hires across applicant pools. Second, we estimate the differential impact of testing across pools with different exception rates of hires, where exception rates can be defined at the application pool, manager-, or location-level. In Section 5.4, we describe additional robustness checks.

5.3 Results

To gain a sense of the correlation between exception rates and outcome of hires, we first summarize the raw data by plotting both variables at the location level. Figure 4 shows a binned scatter plot of average tenure (log of completed duration) post-testing on the y-axis, and average location-level exception rates on the x-axis, by ventiles of exception rate. We see a strong negative relationship: locations that make more exceptions post testing have lower average tenure. Figure 5 demonstrates the same pattern for a variety of other duration related outcomes: duration measured in days, and indicators for whether a worker stays at least 3, 6 and 12 months.

The first two columns of Table 3 present the correlation between exception rates and worker tenure. We use a standardized exception rate with mean 0 and standard deviation 1 and in this panel exception rates are defined at the pool level (based on the set of applicants and hires a manager makes at a particular location in a given month).

Column 1 contains our base specification and indicates that a one standard deviation increase in the exception rate of a pool is associated with a 5% reduction in completed tenure for that group, significant at the 5% level. The coefficient is still sizeable in Column 2 which contains our full set of controls, though it does fall slightly in magnitude and become insignificant.

The remaining columns of Table 3 examine how the impact of testing varies by the extent to which managers make exceptions. Our main explanatory variable is the interaction between the introduction of testing and a post-testing exception rate.

In Columns 3 and 4, we continue to use pool-level exception rates. The coefficient on the main effect of testing represents the impact of testing at the mean exception rate (since the exception rate has been standardized). Including the full set of controls (Column 4),

we find that locations with the mean exception rate experience a 0.22 log point increase in duration as a result of the implementation of testing, but that this effect is offset by a quarter (0.05) for each standard deviation increase in the exception rate, significant at the 1% level.

In Columns 5-8, we aggregate exception rates to the manager- and location-level.[31] Results are quite consistent using these aggregations, and the differential effects are even larger in magnitude. Managers and locations that tend to exercise discretion benefit much less from the introduction of testing. A one standard deviation increase in the exception rate reduces the impact of testing by roughly half to two-thirds.[32]

Figure 4 better illustrates the variation underlying these results. Here we show binned scatter plots with the average exception rate in each of 20 bins on the x-axis, and the bin-specific impact of testing on the y-axis. of the estimated impact of testing we create another set of binned scatter plots, describing how the impact of testing varies with exception rates. Observations in the graph are weighted by the inverse variance of the estimated effect. The relationship is negative, and does not look to be driven by any particular location. Similarly, Figure 5 plots the same figure for alternative measures of duration that do not suffer from as much right censoring.

We therefore find that job durations of hires is lower for applicant pools, managers, and locations with higher exception rates. It is worth emphasizing that with the controls for the size and quality of the applicant pool, our identification comes from comparing outcomes of hires across managers who make different numbers of exceptions when facing similar applicant pools. Given this, differences in exception rates should be driven by a manager's own weighting of his or her private preferences and private information. If managers were making these decisions optimally from the firm's perspective, we should not expect to see (as we do in Table 3) that the workers they hire perform systematically worse. Based on Proposition 4.3, we can infer then that exceptions are largely driven by managerial bias,

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

[32] As we noted above, it is not possible to use these aggregated exceptions rates when examining the post-testing correlation between exceptions and outcomes (as in columns 1 and 2) because they leave little or no variation within locations to also identify location fixed effects, which, as we have argued, are quite important.

rather than private information, and these firms could improve outcomes of hires by limiting discretion.

5.4 Additional Robustness Checks

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

5.4.1 Quality of "Passed Over" Workers

There are several scenarios under which we might find a negative correlation between worker outcomes and exception rates at the pool, manager, or location level, even though managers are unbiased and using their private information optimally. For example, as mentioned above, managers may make more exceptions when green applicants in an applicant pool are idiosyncratically weak. If yellow workers in these pools are weaker than green workers in our sample on average, it will appear that more exceptions are correlated with worse outcomes even though managers are making individual exceptions to maximize worker quality. Similarly, our results in Table 3 show that locations with more exceptions see fewer benefits from the introduction of testing. An alternative explanation for this finding is that high exception locations are ones in which managers have always had better information about applicants: these locations see fewer benefits from testing because they simply do not need the test.

In these and other similar scenarios, it should still be the case that individual exceptions are correct: a yellow hired as an exception should perform better than a green who is not hired. To examine this, we would like to be able to observe the counterfactual performance of all workers who are not hired.

While we cannot observe the performance all non-hired greens, we can proxy for this comparison by exploiting the timing of hires. Specifically, we compare the performance of yellow workers hired as exceptions to green workers from the same applicant pool who are not hired that month, but who subsequently begin working in a later month. If it is the case that managers are making exceptions to increase the worker quality, then the exception yellows should have longer completed tenures than the "passed over" greens.

Table 4 shows that is not the case. The first panel compares individual durations by restricting our sample to workers who are either exception yellows, or greens who are initially passed over but then subsequently hired, and including an indicator for being in the latter group. Because these workers are hired at different times, all regressions control for hire year-month fixed effects to account for mechanical differences in duration. For column 3, 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.[33] The second panel of Table 4 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. In column 3, the best comparison, we find that passed over greens stay about 8% longer than the yellows hired before them in the same pool (top panel Column 3) and greens and yellows stay almost 19% and 12% longer, respectively, compared to the reds they were passed over for. The results in Table 4 mean that it is unlikely that exceptions are driven by better information. When workers with better test scores are at first passed over and then later hired, they still outperform the workers chosen first.

An alternative explanation is that the applicants with higher test scores were not initially passed up, but were instead initially unavailable because of better outside options. However, this seems unlikely for two reasons. First, client firms have a strong desire to fill slots and training classes quickly, so would likely strongly discourage delays. Second, it is at odds with patterns in the data. To see this, Table 5 compares job durations for workers hired immediately (the omitted category) to those who waited one, two, or three months before starting, holding constant test score. Because these workers are hired at different times, all regressions again control for hire year-month fixed effects. Across all specifications, we find no significant differences between these groups. We thus feel more comfortable interpreting the workers with longer delays as having been initially passed over by the manager, rather than initially unavailable because of a better outside option.

[33]Recall that an applicant pool is defined by a manager-location-date. Applicant pool fixed effects thus subsume a number of controls from our full specification from Table 3

Table 5 also provides insights about how much information managers have, beyond the job test. If managers have useful private information about workers, then we would expect them to be able to distinguish quality within test-color categories: greens hired first should be better than greens who are passed up. Table 5 shows that this does not appear to be the case. We estimate only small and insignificant differences in tenure, within color, across start dates. That is, within color, workers who appear to be a manager's first choice do not perform better than workers who appear to be a manager's last choice. This again suggests the value of managerial private information is small, relative to the test.

5.4.2 Extreme Outcomes

We have thus far assumed that the firm would like managers to maximize the average quality of hired workers. Firms may instead instruct managers to take a chance on candidates with poor test scores to avoid missing out on an exceptional hire. This could explain why managers hire workers with lower test score candidates to the detriment of average quality; managers may be using discretion to minimize false negatives.[34] Alternatively, firms may want managers to use discretion to minimize the chance of hiring a worker who leaves immediately.

The first piece of evidence that managers do not effectively maximize upside or minimize downside is the fact, already shown, that workers hired as exceptions perform worse than the very workers they were originally passed over. To address this more directly, Appendix Table A2 repeats our analysis focusing on performance in the tails, using the 90th and 10th percentiles of log completed durations for a cohort as the dependent variables. These results show that more exceptions imply worse performance even among top hires, suggesting managers who make many exceptions are also unsuccessful at finding star workers. We also show that more exceptions decrease the performance of the 10th percentile of completed durations as well.

[34]For example, Lazear (1998) points out that firms may be willing to pay a premium for risky workers.

5.4.3 Heterogeneity across Locations

Another possible concern is that the usefulness of the test varies across locations and that this drives the negative correlation between exception rates and worker outcomes. Our results on individual exceptions already suggest that this is not the case. However, we explore a couple of specific stories here.

In very undesirable locations, green applicants might have better outside options and be more difficult to retain. In these locations, a manager attempting to avoid costly retraining may optimally decide to make exceptions in order to hire workers with lower outside options. Here, a negative correlation between exceptions and performance would not necessarily imply that firms could improve productivity by relying more on testing. However, we see no evidence that the "return" to test score varies across locations. For example, when we split locations by pre-testing worker durations (Appendix Table A3) or by exception rates post-testing (Appendix Table A4) we see no systematic differences in the correlation between test score and job duration of hired workers.

Finally, one might worry that variation in the immediate need to fill slots drives exception rates, if the expected yield rate is higher for reds and yellows. First, fixed differences across locations in either the need to fill slots or the ability to attract greens is incorporated into our controls. Second, we have shown above that, within color, workers who start immediately are similar to those who start later, suggesting that there is no speed-quality trade-off within color. Third, Appendix Table A5 shows that our results hold in a sample of applicant pools where greens are plentiful (at least as many greens as eventual hires), and therefore likely more typical for the location.

5.4.4 Productivity

Our results show that firms can improve the tenures of their workers by relying more on job test recommendations. Firms may not want to pursue this strategy, however, if their HR managers exercise discretion in order to improve worker quality on other metrics. For example, managers may optimally choose to hire workers who are more likely to turn over

if their private signals indicate that those workers might be more productive while they are employed.

Our final set of results provides evidence that this is unlikely to be the case. Specifically, for a subset of 62,494 workers (one-quarter of all hires) in 6 client firms, we observe a direct measure of worker productivity: output per hour.[35] We are unable to reveal the exact nature of this measure but some examples may include: the number of data items entered per hour, the number of standardized tests graded per hour, and the number of phone calls completed per hour. In all of these examples, output per hour is an important measure of efficiency and worker productivity and is fairly homogenous across client firms. Our particular measure has an average of roughly 8 with a standard deviation of roughly 5.

Table 6 repeats our main findings, using output per hour instead of job duration as the dependent variable. We focus on estimates only using our base specification (controlling for date and location fixed effects) because the smaller sample and number of clients makes identifying the other controls difficult.[36]

Column 1 examines the impact of the introduction of testing, which we find leads to a statistically insignificant increase of 0.7 transactions in an hour, or a roughly 8% increase. The standard errors are such that we can rule out virtually any negative impact of testing on productivity with 90% confidence.[37]

Column 2 documents the post-testing correlation between pool-level exceptions and output per hour, and Columns 3-5 examine how the impact of testing varies by exception rates. In all cases, we find no evidence that managerial exceptions improve output per hour. Instead, we find noisy estimates indicating that worker quality appears to be lower on this dimension as well. For example, in Column 2, we find a tiny, insignificant positive coefficient describing the relationship between exceptions and output. Taking it seriously implies that a 1 standard deviation increase in exception rates is correlated with 0.07 more transactions, or a less than 1% increase. In, Columns 3-5, we continue to find an overall positive effect of

[35] We have repeated our main analyses on the subsample of workers that have output per hour data and obtained similar results.

[36] Results are, however, qualitatively similar with additional controls, except where noted.

[37] This result is less robust to adding additional controls, however we can still rule out that testing has a substantial negative effect. For example, adding location-specific time trends, the coefficient on testing falls from 0.7 to 0.26 (with a standard error of about 0.45).

testing on output; we find no evidence of a positive correlation between exception rates and the impact of testing. If anything, the results suggest that locations with more exceptions experience slightly smaller impacts of testing. These effects are insignificant.

Taken together, the results in Table 6 provide no evidence that exceptions are positively correlated with productivity. This refutes the hypothesis that, when making exceptions, managers optimally sacrifice job tenure in favor of workers who perform better on other quality dimensions.

6 Conclusion

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

Our paper contributes a new methodology for evaluating the value of discretion in firms. Our test is intuitive, tractable, and requires only data that would readily be available for firms using workforce analytics. In our setting it provides the stark recommendation that firms would do better to remove discretion of the average HR manager and instead hire based solely on the test. Our results provide evidence that the typical manager underweights the job test relative to what the firm would prefer. Based on such evidence, firms may want to explore a range of alternative options. For example, relative to the status quo, firms may restrict the frequency with which managers can overrule the test (while still allowing a degree of discretion) or, adopt other policies to influence manager behavior such as direct pay for performance or more selective hiring and firing.

These findings highlight the role new technologies can play in reducing the impact of managerial mistakes or biases by making contractual solutions possible. 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, Philippe and Jean Tirole (1997), "Formal and Real Authority in Organizations," *The Journal of Political Economy*, 105(1).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[17] Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2015). "Human Decisions and Machine Predictions," Unpublished.

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

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

[20] Lazear, Edward P. (1998), "Hiring Risky Workers." In *Internal Labour Market, Incentives, and Employment,* edited by Isao Ohashi and Toshiaki Tachibanaki. New York: St. Martin's.

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

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

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

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

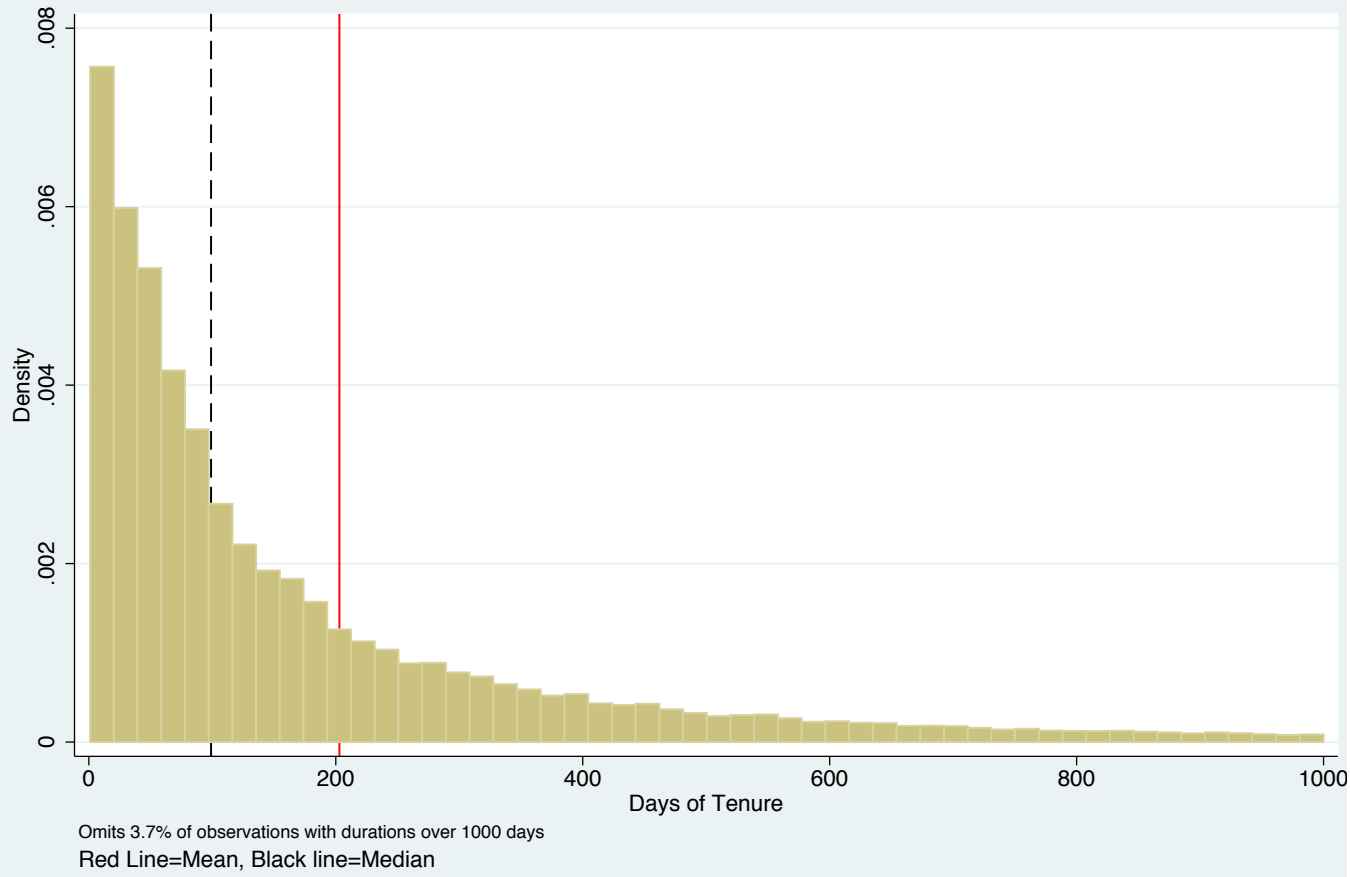
[25] Paravisini, Daniel and Antoinette Schoar (2013) "The Incentive Effect of IT: Randomized Evidence from Credit Committees" NBER Working Paper #19303.

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

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

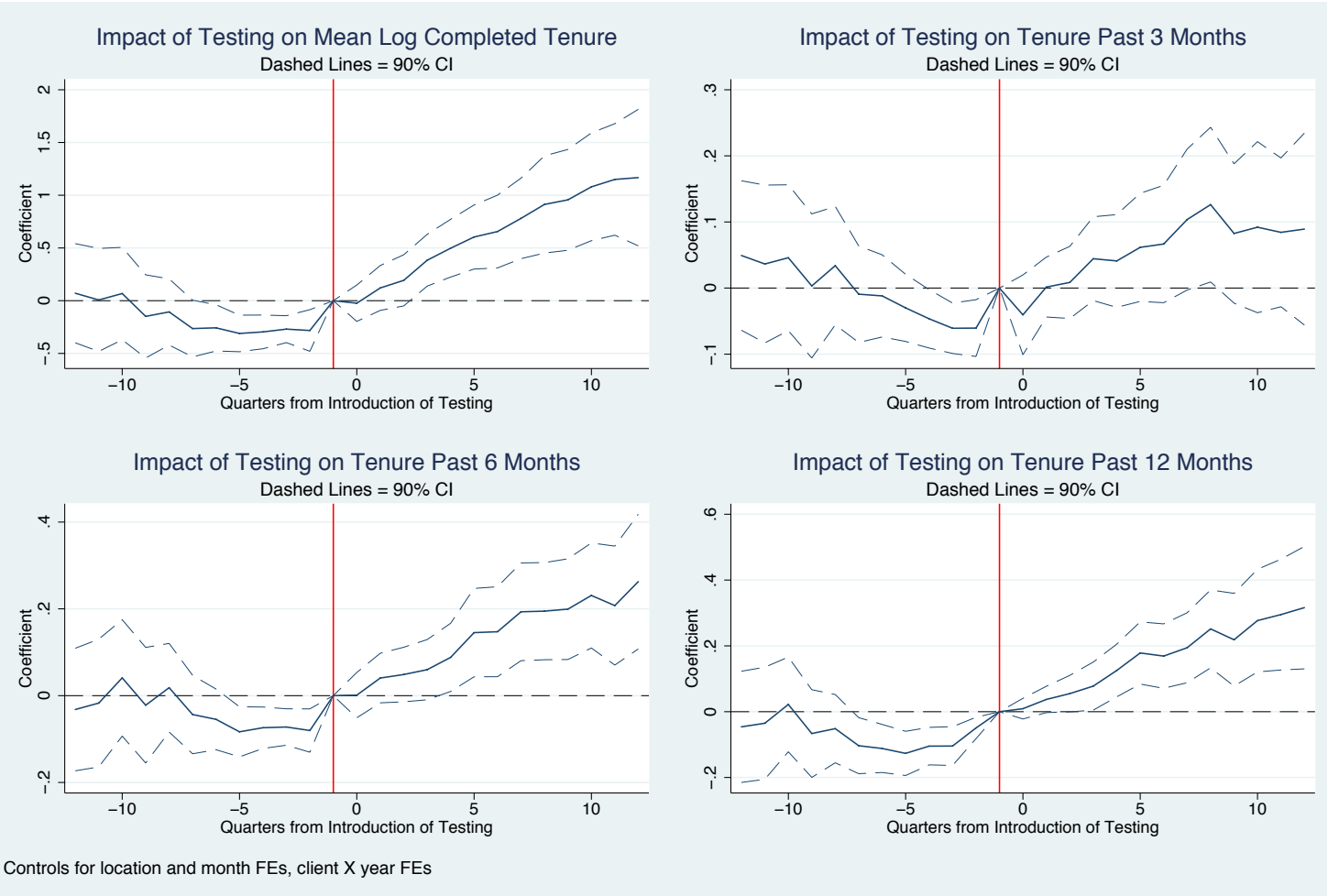
[28] Wang, James (2014), "Why Hire Loan Officers? Examining Delegated Expertise," mimeo University of Michigan.

FIGURE 1: DISTRIBUTION OF LENGTH OF COMPLETED JOB SPELLS



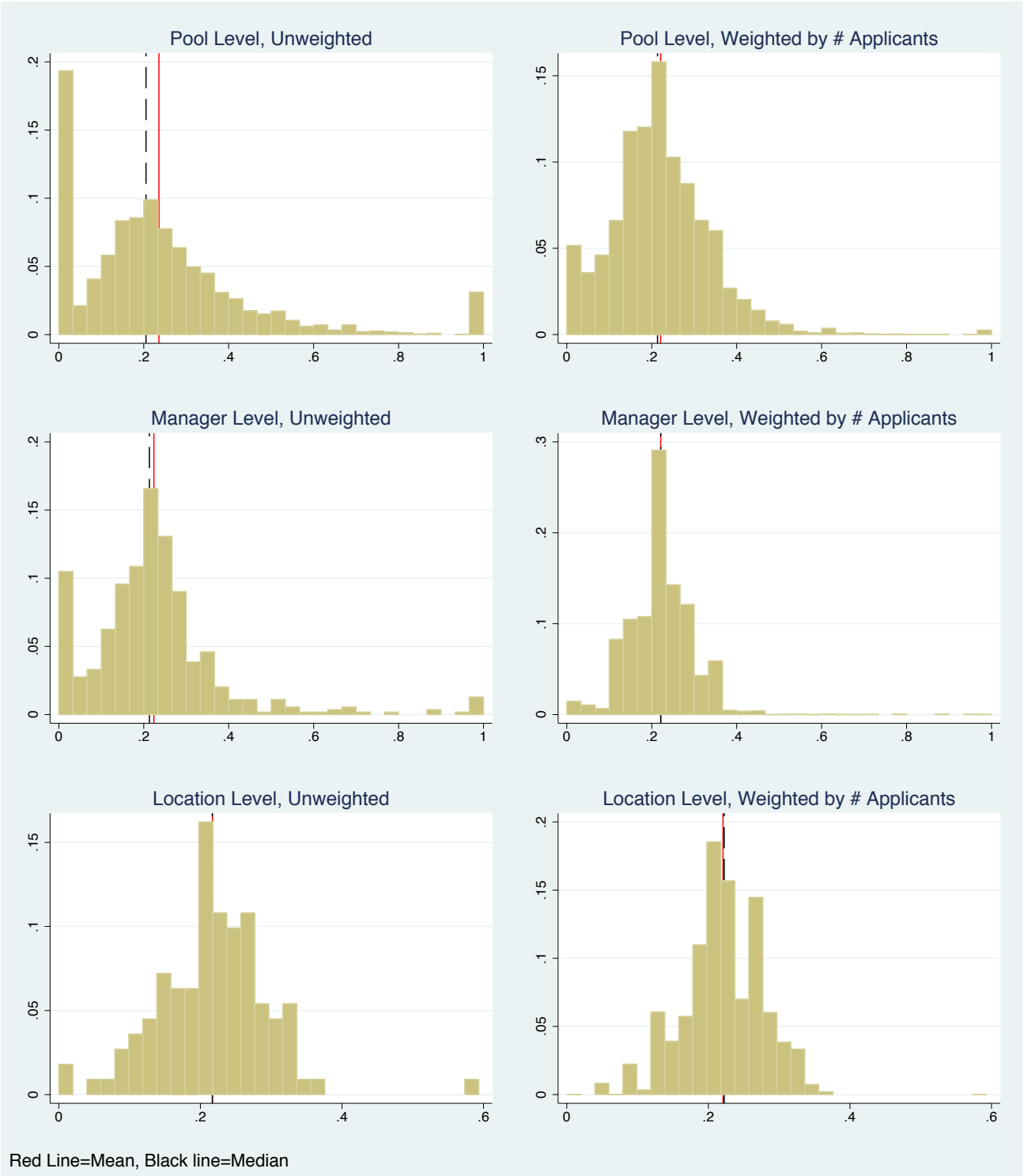
NOTES: Figure 1 plots the distribution of completed job spells at the individual level.

Figure 2: Event Study of Duration Outcomes



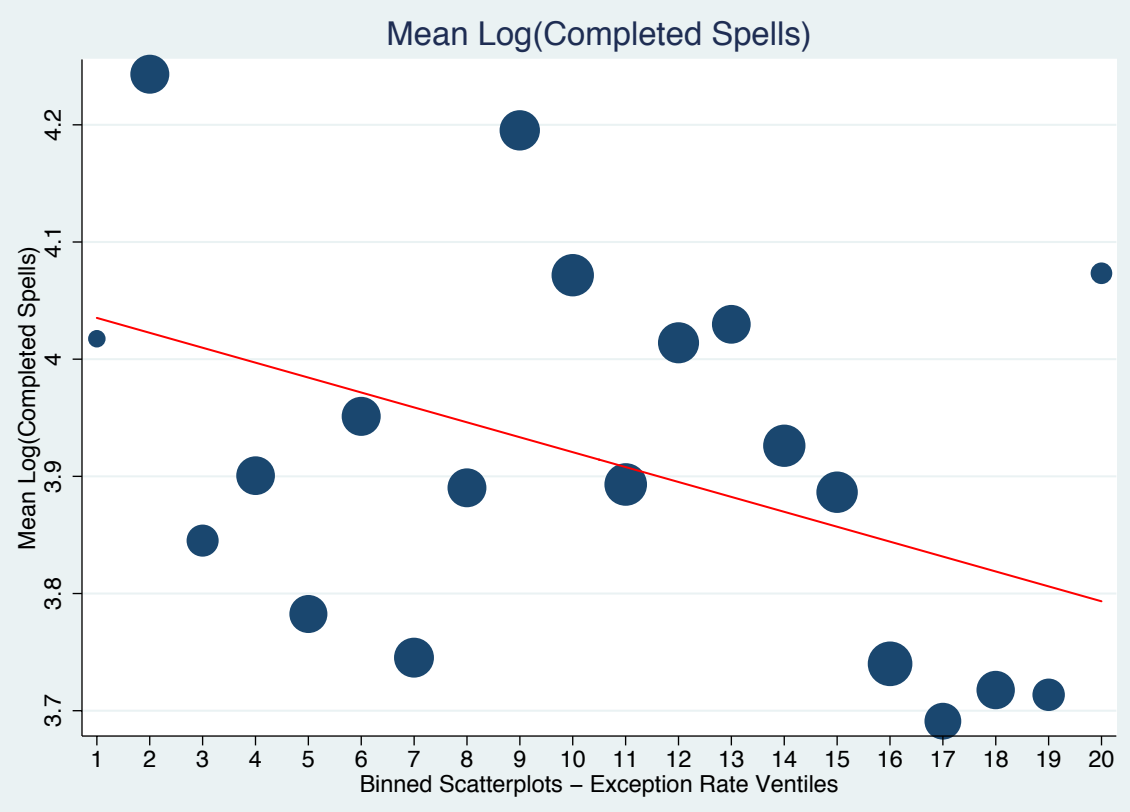
Notes: These figures plot the average duration outcome for entry cohorts by time (in quarters) until or time after testing is adopted. The underlying estimating equation is given by $\text{Log(Duration)}_{lt} = \alpha_0 + I_{lt}^{\text{time since testing}}\alpha_1 + \delta_l + \gamma_t + \epsilon_{lt}$, where $I^{\text{time since testing}}$ is a vector of dummies indicating how many quarters until or after testing is adopted, with one quarter before as the omitted category. This regression includes the base set of controls – location (δ_l) and date (γ_t) fixed effects; it does not control for location-specific time trends.

Figure 3: Distributions of Application Pool Exception Rates



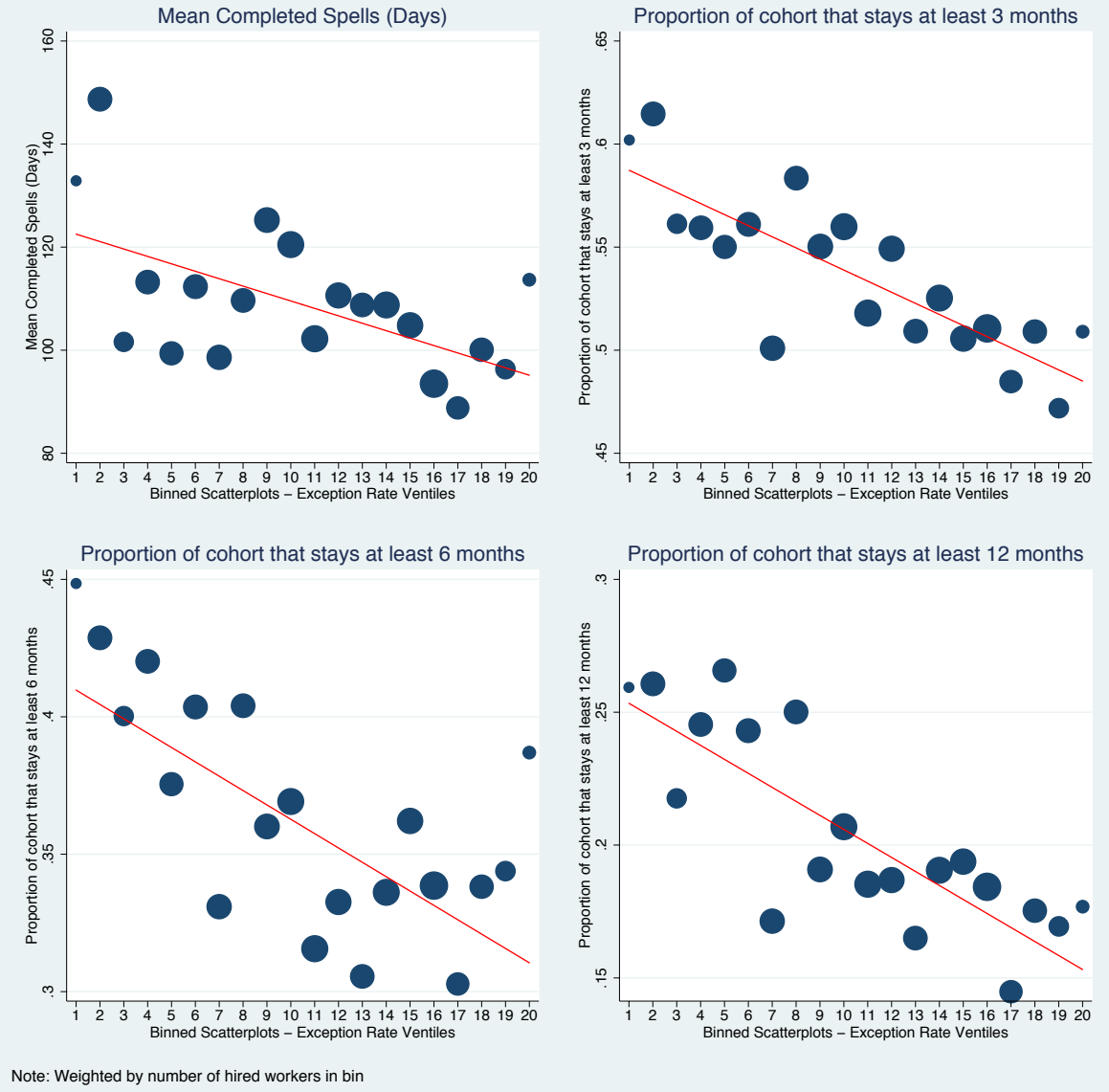
Notes: These figures plot the distribution of the exception rate, as defined by Equation (2) in Section 5. The leftmost panel presents results at the applicant pool level (defined to be a manager–location–month). The middle panel aggregates these data to the manager level and the rightmost panel aggregates further to the location level. Exception rates are only defined for the post-testing sample.

FIGURE 4: LOCATION-LEVEL EXCEPTION RATES AND POST-TESTING JOB DURATIONS



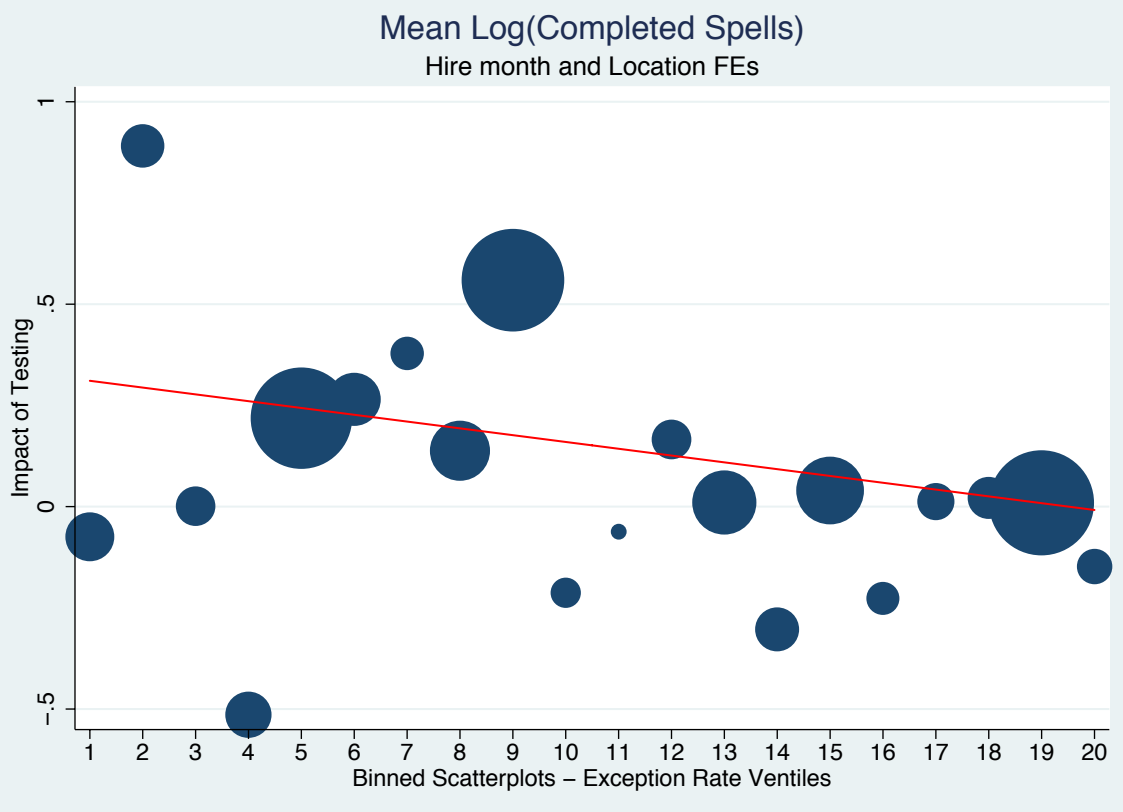
NOTES: The x-axis represents ventiles of exception rates. The y-axis is the mean log completed tenure at a given location after the introduction of testing, for locations in the specified venture. The line shows the best linear fit of the scatter plot, weighted by the number of hires in locations for each ventile.

FIGURE 5: LOCATION-LEVEL EXCEPTION RATES AND POST-TESTING DURATION MILESTONES



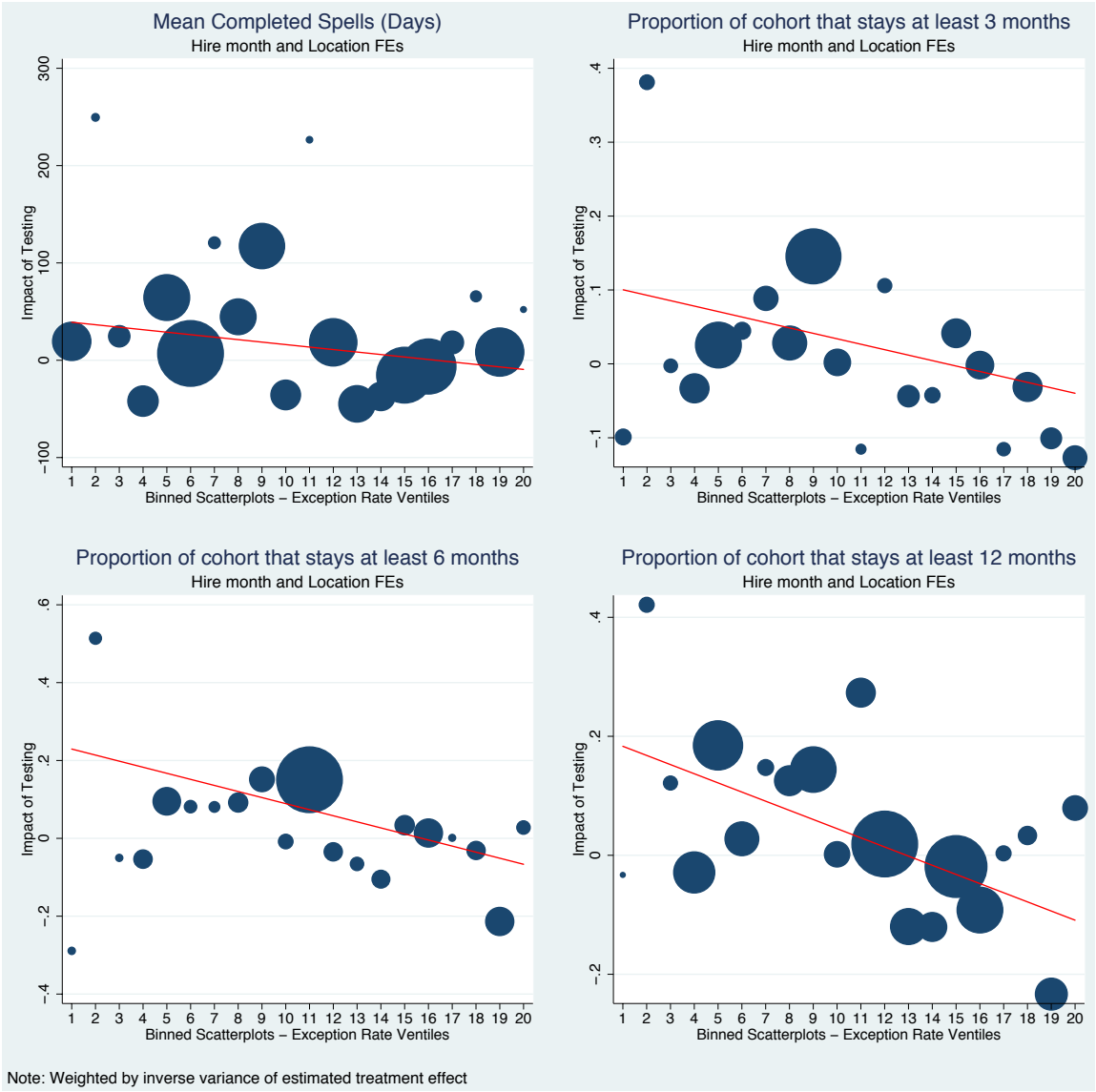
NOTES: The x-axis represents ventiles of exception rates. The y-axis is the mean log completed tenure at a given location after the introduction of testing, for locations in the specified venture. The line shows the best linear fit of the scatter plot, weighted by the number of hires in locations for each ventile.

FIGURE 6: LOCATION-LEVEL EXCEPTION RATES AND THE IMPACT OF TESTING ON JOB DURATIONS



NOTES: The x-axis represents ventiles of exception rates. The y-axis is the estimated impact of the introduction of testing, estimated separately for each ventile of exception rates. The plot is weighted by the inverse variance of the estimate associated with each exception rate bin.

FIGURE 7: LOCATION-LEVEL EXCEPTION RATES AND THE IMPACT OF TESTING ON JOB DURATION MILESTONES



NOTES: The x-axis represents ventiles of exception rates. The y-axis is the estimated impact of the introduction of testing, estimated separately for each ventile of exception rates. The plot is weighted by the inverse variance of the estimate associated with each exception rate bin.

TABLE 1: SUMMARY STATISTICS

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

	Worker Performance			*mean (st dev)*	
	Pre-testing	Post-testing	Green	Yellow	Red
Duration of Completed Spell (Days) (N=202,728)	247 *(314)*	116 *(116)*	122 *(143)*	110 *(131)*	93 *(122)*
Output per Hour (N=62,494)	8.32 *(4.58)*	8.41 *(5.06)*	8.37 *(4.91)*	8.30 *(5.04)*	9.09 *(5.94)*

	Applicant Pool Characteristics			
	Post-testing	Green	Yellow	Red
Share Applicants		0.43	0.29	0.28
Share Hired	0.19	0.22	0.18	0.09
Exception Rate	0.24			

NOTES: Post-testing is defined at the location-month level as the first month in which 50% of hires had test scores, and all months thereafter. An applicant pool is defined at the manager-location-month level and includes all applicants that had applied within four months of the current month and not yet hired. Number of applicants reflects the total number in any pool.

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

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

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

NOTES: In Panel 1, an observation is a location-month. The dependent variable is average log 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 hires in that location-month. Standard errors in parentheses are clustered at the location level. In Panel 2, observations are at the individual level. Testing is defined as whether or not an individual worker has a test score. Regressions are unweighted. Size and composition of the applicant pool controls include separate fixed effects for the number of red, yellow, and green applicants at that location-month.

TABLE 3: RELATIONSHIP BETWEEN JOB DURATIONS AND EXCEPTION RATES

	Log Duration of Completed Spells							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Post-Testing Sample		**Full Sample**					
			Level of Aggregation for Exception Rate					
			Pool		*Manager*		*Location*	
Post-Testing	--	--	0.277**	0.217**	0.281**	0.243***	0.306**	0.252***
			(0.112)	*(0.0876)*	*(0.114)*	*(0.0869)*	*(0.123)*	*(0.0880)*
*Exception Rate*Post-Testing*	-0.0491**	-0.0310	-0.101**	-0.0477***	-0.211**	-0.131***	-0.290**	-0.170**
	(0.0223)	*(0.0213)*	*(0.0446)*	*(0.0182)*	*(0.0907)*	*(0.0335)*	*(0.112)*	*(0.0744)*
N	3,839	3,839	6,869	6,869	6,942	6,942	6,956	6,956
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
Size and Composition of Applicant Pool		X		X		X		X

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

NOTES: Each observation is a manager-location-month, observations are weighted by number of hires, and standard errors are clustered by location. 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. Columns 1 and 2 restrict to the post-testing sample only. The pool-level 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. Columns 3-4 and 5-6 aggregate the pool-level exception rate to the manager- and location-level, respectively. Exception rates are only defined post testing and are set to 0 pre testing. Size and composition of the applicant pool controls include separate fixed effects for the number of red, yellow, and green applicants at that location-manager-month. See text for additional details.

Table 4: Completed Tenure of Exceptions vs. Passed Over Applicants

	Log Duration of Completed Spells		
	(1)	(2)	(3)
Panel 1: Quality of Yellow Exceptions vs. Passed over Greens			
Passed Over Greens	0.0436***	0.0545***	0.0794***
	(0.0140)	*(0.0118)*	*(0.0236)*
N	59,462	59,462	59,462
Panel 2: Quality of Red Exceptions vs. Passed over Greens and Yellows			
Passed Over Greens	0.131***	0.165***	0.213***
	(0.0267)	*(0.0251)*	*(0.0341)*
Passed Over Yellows	0.0732***	0.113***	0.154***
	(0.0265)	*(0.0238)*	*(0.0344)*
N	44,456	44,456	44,456
Hire Month FEs	X	X	X
Location FEs	X	X	X
Full Controls from Table 3		X	X
Application Pool FEs			X

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

Notes: Regressions are at the individual level on the post-testing sample. Standard errors are clustered by location. Panel 1 includes yellow exceptions (who were hired when a green applicant was available but not hired in that month) and passed over green applicants who were later hired. The omitted category are yellow exceptions. The second panel includes red exceptions (who were hired when a green or yellow applicant was available but not hired in that month) and passed over greens and yellows only. Red exceptions are the omitted category. Application pool fixed effects are defined for a given location-manager-month.

TABLE 5: JOB DURATION OF WORKERS, BY LENGTH OF TIME IN APPLICANT POOL

	Log Duration of Completed Spells		
	(1)	(2)	(3)
Green Workers		**Green Workers**	
Waited 1 Month	-0.00908	0.00676	0.00627
	(0.0262)	*(0.0187)*	*(0.0204)*
Waited 2 Months	-0.0822	-0.0502	-0.0446
	(0.0630)	*(0.0365)*	*(0.0385)*
Waited 3 Months	-0.000460	-0.0342	-0.0402
	(0.0652)	*(0.0581)*	*(0.0639)*
N	41,020	41,020	41,020
		Yellow Workers	
Waited 1 Month	-0.00412	0.0144	0.00773
	(0.0199)	*(0.0198)*	*(0.0243)*
Waited 2 Months	-0.0100	-0.0355	-0.0474
	(0.0448)	*(0.0406)*	*(0.0509)*
Waited 3 Months	0.103	0.0881	0.114
	(0.0767)	*(0.0882)*	*(0.0979)*
N	22,077	22,077	22,077
		Red Workers	
Waited 1 Month	0.0712	0.0587	0.0531
	(0.0520)	*(0.0621)*	*(0.0617)*
Waited 2 Months	0.0501	0.0789	0.0769
	(0.0944)	*(0.128)*	*(0.145)*
Waited 3 Months	0.103	0.225	0.149
	(0.121)	*(0.141)*	*(0.168)*
N	4,919	4,919	4,919
Year-Month FEs	X	X	X
Location FEs	X	X	X
Full Controls from Table 3		X	X
Application Pool FEs			X

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

NOTES: Each observation is an individual hired worker for the post-testing sample. 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. Standard errors are clustered by location. Application pool fixed effects are defined for a given location-manager-month.

Table 6: Testing, Exception Rates, and Output Per Hour

	Impact of testing	**Exceptions and outcomes, post testing**	**Impact of testing, by exception rate**		
			Level of aggregation for exception rate		
			Pool	*Manager*	*Location*
	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Output per hour				
Post-Testing	0.704 *(0.430)*		0.703 *(0.432)*	0.721* *(0.429)*	0.711 *(0.435)*
*Exception Rate*Post-Testing*		0.0750 *(0.0851)*	0.0335 *(0.0838)*	-0.112 *(0.250)*	-0.0424 *(0.123)*
N	2,699	1,527	2,665	2,699	2,690
Year-Month FEs	X	X	X	X	X
Location FEs	X	X	X	X	X

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

Notes: This table replicates the baseline specifications in Tables 2 and 3, using the number of transactions per hour (mean 8.38, std. dev. 3.21) as the dependent variable. All regressions are weighted by number of hires and standard errors are clustered by location. Column 1 studies the impact of the introduction of testing. Column 2 examines the correlation between output per hour and pool-level exception rates in the post-testing sample only. Columns 3 through 5 study the differential impact of the introduction of testing, with the exception rate aggregated at the pool, manager, and location levels, respectively.

FOR ONLINE PUBLICATION ONLY

A Data Appendix

Firms in the Data. The data were assembled for us by the data firm from records of the individual client firms. The client firms in our sample employ workers who are engaged in the same job, but there are some differences across the firms in various dimensions. For example, at one firm, workers engage in a relatively high-skilled version of the job we study. At a second firm, the data firm provides assistance with recruiting (beyond providing the job test). Our baseline estimate of the relationship between exceptions and duration is similar when individual firms are excluded one by one.[38] Among the non-tested workers, the data include a small share of workers outside of entry-level positions, but we checked our results are robust when we repeat our analyses controlling for employee position type.

Pre-testing Data. In the pre-testing data, at some client firms, there is information not only on new hires, but also on incumbent workers. This may generate a survivor bias for incumbent workers, relative to new workers. For example, consider a firm that provided pre-testing data on new hires going back to Jan. 2010. For this firm, we would observe the full set of workers hired at each date after Jan. 2010, but for those hired before, we would only observe the subset who survived to a later date. We do not explicitly observe the date at which the firm began providing information on new hires; instead, we conservatively proxy this date using the date of first recorded termination. We label all workers hired before this date as "stock sampled" because we cannot be sure that we observe their full entry cohort. We drop these workers from our primary sample, but have experimented with including them along with flexible controls for being stock sampled in our regressions.[39]

Productivity. In addition to hire and termination dates, with which we calculate our primary outcome measure, some client firms provide data on output per hour. This is available for about a quarter of hired workers in our sample, and is mentioned by our data firm in its advertising, alongside duration. We trim instances where average transaction time in a given day is less than 1 minute.[40]

[38] Specifically, we estimated Column 1 of Table 3 excluding each firm one by one.

[39] In addition to the issue of stock-sampling, the number of months of pre-testing data varies across firms. However, as noted above, our baseline estimate of the relationship between exceptions and duration is robust to excluding individual firms from our sample.

[40] This is about one percent of transactions. Our results are stronger if we do not trim. Some other productivity variables are also shared with our data provider, but each variable is only available for an even smaller share of workers than is output per hour. Such variables would likely face significant statistical power issues if subjected to the analyses in the paper (which involve clustering standard errors at the location level).

Test Scores. As described in the text, applicants are scored as Red, Yellow, or Green. Applicants may receive multiple scores (e.g., if they are being considered for multiple roles). In these cases, we assign applicants to the maximum of their scores.[41]

HR Manager. We do not have data on the characteristics of HR managers (we only see an individual identifier). When applicants interact with more than one HR manager during the recruitment process, they are assigned to the manager with whom they have the most interactions.[42]

In the data provided, information on HR manager is missing for about one-third of the job-tested applicants. To form our base sample of workers in Table 1, we drop job-tested workers where information on HR manager is missing. In forming our sample of applicants and in creating the applicant pool variables, individuals with missing information on HR manager are assigned to a separate category. We have also repeated our main analyses while excluding all job-tested individuals (applicants *and* workers) where HR manager is missing and obtained qualitatively similar results.

Race, Gender, Age. Data on race, sex, and age are not available for this project. However, Autor and Scarborough (2008) show that job testing does not seem to affect worker race, suggesting that changes in worker demographics such as race are not the mechanism by which job testing improves durations.[43]

Location Identifiers. In our dataset, we do not have a common identifier for workplace location for workers hired in the pre-testing period and applicants applying post-testing. Consequently, we develop a crosswalk between anonymized location names (used for workers in the pre-testing period) and the location IDs in the post-testing period. We drop workers from our sample where the merge did not yield a clean location variable.[44]

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

[41] For 1 of the 15 firms, the Red/Yellow/Green score is missing for non-hired applicants in the dataset provided for this project. Our conclusions are substantively unchanged if that firm is removed from the data.

[42] This excludes interactions where information on the HR manager is missing. If there is a tie for most interactions, an applicant is assigned the last manager listed in the data among those tied for most interactions with that applicant. Our main results are also qualitatively robust to setting the HR manager identifier to missing in cases of ties for most interactions.

[43] Autor and Scarborough (2008) do not look at impact of testing on gender. However, they show there is little differential impact of testing by gender (or race).

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

managers to overrule testing recommendations. Information from this survey is referenced in footnotes 12 and 28 in the main text.

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

B Proofs

B.1 Preliminaries

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

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

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

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

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

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

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

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

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

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

B.2 Proof of Proposition 3.1

The following results formalize conditions under which the firm will prefer Discretion or No Discretion.

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

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

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

For this proof we make use of the following lemma:

Lemma B.1 *The expected quality of hires for a given manager, $E[a|Hire]$, is decreasing in managerial bias, k.*

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

That is, some candidates will be hired by both managers, but for $E[a|s_i,t_i] > \hat{a}$, Manager 2 (less bias) will hire some candidates that Manager 1 would not, and for $E[a|s_i,t_i] < \hat{a}$ Manager 1 (more bias) will hire some candidates that Manager 2 would not. The candidates that Manager 2 would hire when Manager 1 would not, have high expected values of a, while the candidates that Manager 1 would hire where Manager 2 would not have low expected values of a. Therefore the average a value for workers hired by Manager 2, the less biased manager, must be higher than that for those hired by Manager 1. $E[a|Hire]$ is decreasing in k.

We next prove each item of Proposition 3.1

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

Proof When $k = 1$, the manager hires based only on b, which is independent of a. So $E[a|Hire] = p_G\mu_G + (1 - p_G)\mu_Y$. The firm would do better under No Discretion (where quality of hires equals μ_G). When $k = 0$, the manager hires only applicants whose expected quality, a, is above the threshold. In this case, the firm will at least weakly prefer Discretion. Since the manager's preferences are perfectly aligned, he or she will always do at least as well as hiring only type G.

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

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

Proof When $1/\sigma_\epsilon^2 = 0$, i.e., the manager has no information, and $k = 0$, he or she will hire based on the test, resulting in an equal payoff to the firm as No Discretion. For all $k > 0$, the payoff to the firm will be worse than No Discretion, thanks to lemma B.1. Thus when the manager has no information the firm prefers No Discretion to Discretion.

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

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

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

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

Define $\Delta(\sigma_\epsilon^2, k)$ as the difference in quality of hires under Discretion, compared to no Discretion, for fixed manager type (σ_ϵ^2, k). We know that $\Delta(\sigma_\epsilon^2, 0)$ is positive and decreasing in σ_ϵ^2, and approaches 0 as σ_ϵ^2 approaches ∞. Also, since the firm's payoff under discretion is continuous in both k and $1/\sigma_\epsilon^2$ (see equation 6 above), $\Delta()$ must also be continuous in these variables.

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

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

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

B.3 Proof of Proposition 3.2

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

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

Above, we defined U_t, a manager's utility of a candidate conditional on t, and showed that it is normally distributed with mean $(1-k)\mu_t$ and variance $\Sigma = (1-k)^2\sigma + k^2\sigma_b^2$. A manager will hire all applicants for whom U_t is above $\underline{u}$ where the latter is chosen to keep the hire rate fixed at W.

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

Let us now define $\tilde{U}_t = \frac{U_t}{\sqrt{\Sigma}}$. This transformation is still normally distributed but now has mean $\frac{(1-k)\mu_t}{\sqrt{\Sigma}}$ and variance 1. This rescaling of course does nothing to the cutoff $\underline{u}$, and it will still be the case that the probability of an exception is decreasing in the difference in expected utilities across $\tilde{U_G}$ and $\tilde{U_Y}$: $\Delta_U = \frac{(1-k)(\mu_G - \mu Y)}{\sqrt{\Sigma}}$.

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

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

B.4 Proof of Proposition 3.3

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

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

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

	Mean Completed Duration (Days, Mean=211; SD=232)		>3 Months (Mean=0.62; SD=0.21)		>6 Months (Mean=0.46; SD=0.24)		>12 Months (Mean=0.32; SD=0.32)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post-Testing	88.89**	29.66*	0.0404***	0.0895**	0.0906***	0.107**	0.107***	0.0652*
	(35.91)	*(15.40)*	*(0.00818)*	*(0.0379)*	*(0.00912)*	*(0.0463)*	*(0.00976)*	*(0.0363)*
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
Size and Composition of Applicant Pool		X		X		X		X

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

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

Appendix Table A2: Testing, Exceptions, and Tail Outcomes

	Impact of testing	**Exceptions and outcomes, post testing**	**Impact of testing, by exception rate**		
			Level of aggregation for exception rate		
			Pool	*Manager*	*Location*
	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: 90th Percentile of Log Duration of Completed Spells				
Post-Testing	0.126*		0.150**	0.141**	0.131*
	(0.069)		*(0.0703)*	*(0.0697)*	*(0.0716)*
*Exception Rate*Post-Testing*		-0.0426*	-0.0461**	-0.0764***	-0.0311
		(0.0242)	*(0.0215)*	*(0.0270)*	*(0.0563)*
	Dependent Variable: 10th Percentile of Log Duration of Completed Spells				
Post-Testing	0.239*		0.186	0.260*	0.314**
	(0.144)		*(0.145)*	*(0.144)*	*(0.141)*
*Exception Rate*Post-Testing*		-0.0322	-0.0540*	-0.204***	-0.456***
		(0.0336)	*(0.0322)*	*(0.0658)*	*(0.170)*
N	6,956	3,839	6,869	6,942	6,956
Year-Month FEs	X	X	X	X	X
Location FEs	X	X	X	X	X
Client Firm X Year FEs	X	X	X	X	X
Location Time Trends	X	X	X	X	X
Size and Composition of Applicant Pool	X	X	X	X	X

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

Notes: See notes to Tables 2 and 3.

Appendix Table A3: Impact of Color Score on Job Duration by Pre-testing Location Duration

	Log Duration of Completed Spells					
	High Duration	Low Duration	High Duration	Low Duration	High Duration	Low Duration
	(1)	(2)	(3)	(4)	(5)	(6)
Green	0.165***	0.162***	0.175***	0.165***	0.175***	0.170***
	(0.0417)	*(0.0525)*	*(0.0527)*	*(0.0515)*	*(0.0535)*	*(0.0528)*
Yellow	0.0930**	0.119**	0.105**	0.107**	0.108**	0.112**
	(0.0411)	*(0.0463)*	*(0.0505)*	*(0.0483)*	*(0.0513)*	*(0.0495)*
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
Full Controls from Table 3			X	X	X	X
Application Pool FEs					X	X

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

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

Appendix Table A4: Impact of Color Score on Job Duration by Location-Specific Exception Rates

	Log Duration of Completed Spells					
	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***	0.182***	0.151**	0.181***	0.171***
	(0.0317)	*(0.0689)*	*(0.0312)*	*(0.0628)*	*(0.0331)*	*(0.0642)*
Yellow	0.112***	0.182**	0.116***	0.109	0.117***	0.128*
	(0.0287)	*(0.0737)*	*(0.0279)*	*(0.0696)*	*(0.0296)*	*(0.0711)*
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
Full Controls from Table 3			X	X	X	X
Application Pool FEs					X	X

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

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

Appendix Table A5: Exception Rates and Duration Outcomes
Applicant Pools with # Green Applicants > # Hires

	Log Duration of Completed Spells							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Post-Testing Sample		**Full Sample**					
			Level of Aggregation for Exception Rate					
			Pool		*Manager*		*Location*	
Post-Testing	--	--	0.286**	0.250***	0.284**	0.264***	0.298**	0.286***
			(0.116)	*(0.0883)*	*(0.115)*	*(0.0888)*	*(0.120)*	*(0.0901)*
*Exception Rate*Post-Testing*	-0.0520*	-0.0359	-0.117*	-0.0313	-0.267**	-0.140***	-0.334**	-0.195***
	(0.0272)	*(0.0261)*	*(0.0594)*	*(0.0252)*	*(0.124)*	*(0.0441)*	*(0.134)*	*(0.0684)*
N	3,343	3,343	6,373	6,373	6,409	6,409	6,417	6,417
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
Size and Composition of Applicant Pool		X		X		X		X

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