DISCRETION IN HIRING*

MITCHELL HOFFMAN
LISA B. KAHN
DANIELLE LI

Job-testing technologies enable firms to rely less on human judgment when making hiring decisions. Placing more weight on test scores may improve hiring decisions by reducing the influence of human bias or mistakes but may also lead firms to forgo the potentially valuable private information of their managers. We study the introduction of job testing across 15 firms employing low-skilled service sector workers. When faced with similar applicant pools, we find that managers who appear to hire against test recommendations end up with worse average hires. This suggests that managers often overrule test recommendations because they are biased or mistaken, not only because they have superior private information. *JEL Codes:* M51, J24.

I. INTRODUCTION

Hiring the right workers is one of the most important and difficult problems that a firm faces. Résumés, 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. Furthermore, the managers that firms employ to gather and interpret this information may have poor judgment or preferences that are imperfectly aligned with firm objectives.[1] Firms may thus face both information and agency problems when making hiring decisions.

*We are grateful to Jason Abaluck, Ajay Agrawal, Ricardo Alonso, Pol Antràs, Ian Ball, David Berger, Arthur Campbell, David Deming, Alex Frankel, Avi Goldfarb, Lawrence Katz, Harry Krashinsky, Peter Landry, Jin Li, Liz Lyons, Steve Malliaris, Mike Powell, Kathryn Shaw, Steve Tadelis, numerous seminar participants, and anonymous referees for helpful comments. We are grateful to the anonymous data provider for providing access to proprietary data. Hoffman acknowledges financial support from the Social Science and Humanities Research Council of Canada. All errors are our own.

1. For example, a manager could have preferences over demographics or family background that do not maximize productivity. In a case study of elite professional services firms, Rivera (2012) provides evidence that an important determinant of hiring is the presence of shared leisure activities.

The Quarterly Journal of Economics (2018), 1–36. doi:10.1093/qje/qjx042.
Advance Access publication on October 10, 2017.

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

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

In this article, we evaluate the introduction of a job test and analyze the consequences of making hiring decisions that deviate from test score recommendations. We use a unique personnel data set consisting of 15 firms that employ workers in the same low-skilled service sector. Prior to the introduction of testing, firms employed HR managers who were involved in hiring new workers. After the introduction of testing, HR managers were also given access to a test score for each applicant: green (high-potential candidate), yellow (moderate-potential candidate), or red (lowest rating). Managers were encouraged to factor the test into their hiring decisions, but were not required to hire strictly according to test recommendations.

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

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

in paid training. This finding suggests that this job test contains useful information about the quality of candidates.

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

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

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

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

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

As data analytics is more frequently applied to HR management decisions, it becomes increasingly important to understand how these new technologies impact the organizational structure of the firm and the efficiency of worker–firm matching. While a large theoretical literature has studied how firms should allocate authority, and a smaller empirical literature has examined discretion and rule making in other settings, empirical evidence on discretion in hiring is scant.[3] Our article provides a first step toward an empirical understanding of the potential benefits of discretion in hiring. Our findings provide evidence that screening

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

technologies may improve information symmetry between firms and managers. In this spirit, our article is related to the classic Baker and Hubbard (2004) analysis of the adoption of onboard computers in the trucking industry.

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

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

II. Setting and Data

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

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

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

II.A. Job Test and Testing Adoption

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

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

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

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

The HR managers in our data are referred to as recruiters by our data provider and are unlikely to manage day-to-day production. Prior to the introduction of job testing, our client firms gave their HR managers discretion to make hiring recommendations

based on interviews and résumés.[6] After adopting this job test, firms made applicant test scores available to managers and encouraged them to factor scores into hiring recommendations, but managers were still permitted to hire their preferred candidate.[7]

II.B. Applicant and Worker Data

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

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

Our primary worker outcome is job duration. We focus on turnover for three main reasons. Foremost, turnover is a perennial challenge for firms employing low-skilled service sector workers. Hence, tenure is an important measure of worker quality for our sample firms. To illustrate this concern, Figure I shows a histogram of job tenure for completed spells (79% of the spells in our data) among employees in our sample. The median worker (solid line) stays only 99 days, or just over 3 months. One in six workers leave after only a month. Despite these short tenures, hired workers in our sample spend the first several weeks of their employment in paid training.[9] Our data firm and its client firms are

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

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

8. See Section A.7 of the Online Appendix regarding sample restrictions.

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

TABLE I

SUMMARY STATISTICS

			All	Pre-testing	Post-testing
Sample coverage					
# Locations			127	113	97
# Hired workers			265,648	174,329	91,319
# Applicants					403,006
# HR managers					445
# Pools					3,698
# Applicants/pool					260
	Pre-testing	Post-testing	Green	Yellow	Red
Worker characteristics mean (std. dev.)					
Duration of completed spell (days) ($N = 209{,}808$)	252 (323)	116 (138)	122 (143)	110 (130)	92 (121)
Duration of censored spell (days) ($N = 55{,}840$)	807 (510)	252 (245)	265 (252)	235 (232)	223 (223)
Share censored	0.19 (0.39)	0.25 (0.43)	0.24 (0.43)	0.26 (0.44)	0.25 (0.43)
Output per hour ($N = 62{,}427$)	8.35 (4.66)	8.44 (5.16)	8.39 (5.01)	8.32 (5.11)	9.16 (6.08)
		Post-testing	Green	Yellow	Red
Applicant pool characteristics					
Share applicants			0.48	0.32	0.20
Hire probability		0.19	0.23	0.18	0.08

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

aware of these concerns: in its marketing materials, our data firm emphasizes the ability of its job test to reduce turnover. Second, in addition to its importance for our sample firms, in many canonical models of job search (e.g., Jovanovic 1979), worker tenures can be thought of as a proxy for match quality. As such, job duration is a commonly used measure of worker quality. For example, it is the primary worker quality measure used by Autor and Scarborough (2008), who also study the impact of job testing in a low-skilled service sector setting (retail). Finally, job duration is available for all workers in our sample.

For a subset of our client firms, we also observe a direct measure of worker productivity: output per hour.[10] Again, we are not able to reveal the exact nature of the job. That said, output per hour measures the number of primary tasks that an individual worker is able to complete. For example, this would be number of words entered per hour in data entry, number of tests graded

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

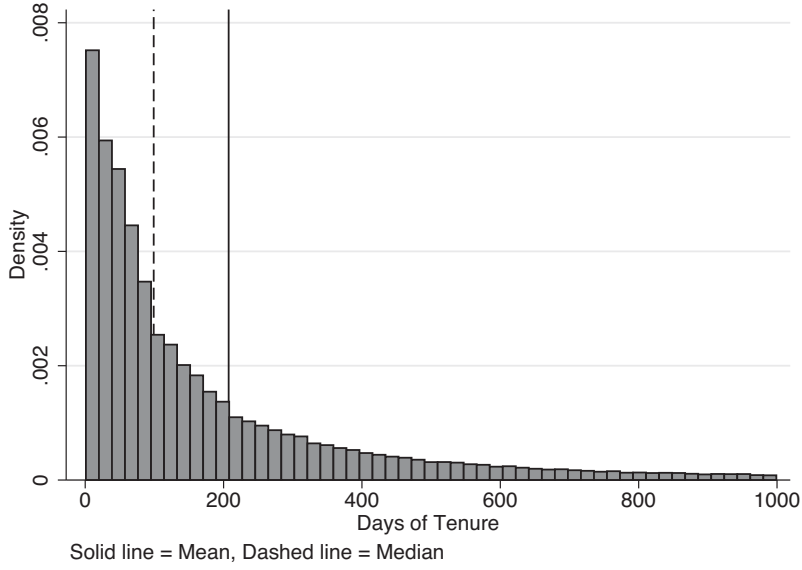

FIGURE I

Distribution of Length of Completed Job Spells

Figure I plots the distribution of completed job spells at the individual level. For legibility, this histogram (though not the computed mean or median) omits 3% of observations with durations over 1,000 days.

in test grading, or number of calls handled in call centers. Recall that in our setting, individuals perform essentially one major task and engage in individual production. Because of the discretized nature of the work, output per hour is a very common performance metric for the type of job we study and is easily measured. However, our data firm was only able to provide us with this measure for a subset of client firms (roughly a quarter of hired workers). We report these findings separately when we discuss alternative explanations.

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

Table I shows that both censored and uncensored job durations increase in color score. For example, among those with

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

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

III. The Impact of Testing

III.A. Empirical Strategy

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

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

Equation (1) compares outcomes for workers hired with and without job testing. We estimate censored normal regressions with individual-specific truncation points to account for the fact that not all workers are observed through the end of their employment spell. We regress log duration ($\log(\text{Duration})_{ilt}$) for a worker i, hired to a location l, at time t, on an indicator for whether the location l had testing at time t (Testing_{lt}). Recall, we assign the

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

test adoption date as the first month in which the modal hire at a location had a test score. After that point, the location is always assigned to the testing regime. We choose to define testing at the location, rather than individual, level to avoid the possibility that whether an individual worker is tested may depend on observed personal characteristics (Table A1 of the Online Appendix shows that our results are robust to defining testing at the individual level or at the first date on which any worker is tested).

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

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

III.B. Results

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

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

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

TABLE II
IMPACT OF JOB TESTING ON JOB DURATIONS

	Dependent variable: log(Duration)			
	(1)	(2)	(3)	(4)
Post-testing	0.368***	0.244**	0.248***	0.233***
	(0.120)	(0.113)	(0.0754)	(0.0637)
N	265,648	265,648	265,648	265,648
Year–month FEs	X	X	X	X
Location FEs	X	X	X	X
Position-type FEs	X	X	X	X
Client firm × year FEs		X	X	X
Local unemployment controls			X	X
Location time trends				X

Notes. We regress log durations on an indicator for testing availability (this equals 1 in the first month in which the modal hire at a location was tested, and in all months thereafter for that location) and the controls indicated. We use censored-normal regressions with individual-specific truncation points (using "cnreg" in Stata) to account for the fact that 21% of hired workers had not yet left their job at the end of our data collection. Standard errors are in parentheses and are clustered at the location level. FE = fixed effects. *** $p < .01$, ** $p < .05$, * $p < .1$.

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

With full controls, we find that testing improves completed job tenures by 0.23 log points, or just over 25%. These results are broadly consistent with previous estimates from Autor and Scarborough (2008).[15] These estimates reflect the treatment on the

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

15. Although our estimates are larger, the Autor and Scarborough (2008) estimate of 12% is inside the range of our 95% confidence interval with full controls.

treated effect for the sample of firms that select into receiving the sort of test we study. Given that firms often select into receiving technologies based on their expected returns (e.g., Griliches 1957), it is quite possible that other firms (e.g., those that are less open to new technologies) might experience less of a return.

In addition to log duration, we examine whether testing affects the probability that hires reach particular tenure milestones: staying at least 3, 6, or 12 months. For these samples, we restrict to workers hired 3, 6, or 12 months, respectively, before the data end date. We estimate OLS models because censoring for these variables is based only on start date and not survival time. The top panel of Online Appendix Table C1 reports results using these milestone measures. We find a positive impact of testing for all these variables, with the most pronounced effects at longer durations. For example, using our full set of controls, we find that testing increases the probability of workers surviving at least 6 months by 6 percentage points (13%) and 12 months by 7.5 percentage points (23%).

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

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

IV. Model

In this section, we formalize a model in which a firm makes hiring decisions with the help of an HR manager and a job test. As in our sample firms, managers in this model observe job test

We also estimated Cox proportional hazard models, and obtained coefficients a bit smaller in magnitude than those from censored normal models, but that were qualitatively similar. Our results are also robust to performing OLS on the length of completed job spells (as in Autor and Scarborough 2008).

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

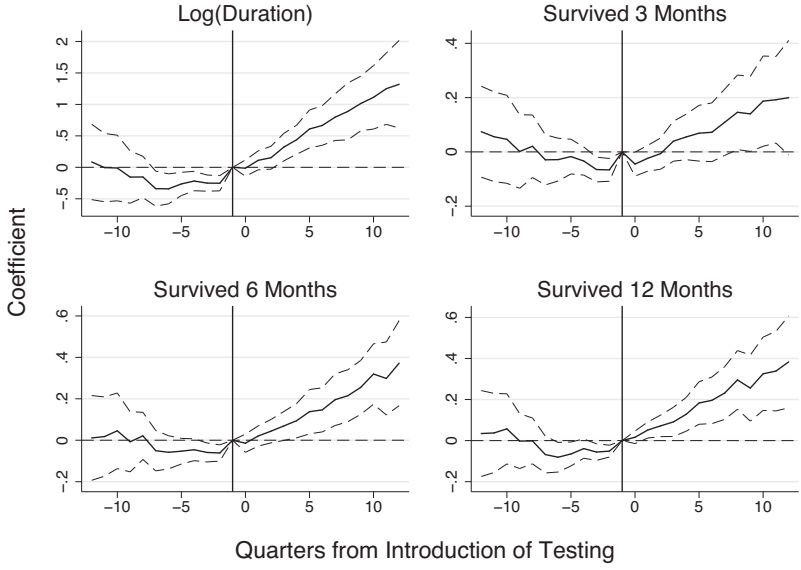

FIGURE II

Event Study of Duration Outcomes

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

recommendations, but are not required to hire strictly by the test. The main purpose of this model is to highlight the trade-offs involved in granting discretion to managers: discretion enables HR managers to take advantage of their private information but also gives them scope to make hires based on biases or incorrect beliefs that are not in the interest of the firm. The model also provides intuition for how we might empirically assess the roles of information and bias/mistakes in hiring. All proofs are in Section B of the Online Appendix.

IV.A. Setup

A mass 1 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 that depends on a worker's type, $t_i \in \{G, Y\}$; a share of workers p_G are type G, a share $1 - p_G$ are type Y, and $a|t \sim N(\mu_t, \sigma_a^2)$ with $\mu_G > \mu_Y$ and $\sigma_a^2 \in (0, \infty)$. This worker-quality distribution enables us to naturally incorporate the discrete test score into the hiring environment. We do so by assuming that the test publicly reveals t.[17]

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

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

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

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

Managers may value the firm's payoff a because they are directly incentivized to, because they risk termination or desire promotion, or because they are simply altruistic.[20] The additional quality b can be thought of in two ways. First, it may capture

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

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

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

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

managerial preferences for certain workers (e.g., for certain demographic groups or those with shared interests). Second, b can represent manager mistakes, such as overconfidence, that lead them to prefer the wrong candidates.[21]

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

The manager privately observes information about a_i and b_i. For simplicity, we assume that the manager directly observes b_i, which is distributed in the population by $N(0, \sigma_b^2)$ with $\sigma_b^2 \in \mathbb{R}_+$. Second, we assume the manager observes a noisy signal of a_i:

$$s_i = a_i + \epsilon_i,$$

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

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

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

22. We assume that managers observe the same distribution of other qualities b, but value them differently, based on the parameter k. In contrast, the

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

IV.B. Model Discussion

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

In practice, firms cannot directly observe a manager's bias or information. However, firms generally do observe the hiring choices of managers and the quality of workers that are hired. We say that when a manager chooses to overrule the test by hiring a Y worker over a G one, the manager is making an "exception." The frequency of such exceptions is increasing in both a manager's bias and the precision of their private information (Proposition 1 in Online Appendix B). That is, managers are more likely to make exceptions both when their preferences differ from those of the firm and when they have information that is not captured by the test.

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

The presence of bias raises the possibility that firms may be able to improve outcomes by placing some limits on managerial discretion. To illustrate an example, Proposition 3 (in Online Appendix B) summarizes theoretical conditions under which a firm would prefer no discretion to full discretion. In general, greater bias pushes the firm to prefer no discretion, while

---

distribution of s_i will vary across managers, based on the value of their private information.

better information pushes it toward allowing discretion. These extreme cases need not be optimal and firms may also wish to consider intermediate policies such as limiting the number of exceptions that managers can make.[23]

V. Empirical Analysis on Discretion

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

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

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

We discuss how we address these issues in the next two subsections. We first define an exception rate that takes into account

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

observable differences in applicant pools. Second, we discuss a range of empirical specifications that help deal with unobserved differences across applicant pools (i.e., differences within color or across locations).

V.A. *Defining Exceptions*

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

We then measure how often managers overrule the recommendation of the test by either (i) hiring a yellow when a green had applied and is not hired, or (ii) 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

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

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

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

However, the total number of order violations in a pool depends both on the manager's choices and on factors related to the applicant pool, such as size and color composition. For example, if a pool has only green applicants, it is impossible to make any exceptions. Similarly, if the manager needs to hire all available applicants, then there can be no exceptions. These variations were

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

implicitly held constant in our model, but need to be accounted for in the empirics. To control for pool characteristics that may mechanically impact the number of exceptions, we normalize the number of order violations by the maximum number of violations that could occur, given the applicant pool that the recruiter faces and the number of hires required.[25] This results in an exception rate that ranges from 0 if the manager never made any exceptions, to 1 if the manager made all possible exceptions. Importantly, although the propositions in Section IV are derived for the probability of an exception, their proofs hold equally for this definition of an exception rate. In Online Appendix A.6, we show that our results are also robust to alternative definitions of exception rates.

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

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

We then aggregate exception rates to the manager level by averaging over all pools a manager hired in, weighting by the number of hires in the pool. The middle panels of Figure III show histograms of manager-level exception rates: these have the same

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

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

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

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

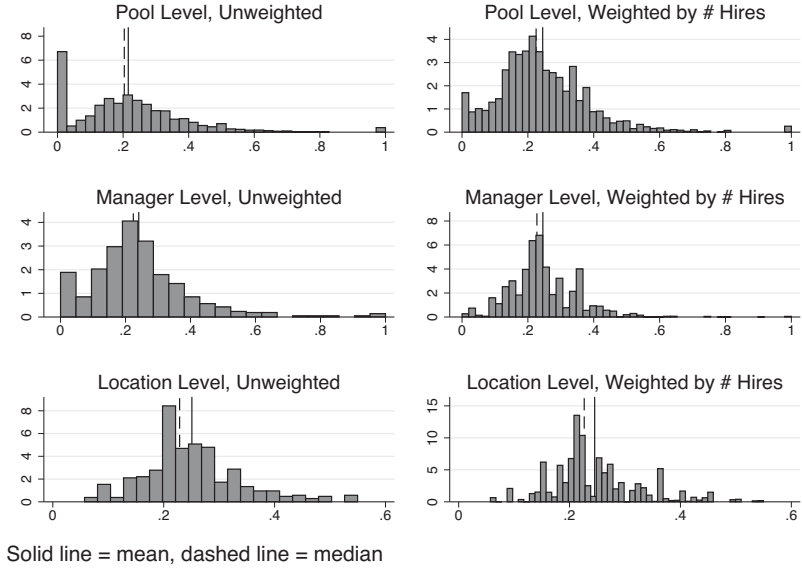

FIGURE III

Distributions of Exception Rates

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

mean and a slightly smaller standard deviation of 10 percentage points. This means that managers very frequently make exceptions, and some managers consistently make more exceptions than others. The bottom panels of Figure III aggregate exception rates to the location level and show that there is also systematic variation in exception rates across locations.

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

V.B. Empirical Specifications

1. *Post-Testing Correlation between Exception Rates and Outcomes.* We first examine the relationship between the manager-level exception rate and the realized durations of hires in the post-testing period:

$$\begin{aligned}\text{Log(Duration)}_{imlt} &= a_0 + a_1 \text{Exception Rate}_m + \delta_l + \gamma_t \\ &\quad + \text{Position}_i \beta + \epsilon_{imlt}.\end{aligned} \tag{3}$$

Our variation comes from differences in manager-level exception rates for managers employed at the same location. In our data, 99.1% of workers are hired at locations with more than 1 manager. The average location in our sample has nearly 7 managers and the average worker is at a location with 11 managers. The coefficient of interest is a_1. A negative coefficient, $a_1 < 0$, indicates that the quality of hires is decreasing in the manager's exception rate. Such a finding suggests that managers may be making exceptions because they are biased or mistaken, and not solely because they have useful private information. We cluster standard errors at the location level, again to take into account any correlation in observations within a location over time.[27]

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

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

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

In both cases, accounting for controls may alleviate some concerns. For example, location fixed effects control for fixed differences across locations in unobserved applicant quality; location-specific time trends further control for smooth changes in these characteristics. Controlling for local labor market conditions reduces the likelihood that our results are driven by "false exceptions," because such exceptions may be more common when green workers have better outside options. Our full set of controls includes location, time, and position-type fixed effects; client-year fixed effects; local labor market variables; location-specific time trends; and detailed controls for the quality and number of applicants in an application pool (fixed effects for each decile of the number of applicants, hire rate, share of applicants that are green, and share that are yellow).[28]

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

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

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

Section III.A:

$$\log(\text{Duration})_{imlt} = b_0 + b_1 \text{Testing}_{lt} \times \text{Exception Rate}_l + b_2 \text{Testing}_{lt} + \delta_l + \gamma_t + \text{Position}_i \beta + \epsilon_{imlt}. \quad (4)$$

Equation (4) includes the main effect of testing but allows testing to interact with the location-specific exception rate. We consider location-level variation in this case because we do not observe manager identifiers in the preperiod. The coefficient of interest, b_1, thus estimates how the impact of testing differs at locations that subsequently make more or fewer exceptions. A value of $b_1 < 0$ indicates that exceptions attenuate gains from testing. Unlike equation (3), which can be estimated only on post-testing data, this specification uses our full data set, making it possible for us to more precisely identify location fixed effects and other controls.

V.C. Results

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

Table III presents the accompanying regression analysis. In these regressions, we standardize exception rates to be mean 0 and standard deviation 1 so that the units are easier to interpret. Column (1) contains our base specification and indicates that a one standard deviation increase in the exception rate is associated with a 7% reduction in job durations, significant at the 5% level. Adding controls reduces the size of the standard error and the coefficient slightly. In our full-controls specification, a one standard deviation higher exception rate is associated with 6% shorter durations, still significant at the 5% level. This says

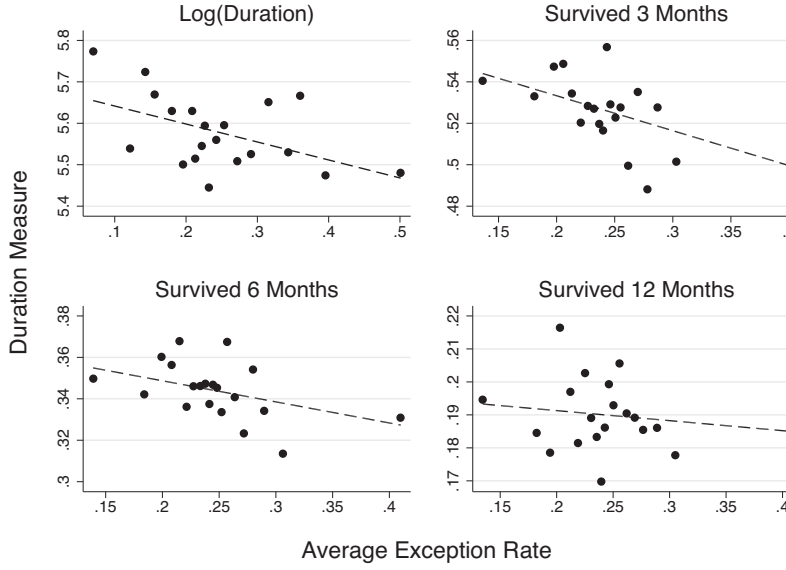

FIGURE IV

Manager-Level Exception Rates and Post-Testing Job Durations

We relate post-testing job durations to manager-level exception rates across 20 equally sized bins (weighted by number of hires). In the top left panel, we estimate a censored normal regression of log duration on the 20 bins and control variables (location, hire month, and position fixed effects), where the x-axis represents the average exception rate within each bin. On the y-axis, we plot the coefficient estimates on the 20 bins. This is implemented in Stata using "cnreg" (with no constant/intercept term included). The panel also shows the line of best fit for the 20 points. In the top right, lower left, and lower right panels, we plot means of whether workers stay 3, 6, or 12 months as a function of exception rates (controlling for location, hire month, and position fixed effects) using a binned scatter plot (using "binscatter" in Stata (Stepner 2013)). In these three panels, we restrict attention to workers hired at least 3, 6, or 12 months before the data end date for each of the 15 client firms, and we show the line of best fit.

that even when we analyze managers at the same location hiring the same number of workers out of pools that have the same share of red, yellow, and green applicants, we continue to find that managers who make more exceptions do worse. The middle panel of Online Appendix Table C1 summarizes regressions for the milestone measures, where we also find significant negative relationships.

Next, Table IV examines how the impact of testing varies by the extent to which locations make exceptions. Estimates are based on equation (4). Including the full set of controls (column

TABLE III
EXCEPTION RATES AND POST-TESTING DURATION

	Dependent variable: log(Duration)				
	(1)	(2)	(3)	(4)	(5)
Manager exception rate	−0.0682**	−0.0658**	−0.0661**	−0.0607**	−0.0557**
	(0.0346)	(0.0321)	(0.0322)	(0.0292)	(0.0283)
N	91,319	91,319	91,319	91,319	91,319
Year–month FEs	X	X	X	X	X
Location FEs	X	X	X	X	X
Position-type FEs	X	X	X	X	X
Client firm × year FEs		X	X	X	X
Local unemployment controls			X	X	X
Location time trends				X	X
Applicant pool controls					X

Notes. This table reports censored normal regressions of manager-level exception rates and tenure outcomes restricted to the post-testing sample. The exception rate is defined as the number of times a yellow is hired above a green plus the number of times a red is hired above a green or yellow, divided by the maximum exceptions possible in that applicant pool. It is then aggregated to the manager level and standardized to be mean 0 and standard deviation 1. Applicant pool controls include fixed effects for deciles of each of the following variables: number of applicants, hire rate, share of applicants that are green, and share that are yellow. Standard errors are clustered by location. *** $p < .01$, ** $p < .05$, * $p < .1$.

(4)), we find that at the mean exception rate (recall that we standardize exception rates), testing increases durations by 0.23 log points, but this effect is substantially offset (by 0.14 log points) for each standard deviation increase in the exception rate, significant at the 5% level.[29] The bottom panel of Online Appendix Table C1 shows that these results are robust to OLS estimates using milestone measures as dependent variables.

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

Across a variety of specifications, we consistently find that worker tenure is lower for managers who made more exceptions to test recommendations. The magnitude of this estimate implies

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

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

TABLE IV
THE IMPACT OF TESTING BY EXCEPTION RATE

	Dependent variable: log(Duration)			
	(1)	(2)	(3)	(4)
Post-testing	0.366***	0.234**	0.242***	0.234***
	(0.119)	(0.107)	(0.0696)	(0.0589)
Location exception	−0.142**	−0.150**	−0.152**	−0.142**
Rate * post-testing	(0.0652)	(0.0679)	(0.0655)	(0.0573)
N	265,648	265,648	265,648	265,648
Year–month FEs	X	X	X	X
Location FEs	X	X	X	X
Position-type FEs	X	X	X	X
Client firm × year FEs		X	X	X
Local unemployment controls			X	X
Location time trends				X

Notes. This table reports censored normal regressions of the differential impact of testing adoption, by location-level exception rate. We use the same sample as defined by the notes to Table II. The exception rate is defined as the number of times a yellow is hired above a green plus the number of times a red is hired above a green or yellow, divided by the maximum exceptions possible in that applicant pool. It is then aggregated to the location level and standardized to be mean 0 and standard deviation 1. *** $p < .01$, ** $p < .05$, * $p < .1$.

that a firm made up of managers at the 10th percentile of the exception distribution would have approximately 15% longer worker durations in the post-testing period, relative to a firm made up of 90th percentile managers (higher exception rates are worse). Further, locations at the 10th percentile of the exception distribution experience duration improvements with the adoption of testing that are multiple times larger than improvements at 90th percentile locations.

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

V.D. Additional Robustness Checks

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

1. *Quality of "Passed Over" Workers.* There are some scenarios under which we may find a negative correlation between worker outcomes and exception rates, even when managerial discretion improves hiring. For example, as discussed earlier, a manager may

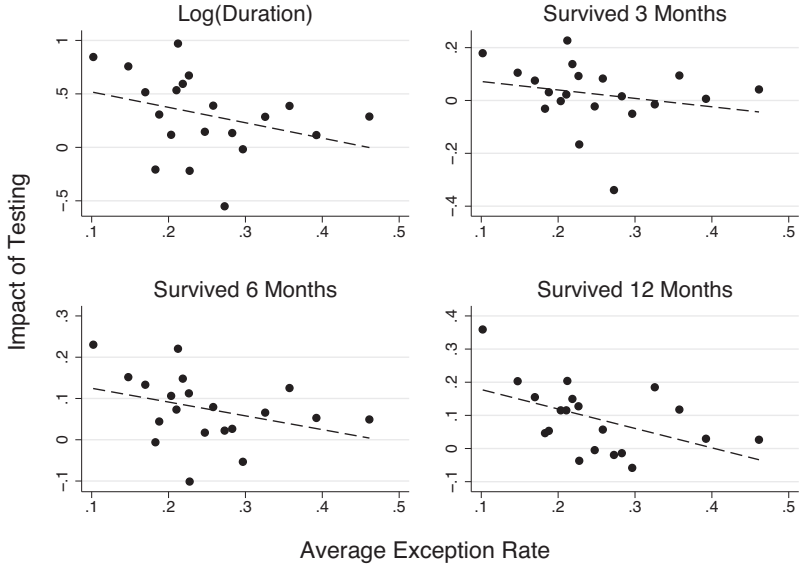

FIGURE V

Location-Level Exception Rates and the Impact of Testing on Job Durations

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

tend to make more exceptions because he or she sees idiosyncratically weak green applicants, relative to the typical applicants at the location. As another example, locations with high exception rates may benefit less from the test because its managers always have better private information.

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

TABLE V

TENURE OF EXCEPTIONS VERSUS PASSED OVER APPLICANTS

	Dependent variable: log(Duration)	
	(1)	(2)
Quality of yellow exceptions versus passed over greens		
Passed over greens	0.0402*	0.0449
	(0.0220)	(0.0357)
N	53,166	53,166
Quality of red exceptions versus passed over greens and yellows		
Passed over greens	0.159***	0.143**
	(0.0543)	(0.0634)
Passed over yellows	0.143***	0.121**
	(0.0546)	(0.0597)
N	25,782	25,782
Base controls	X	X
Comparison pool FEs		X

Notes. Regressions are restricted to the post-testing sample, adjust for censoring, and standard errors are clustered at the location level. The top (bottom) panel compares yellow (red) exceptions—the omitted category—to passed over greens (and yellows) who were available at the same time but hired in a later month. Observations are restricted to pools with at least one exception and one passed over worker, and are further restricted to locations and pools with at least 10 and 5 observations, respectively. Base controls are location, hire month, and position-type fixed effects. Comparison pool fixed effects are defined by the manager–location–month for the applicant pool in which candidates were considered together. *** $p < .01$, ** $p < .05$, * $p < .1$.

working in a later month. If managers make exceptions when they have better information, then exception yellows should have longer tenures than "passed over" greens.

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

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

comparing exception reds (the omitted group) to passed over yellows and greens.[32]

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

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

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

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

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

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

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

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

Our final set of results provides evidence that this is unlikely to be the case. For a subset of client firms, we observe a direct measure of worker productivity: output per hour. Recall that in our setting, individuals perform essentially one major task and engage in individual production. Some examples may include the number of data items entered per hour, the number of standardized tests graded per hour, and the number of phone calls completed per hour. As in these examples, output per hour is an important measure of productivity for the fairly homogeneous task we study.

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

This measure is available for 62,427 workers (one-quarter of all hires) in six client firms. The primary reason for missing output is that the metric is not made available to us for many locations, time periods, and end clients (i.e., the ones purchasing services from the client firms).[35] In addition, its availability depends on

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

35. We can account for half of the variation in whether an output measure is available for an individual worker with location, time, and position controls.

TABLE VI
TESTING, EXCEPTION RATES, AND OUTPUT PER HOUR

	Dependent variable: output per hour			
	(1)	(2)	(3)	(4)
	Post-testing sample		Introduction of testing	
Post-testing			0.343	0.156
			(0.366)	(0.327)
Exception rate × post-testing	−0.0659	−0.111	0.137	−0.137
	(0.134)	(0.0953)	(0.153)	(0.217)
N	28,858	28,858	62,421	62,421
Base controls	X	X	X	X
Full controls		X		X

Notes. See notes to Tables III and IV. The dependent variable in this case is output per hour and regressions are estimated with OLS. In columns (1) and (2), we examine the post-testing correlation between the manager-level exception rate and output per hour. In columns (3) and (4), we examine the differential impact of testing as a function of location-level exception rates. Base controls include location, hire month, and position fixed effects, as well as controls for the number of tasks in a day that are used to measure a worker's output per hour. Full controls add client-by-year effects, local unemployment rates, and location-specific time trends. For the post-testing sample regressions (columns (1) and (2)), full controls also include applicant pool controls. *** $p < .01$, ** $p < .05$, * $p < .1$.

workers completing their training and being permitted to perform the job task: this is the period in which workers become valuable to client firms.

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

Table VI summarizes our main analyses using output per hour as the dependent variable. Columns (1)–(2) document the post-testing correlation between manager-level exceptions and output per hour. For both our base and full sets of controls, we obtain negative coefficients that are not significant. For example in column (2), the estimate is −0.11 with a standard error of 0.095.

According to the data firm, certain lines of business within a firm do not make their productivity data available.

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

Recall the manager-level exception rates are standardized to be mean 0 and standard deviation 1 in the full post-testing sample. The estimate therefore implies that a one standard deviation higher exception rate manager hires workers who perform 0.11 fewer units of output per hour, on average. This is 2.3% of the standard deviation for output per hour (4.7, mentioned above). Based on the standard error, we can rule out positive effects beyond 1.7% and negative effects beyond −6.4% of a standard deviation with 95% confidence.

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

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

VI. Conclusion

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

a range of specifications, we show that hiring against test recommendations is associated with worse outcomes.

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

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

More broadly, our findings highlight the role that technology can play in reducing the impact of managerial mistakes or biases

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

by changing how decision making is structured within the firm. As workforce analytics become an increasingly important part of HR management, more work needs to be done to understand how such technologies interact with organizational structure and the allocation of decision rights within the firm. This article makes an important step towards understanding and quantifying these issues.

UNIVERSITY OF TORONTO AND NATIONAL BUREAU OF ECONOMIC RESEARCH
YALE UNIVERSITY AND NATIONAL BUREAU OF ECONOMIC RESEARCH
MIT AND NATIONAL BUREAU OF ECONOMIC RESEARCH

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online. Code replicating tables and figures in this article can be found in Hoffman, Kahn, and Li (2017) in the Harvard Dataverse, doi:10.7910/DVN/DWPXXT.

REFERENCES

Aghion, Philippe, and Jean Tirole, "Formal and Real Authority in Organizations," *Journal of Political Economy*, 105 (1997), 1–29.
Alonso, Ricardo, and Niko Matouschek. "Optimal Delegation," *Review of Economic Studies*, 75 (2008), 259–293.
Altonji, Joseph, and Charles Pierret, "Employer Learning and Statistical Discrimination," *Quarterly Journal of Economics*, 116 (2001), 313–350.
Autor, David, "Why Do Temporary Help Firms Provide Free General Skills Training?," *Quarterly Journal of Economics*, 116 (2001), 1409–1448.
Autor, David, and David Scarborough, "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments," *Quarterly Journal of Economics*, 123 (2008), 219–277.
Baker, George, and Thomas Hubbard, "Contractibility and Asset Ownership: On-Board Computers and Governance in U.S. Trucking," *Quarterly Journal of Economics*, 119 (2004), 1443–1479.
Bolton, Patrick, and Mathias Dewatripont "Authority in Organizations," in *The Handbook of Organizational Economics*, Robert Gibbons and John Roberts, eds. (Princeton, NJ: Princeton University Press, 2013).
Brown, Meta, Elizabeth Setren, and Giorgio Topa, "Do Informal Referrals Lead to Better Matches? Evidence from a Firm's Employee Referral System," *Journal of Labor Economics*, 34 (2016), 161–209.
Burks, Stephen, Bo Cowgill, Mitchell Hoffman, and Michael Housman, "The Value of Hiring through Employee Referrals," *Quarterly Journal of Economics*, 130 (2015), 805–839.
Dana, Jason, Robyn Dawes, and Nathaniel Peterson, "Belief in the Unstructured Interview: The Persistence of an Illusion," *Judgment and Decision Making*, 8 (2013), 512–520.
Dessein, Wouter, "Authority and Communication in Organizations," *Review of Economic Studies*, 69 (2002), 811–838.
Diamond, Rebecca, and Petra Persson, "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests," Mimeo, Stanford University, 2016.

Farber, Henry, and Robert Gibbons, "Learning and Wage Dynamics," *Quarterly Journal of Economics*, 111 (1996), 1007–1047.
Frankel, Alexander, "Selecting Applicants," Mimeo, University of Chicago, 2017.
Frederiksen, Anders, Lisa B. Kahn, and Fabian Lange, "Supervisors and Performance Management Systems," NBER Working Paper No. 23351, 2017.
Griliches, Zvi, "Hybrid Corn: An Exploration in the Economics of Technological Change," *Econometrica*, 25 (1957), 501–522.
Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li. "Replication Data for 'Discretion in Hiring'," *Harvard Dataverse* (2017), doi:10.7910/DVN/DWPXXT.
Hoffman, Mitchell, and Steven Tadelis. "People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis," Working Paper, University of Toronto, 2017.
Horton, John, "The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment," *Journal of Labor Economics*, 35 (2017), 345–385.
Jovanovic, Boyan, "Job Matching and the Theory of Turnover," *Journal of Political Economy*, 87 (1979), 972–990.
Kahn, Lisa B., and Fabian Lange, "Employer Learning, Productivity and the Earnings Distribution: Evidence from Performance Measures," *Review of Economic Studies*, 81 (2014), 1575–1613.
Kahneman, Daniel, *Thinking Fast and Slow* (New York: Farrar, Strauss, and Giroux, 2011).
Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 133 (2018), 237–293.
Kuncel, Nathan, David Klieger, Brian Connelly, and Deniz Ones, "Mechanical Versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis," *Journal of Applied Psychology*, 98 (2013), 1060–1072.
Kuziemko, Ilyana, "How Should Inmates Be Released from Prison? An Assessment of Parole versus Fixed Sentence Regimes," *Quarterly Journal of Economics*, 128 (2013), 371–424.
Lazear, Edward, Kathryn Shaw, and Christopher Stanton, "The Value of Bosses," *Journal of Labor Economics*, 33 (2015), 823–861.
Li, Danielle, "Expertise and Bias in Evaluation: Evidence from the NIH," *American Economic Journal: Applied Economics*, 9 (2017), 60–92.
Li, Danielle, and Leila Agha, "Big Names or Big Ideas: Do Peer Review Panels Select the Best Science Proposals?," *Science*, 348 (2015), 434–438.
Pallais, Amanda, and Emily Sands, "Why the Referential Treatment? Evidence from Field Experiments on Referrals," *Journal of Political Economy*, 124 (2016), 1793–1828.
Paravisini, Daniel, and Antoinette Schoar, "The Incentive Effect of IT: Randomized Evidence from Credit Committees," NBER Working Paper No. 19303, 2013.
Rivera, Lauren, "Hiring as Cultural Matching: The Case of Elite Professional Service Firms," *American Sociological Review*, 77 (2012), 999–1022.
Stanton, Christopher, and Catherine Thomas, "Landing the First Job: The Value of Intermediaries in Online Hiring," *Review of Economic Studies*, 83 (2016), 810–854.
Stepner, Michael, "BINSCATTER: Stata module to generate binned scatterplots," Statistical Software Components, Boston College Department of Economics, 2013.
Wang, James. "Why Hire Loan Officers? Examining Delegated Expertise," Mimeo, University of Michigan, 2014.