Discretion in Hiring: Online Appendix

Mitchell Hoffman
University of Toronto & NBER

Lisa B. Kahn
Yale University & NBER

Danielle Li
MIT & NBER

Section A is our Data Appendix. Section B is our Theory Appendix, accompanying Section IV in the main text. Section C contains supplemental tables and figures.

A Data Appendix

The 15 client firms in our sample each obtained testing services from our data provider. In this section, we describe the introduction of testing across locations within these firms, as well as give further information on the data. We first provide some details about the test itself (Section A.1). We then discuss how we assign the date at which testing is introduced to a location and demonstrate the robustness of our main results to this definition (Section A.2). We also describe sample coverage over time across client firms and show robustness to using more balanced panels (Section A.3). We then discuss the timing of test adoption (Section A.4). We further provide a discussion of heterogeneity in test accuracy across locations (Section A.5) and discuss alternative exception rate definitions (Section A.6). Finally, we provide details on sample restrictions (Section A.7), as well as provide additional information about the setting and data set (Section A.8).

A.1 The Job Test

The test is designed to take around 30-60 minutes, though its intended length varies by firm (e.g., according to whether the test covers multiple positions) and consists of several sections. Applicants generally take the test in addition to submitting standard application information (such as a resume). The test includes an introductory section describing the job and work environment, and asks the applicant if he/she thinks they are well-suited for the job and about eligibility. Following this section, there are questions on many dimensions, including those on work experience, computer/technical skills, personality traits, cognitive skills, hypothetical job scenarios, and workplace simulations. The hypothetical job scenarios reflect issues that may arise in performing the specific task we study: for example, if this were a data entry job, it may ask what the employee would do if she were unable to understand the data entry interface. In the workplace simulations, applicants are asked to perform part of the job itself. For example, if this were a data entry job, the applicant may be asked to read an input file and enter the relevant data.

Our data firm uses a proprietary algorithm based on candidates' responses to generate test scores. This algorithm varies somewhat by client firm, but there are commonalities, and the algorithm is updated over time as more data arrives. The algorithm used by a given firm will include data from that particular firm, as well as data from other firms. Correlations are analyzed between various questions and employee attrition (a key outcome), as well as between the various questions and other outcomes (depending on the client firms), particularly output per hour, as well as output quality. In its promotional materials as well

as in its conversations with us, our data provider has stressed the importance of attrition as a key outcome.

The central output of the test is a Red/Yellow/Green score (or scores if the test covers multiple positions) for each applicant. Recruiters observe overall job test scores, but do not observe underlying information on data such as cognitive skills, personality, or how applicants would handle various job scenarios.[1]

A.2 Assigning Testing Adoption Dates to Locations

We observe the date at which test scores appear in our data, but not all workers are tested immediately. Our preferred definition assigns testing to begin at a location when the modal hire in a cohort has a test score. At this point, testing "turns on" for the location for the remainder of our sample period.

Within locations, testing appears to be adopted quickly. Appendix Figure A1 plots the share of hires who are tested as a function of time relative to when the modal hire at that location is tested. This shows that testing ramps up very quickly within a location, reaching roughly 80% coverage almost immediately and continuing to increase to nearly 100% by the end of our sample period.[2] This supports our defining test-adoption as the first month in which the modal hire at a location is tested.

Appendix Table A1 shows that our results are robust to this testing definition. Column 1 replicates our base specifications from Tables II and IV in the main text for the introduction of testing (top panel) and the differential impact of testing across exception rates (bottom panel).[3] These results are very similar when the alternative testing definitions used in Columns 2 and 3. Column 2 defines testing adoption as the first date in which any hire is tested, while column 3 assigns testing at the individual level.

[1] Beyond the central Red/Yellow/Green score (or scores), recruiters observe information on typing speed and accuracy, and, for some firms and time periods, information on an additional job-related skill, but these do not enter into the scoring of Red/Yellow/Green. Results are robust to controlling for typing variables where available, which accounts for the possibility that some locations may have had typing threshold hiring rules. In addition, recruiters could observe information on responses submitted during the introductory section (e.g., whether applicants may have a work schedule issue). Further, recruiters had the option to observe several performance prediction scores that go into the final Red/Yellow/Green score; however, these also represent overall job test scores (as opposed to underlying information on data such as cognitive skills, personality, and job scenarios).

[2] According to the data firm, non-tested individuals are primarily those hired from job fairs. Also, our data contain a small number of non-frontline workers (such as managers and professionals) who are not tested. These workers are distinguished in our position controls. Last, it is possible that testing could be rolled out to hiring for particular end clients within a location (but not for others).

[3] Results from Table III from the main text on the correlation between manager-level exception rates and outcomes of hires do not rely on a comparison of pre and post-testing data so are not included.

A.3 Sample Coverage within Locations over Time

Based on our preferred definition of testing, 97 out of 127 locations receive testing at some point during our sample period; 83 locations are observed both before and after testing. Locations observed only before or after testing are included in our regressions and help identify coefficients on controls. However, Column 4 of Appendix Table A1 shows that our results on the impact of testing are robust to restricting to a balanced panel of locations that are observed both before and after testing.

Appendix Figure A2 provides a summary of our sample coverage over time for all locations. We collect locations by client firm on the y-axis and plot a dot for each month the location hires in, with calendar time indicated on the x-axis. Hollow circles indicate that testing had not yet been introduced to the location, based on our preferred measure; filled in circles indicate the post-testing period. A gap between circles indicates no hires were made in that month.

This figure indicates that we observe cohorts of workers for many periods both before and after testing for most locations. Specifically, among the 83 locations that hire both before and after testing, the average observation window post-testing is 15 months and the average pre-testing observation window is 3.5 years (worker weighted). Furthermore, 90% of hires in this sample are to a location that can be observed for at least 6 months before and after testing, 60% are hired to locations with at least a 1 year window around testing. Of course, the panel is highly unbalanced and there is a range of observation windows for clients and locations.[4]

From the figure, locations also appear to hire in most months during their observation window. In fact, of the locations that can be observed for at least a full year before and after testing, three-quarters (worker weighted) hire in every single quarter in that window. Column 5 of Appendix Table A1 shows that results on the impact of testing are robust to restricting to this very balanced panel of locations. Results are similar across a wide range of balanced panels.

Furthermore, Appendix Figure A3, replicates the event study for the impact of testing, restricting to locations that hire in each of the four quarters before and after testing, and shows a very similar picture to Figure II of the main text.

Finally, as noted in the main text, the data firm informed us that a number of client firms had some other form of testing before the introduction of our data firm's test. While information about whether a client firm had testing before our data provider is not part of our dataset, we asked our data provider to collect information about this on our behalf by surveying managers and executives at the data firm. From this, the data firm reported

[4]For example, client #13 has no pre-testing data.

that 5 firms had pre-sample testing (and not just in one part of its business), 1 firm had pre-sample testing in one part of its business, 1 firm was believed to have pre-sample testing (but our data firm was not certain), and 8 firms were regarded as either not having testing or believed not to have testing.

This survey does not provide certainty for all 15 client firms in our data. However, column 6 of Appendix Table A1 shows that key coefficients are larger on the sample of firms who likely did not have pre-sample testing. This is consistent with testing being more of an improvement for firms that had no alternative test in the pre-period, as well as it being more useful for managers to follow test recommendations rather than make exceptions at these firms.

A.4 Timing of Testing and Location Observables

Appendix Figure A4 describes how testing enters our sample across both client firms and locations. Circles indicate the date at which testing is adopted for the 97 locations that ever receive testing during our sample (x-axis). Locations are collected by client firm and lined up on the y-axis in the order of their specific test adoption date. The size of each circle reflects the location's size.[5] Among client firms with more than one location (11 out of 15 firms, accounting for 94% of hires in our data), most firms adopt testing across all their locations in our sample in under 2 years. There does not appear to be a systematic relationship between the size of a location and the time at which it receives testing.

In Section III of the main paper, we exploited this gradual roll-out of testing across locations within client firms to estimate the impact of testing on job durations, while controlling for location and hire date fixed effects. Naturally, one may be concerned about factors leading clients to introduce testing in some locations before others. However, based on qualitative and quantitative information, we see no evidence that the timing of this roll out would bias our results.

On the qualitative side, we had discussions involving different individuals from our data provider (including one person who worked closely with different firms on rolling out testing), as well as managers from a large client firm in our dataset. Representatives mentioned several possible drivers of testing adoption, including the availability/"bandwith" of managers to oversee the adoption of testing, considerations of geography, the openness of end clients (i.e., the ones paying for the services provided by our client firms) to testing, and whether a location had historically high attrition. Importantly, representatives did not say that

[5] We define the size of the location as the number of workers currently employed in July 2013. For one location we must use July 2012 instead. This snapshot date avoids overweighting locations that have high churn.

firms may have adopted testing in ways that reflect time-varying differences in a location's attrition risk. For example, no one mentioned bringing in testing to a location that was recently experiencing or expecting a retention problem.

On the quantitative side, we have examined the correlation between location-level observables and the timing of testing adoption. For example, Appendix Figure A5 plots location characteristics as a function of testing adoption date for several key variables. Circles and the fitted regression line are again weighted by location size, and durations are censoring adjusted.

The top panels show relationships for pre-testing characteristics at the location level. In the top left panel, we find no systematic relationship between a location's average pre-testing duration (censoring adjusted) and the date at which it adopts testing. The top middle panel considers a location-specific time trend in censoring-adjusted durations pre-testing.[6] This gradient is also quite flat: testing does not arrive earlier or later for locations that are on a stronger or weaker trend in worker duration. Finally, the top left panel plots the average unemployment rate among workers with exactly a High School Diploma pre-testing. Here, there is again no relationship between the testing date and local labor market conditions. We use the state-level unemployment rate for high school graduates (a high school diploma is a typical educational requirement for many low-skill service jobs), but the graph looks similar for unemployment rates for other groups.[7]

The bottom panel of Appendix Figure A5 focuses on variables that are available only after testing: the share of applicants with a green test score, the average number of applicants per month, and the average exception rate across HR managers at that location (see Equation 2). Again, we do not find a clear pattern for these dependent variables.

We also point out that the linear relationships in these graphs tend to be statistically insignificant and small in magnitude. For example, we can rule out a plus or minus 1.5% change in pre-testing average durations with each month that testing is delayed with 95% confidence. We can similarly rule out a plus or minus 0.2% change in the share of applicants that are green. We can also rule out a plus or minus 0.004 variation in the location exception rate. We have examined a wide range of location characteristics and similarly find little systematic or robust relationships with timing of testing. Notably, these include pre-testing averages for the share of months that the location is active in hiring and the location-specific churn rate.

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

[7] The graph also looks similar when using aggregated unemployment rates for the 25% of international locations and when using U.S.-level unemployment rates for each education group for the non-standard location identifiers where state cannot be easily assigned.

A.5 The Accuracy of Test Scores Across Locations

One may be worried that the test does not predict worker quality equally well across locations. For example, worse establishments may be especially undesirable for more skilled workers, resulting in lower durations among greens.

Appendix Figure A6 plots the relationship between manager-level exception rates and worker duration, separately by color. We estimate censored normal regressions of log duration on indicators for each of 20 equally sized (based on number of hires) exception rate bins and base controls (location, hire month, and position fixed effects, and with no constant/intercept term included), separately by color. There are two main patterns to notice. First, greens perform better than yellows who in turn perform better than reds across exception rate bins. Second, the overall quality of hired yellows and greens is broadly stable across exception rates. This means that, among workers a manager is able to hire, color score is predictive of performance, even across varying manager exception rates.

We do see some evidence that reds hired by managers who make many exceptions appear worse than reds hired by managers who make few exceptions. This could be because reds in these locations are worse or because managers with high exception rates are especially bad at picking out reds. In either case, this reinforces the point that, in high exception locations, managers may do better by hiring more greens and yellows, relative to reds: the greens and yellows they are able to hire are broadly comparable to the quality of greens and yellows in low exception locations, while the reds they hire appear somewhat worse.

Appendix Figure A7 provides more information along these lines. We plot the relationship between color score and job duration as a function of the same set of location-level characteristics reported above in Appendix Figure A5. We divide locations into 20 equally sized bins (based on number of hires post-testing).[8]

Across the bins for the panels in Appendix Figure A7, we find that color score is predictive of job durations. For example, the top left panel plots the relationship for the average duration of the location pre-testing and shows three upward sloping parallel lines. This means that job durations are increasing in the average quality of the location pre-testing, naturally. However, the gap between job durations by color is roughly constant across bins. The other panels suggest a similar conclusion: we fail to see systematic evidence that the predictiveness of the color score varies much with location characteristics.[9]

[8] We estimate these figures in a similar manner as Appendix Figure A6, with 20 bins for location-level characteristics instead of manager-level exception rates (and we exclude location fixed effects because they are collinear with the location characteristics).

[9] For each subfigure in Figure A7 (as well as for Figure A6), if the 3 regressions are run without control variables, or if the 3 regressions are run instead as one regression with color x ventile interactions and control variables, the differences in duration between the 3 colors become smaller. However, the same qualitative

A.6 Alternative exception rate definitions

In Appendix Table A2, we examine the robustness of our main results in Tables III and IV from the main text to alternative ways of defining an exception rate. Recall that we construct our exception rate by counting the number of order violations (the number of greens that are passed over by each hired yellow, plus the number of greens and yellows that are passed over by each hired red) and normalizing by the maximum number possible, given the same color composition of applicants and total number of hires.

First, we consider an alternative normalization: the number of order violations that would occur if managers hired at random. The random benchmark is interesting because this is the number of exceptions that would occur if managers ignored the test and the test were uninformative for quality. In our data, 86% of workers are hired from application pools in which this exception rate is less than 1, indicating that, in the vast majority of pools, managers' decisions align with test recommendations to some extent. Next, we consider a different way of conceptualizing the exception rate, using the idea of a "score" rather than a violation: 2 points for every green hire, 1 point for every yellow hire, and no points for red hires. We count up scores per applicant pool and normalize by either the maximum possible score, or the score that would obtain under random hiring. The score measure differs conceptually from the order violation approach because it is less sensitive to the number of unhired applicants. For example, the score is the same if a single yellow worker is hired over 20 greens, or over only one green.[10] We negate the score metrics so that a larger number means more exceptions, to align with the order violation measure. All three of these measures are aggregated to the manager and location-levels and then standardized. Appendix Table A2 shows that all of these metrics tell similar stories. Results are robust quantitatively and generally in terms of statistical significance as well.

A.7 Sample Restrictions

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

1. We drop roughly one third of applicants because they have a missing identifier for their HR manager.[11]

message remains that color is predictive of duration, and this holds across varying levels of manager exception rates or location characteristics.

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

[11] To assess the possibility of selection bias, we regressed whether HR manager is missing on duration (or log duration), a dummy for being censored, location controls, month-year of hire dummies, and position dummies, using the full sample of tested hires. In the two regressions, the coefficients on duration and log duration are statistically insignificant, suggesting that selection bias is not a main concern for our analysis. In forming the applicant pool, we also drop about 200 people (almost all from one firm) where there appears

2. We drop 2% of hires that are part of pools with less than 3 applicants.

3. We drop locations that do not have at least two managers because part of our exception rate analysis (Equation (3)) relies on within-location variation in manager-level exception rates. This drops 2% of remaining managers associated with 0.9% of remaining hires.

4. We drop pools that hire only exceptions because we worry that an idiosyncratic shock drives the lack of matriculation of higher scoring applicants. This reflects 8% of the remaining pools associated with 0.6% of remaining hires.

5. We drop managers that hire in only 1 pool to clean out some noise in the manager-level exception rates. This reflects 16% of the remaining managers associated with 0.55% of remaining hires.

6. We drop observations with missing manager-level exception rates, which occur when all pools a manager hires to have a value of 0 for the maximum number of possible exceptions. This reflects 1.5% of the remaining pools associated with 0.06% of remaining hires.

We implement these restrictions for the post-testing period in all analyses, even those that do not use exception rates, to keep the sample consistent. However, results from Section III on the impact of testing (which do not use exception rates) are similar without the restrictions. We do not exclude observations in the pre-testing period on the basis of being associated with locations that do not meet our post-testing criteria, as these observations help identify cohort, client, and position controls. Further, for all analyses, we drop the four locations (reflecting 0.04% of remaining hires) with less than 50 hires over the sample period. Section A.8 below describes a few further sample restrictions.

A.8 Further Information on Setting and Data

Firms in the Data. The data were assembled for us by the data firm from records of the individual client firms. The client firms in our sample employ workers who are engaged in the same job, but there are some differences across the firms along various dimensions. For example, at one firm, workers engage in a relatively high-skilled version of the job we study.[12]

to be an error in the application date; also, for about 1,500 people, we fill in missing application month using the hire month minus one.

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

At a second firm, the data firm provides assistance with recruiting (beyond providing the job test). Our baseline key results are similar when individual firms are excluded one by one.[13]

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 hypothetical firm that provided pre-testing data on new hires going back to Jan. 2010. For this firm, for workers hired before Jan. 2010, we would only observe the subset of workers 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. In forming our regression sample, we also drop a couple thousand hired workers who have a missing duration variable in the data, most of whom are from the pre-testing period.

Productivity. In addition to the information on job durations, some client firms provide data on output per hour. This is available for about a quarter of hired workers in our sample. We trim instances where average transaction time in a given day is less than 60 seconds.[14]

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.[15]

For candidates in our data with at least one Red/Yellow/Green score, roughly one third have one score, roughly half have two scores, and the remainder have more than two scores in our data. Among candidates with multiple scores, the scores are very highly correlated with one another. For example, scores for the two most common positions have a correlation

[13] Specifically, we estimated base specifications of Tables 2, 3, and 4 from the main text excluding each firm one by one.

[14] This is about one percent of transactions. 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).

[15] For 1 of the 15 client 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. For another 1 of the 15 client firms, we fill in about 500 missing observations using a constructed test score variable from the data firm that exists primarily for hires. Our conclusions are also substantively unchanged if these observations are omitted.

coefficient of 0.88 (for Red=0, Yellow=1, Green=2).[16] Our focus on the maximum of scores thus seems without much loss of generality.[17]

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

Other managers may take part in hiring decisions as well. As noted in footnote 6 of the main text, one firm said that its HR managers will typically endorse candidates to an additional manager (e.g., a manager in operations one rank above the frontline supervisor) to make a "final call." That said, HR managers play a critical role in deciding who gets hired. For low-skilled jobs of the type we study, past work suggests that HR managers play an active role in hiring; for example, in a detailed study by sociologists of a call-center at a bank, Fernandez, Castilla, and Moore (2000) report that HR managers played an important role in the recruiting process, even though there was a second interview that was done by line managers during their study period. In fact, the importance of HR managers at this particular firm happened to grow after the study: HR managers were granted authority to make hiring decisions on their own.

Also, applicants may interact with more than one HR manager during the recruitment process. In such cases, we assign an applicant to the HR manager with whom they have the most interactions.[18] Most managers are primarily associated with one location, but some are associated with multiple locations.

We do not observe manager incentives in our data. However, a manager from our data provider informed us that recruiters in our setting often receive a financial incentive to meet or exceed several targets (while pointing out that such pay structures are highly variable by firm). He said that recruiters always have targets with respect to fill rate (e.g., a requisition of 20 new hires to begin work on March 1st), and often have targets with respect to short-term tenure (e.g., a certain share of people graduating training, or of staying some length of time, such as 90 days) or activities (e.g., conducting X interviews or reaching out to Y candidates).

Race, Gender, Age. Data on race, sex, and age are not available for this project. However, Autor and Scarborough (2008) show that job testing does not seem to affect worker

[16] This is the correlation in the raw data before imposing data restrictions.

[17] Applicants may be considered for multiple positions and it is difficult to discern which is the most relevant score (or scores) for a given applicant.

[18] This excludes interactions where information on the HR manager is missing. If there is a tie for most interactions, we assign an applicant to only one manager. Our main results are also qualitatively robust to setting the HR manager identifier to missing in cases of ties for most interactions.

race, suggesting that changes in worker demographics such as race are not the mechanism by which job testing improves durations.

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.[19]

As explained in the main text, there are a small number of non-standard location identifiers (e.g., those where workers generally work off-site in different states). We assign these locations to unemployment rates using education-specific, US national unemployment data. We do so even though a small share of workers associated with non-standard location identifiers may be outside the US.

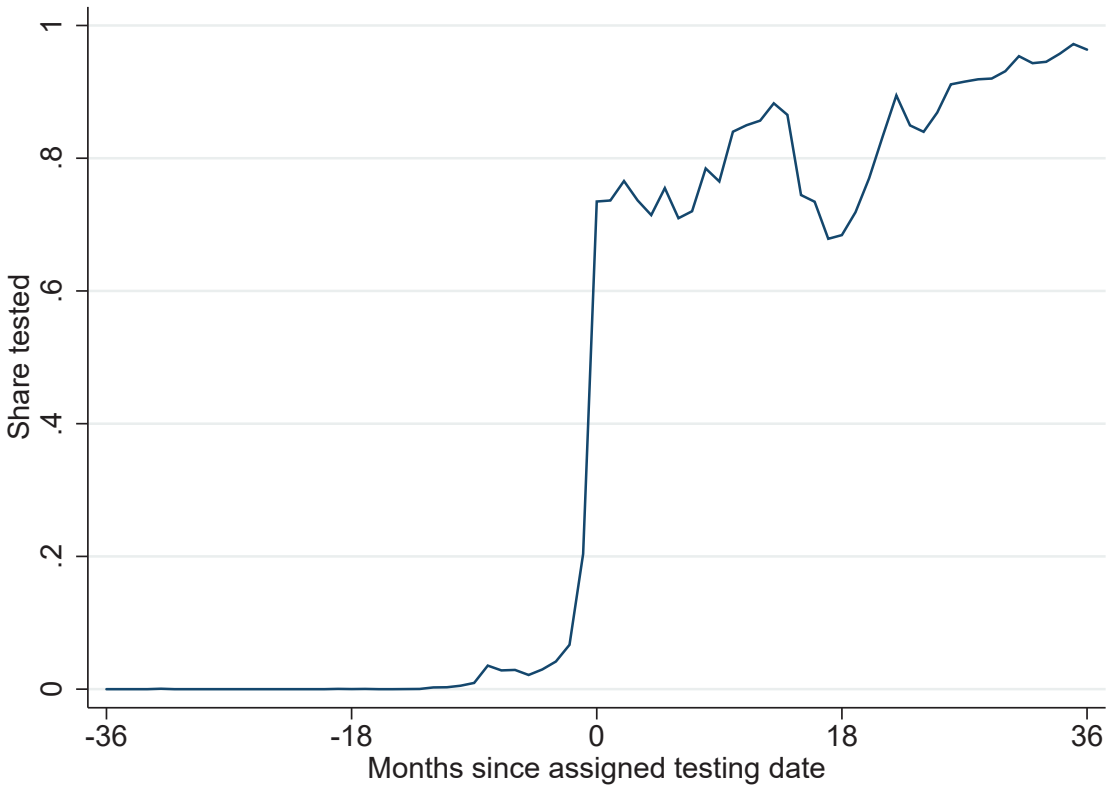
Hiring Practices Information. For several client firms, our data firm surveyed its account managers (who interact closely with the client firms regarding job testing matters), asking them to provide us with information on hiring practices once testing was adopted. The survey indicated that firms encouraged managers to hire workers with higher scores (and some firms had policies on not hiring low-scored candidates), but left substantial leeway for managers to overrule testing recommendations. Information from this survey is referenced in footnote 7 of the main text.

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

Position Controls. Position is measured in our data using the last position that a worker held when the data file was created.

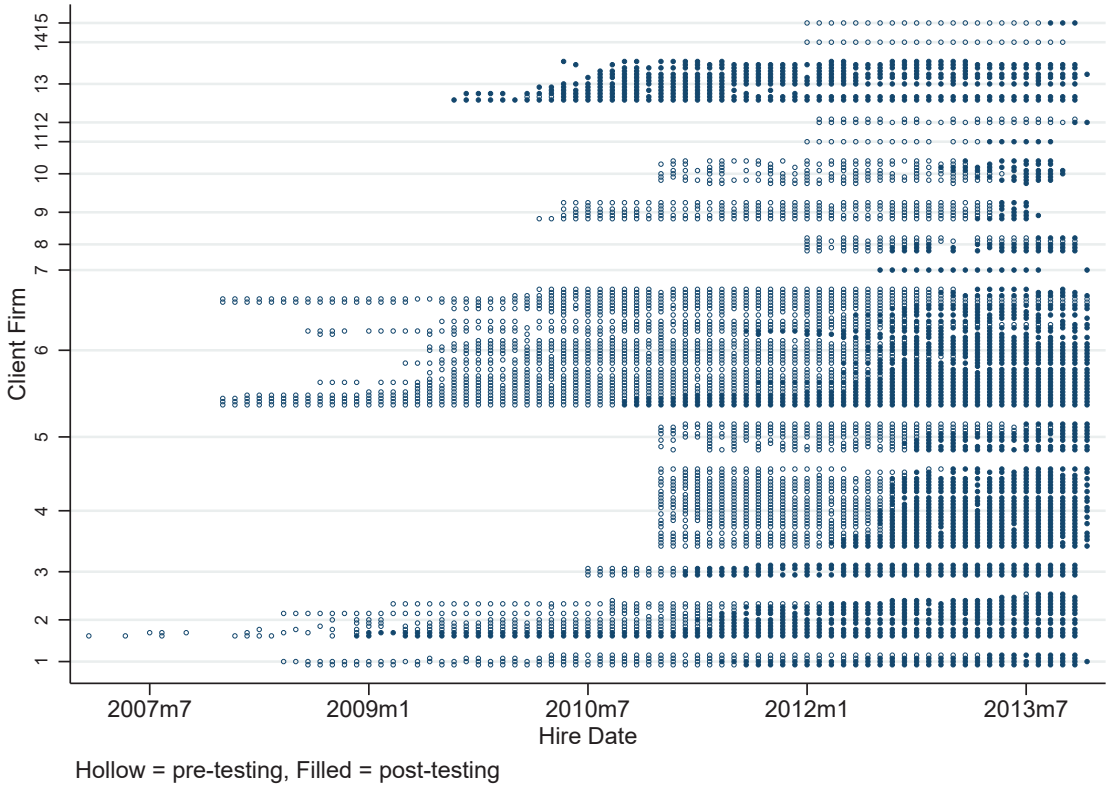
[19] This includes some locations in the pre-testing data where testing is never later introduced. Our sample is not all locations within the firms.

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



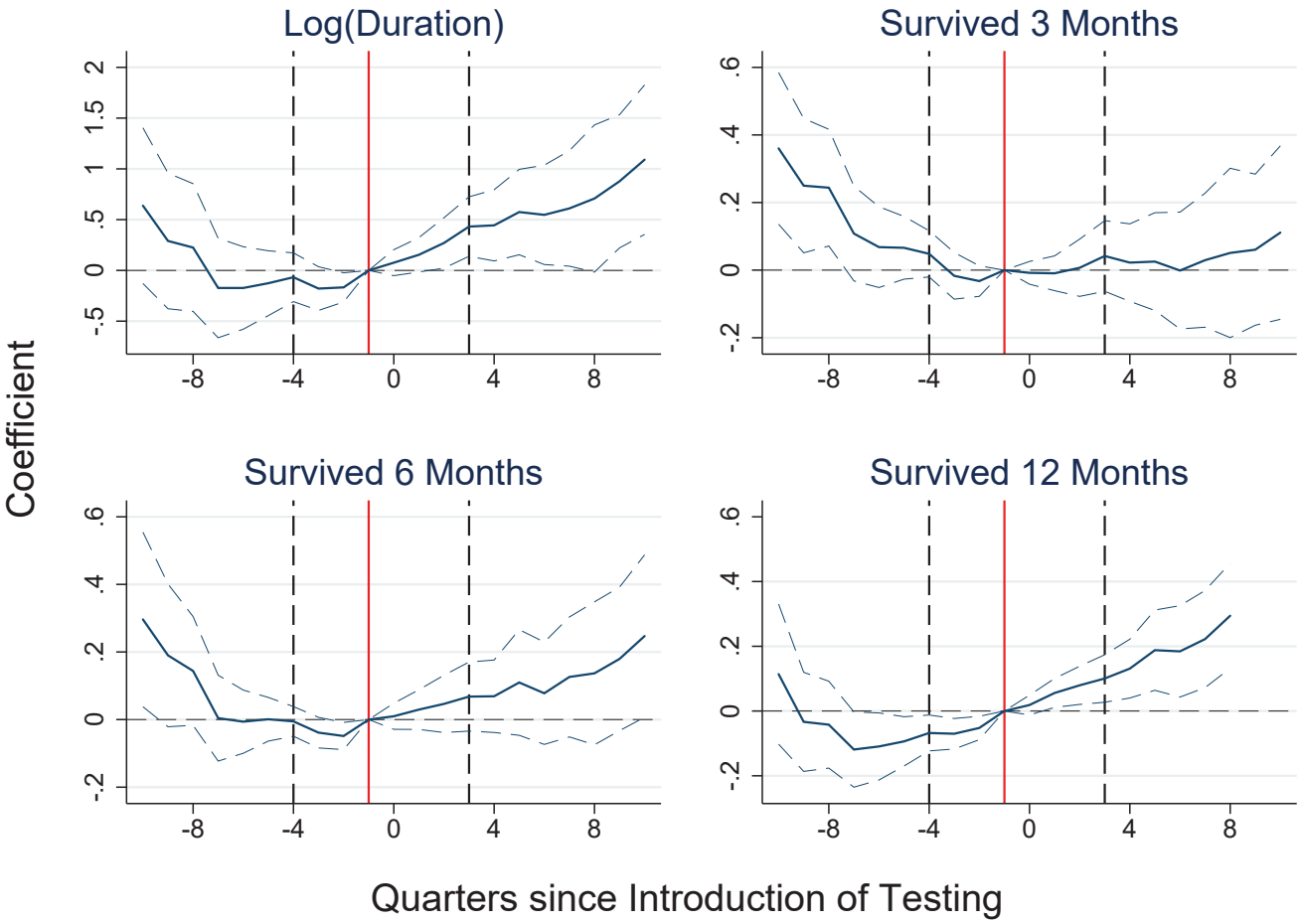
Notes: Figure A1 plots the share of hired workers with a test score as a function of time since the location-specific assigned testing date, averaged across locations. The testing date is defined at the location-month level as the first month in which the modal hire is tested. This graph is restricted to locations that receive testing. For figure clarity, we further restrict to the 89% of workers hired within 3 years of the introduction of testing.

Appendix Figure A2: Location Coverage by Date



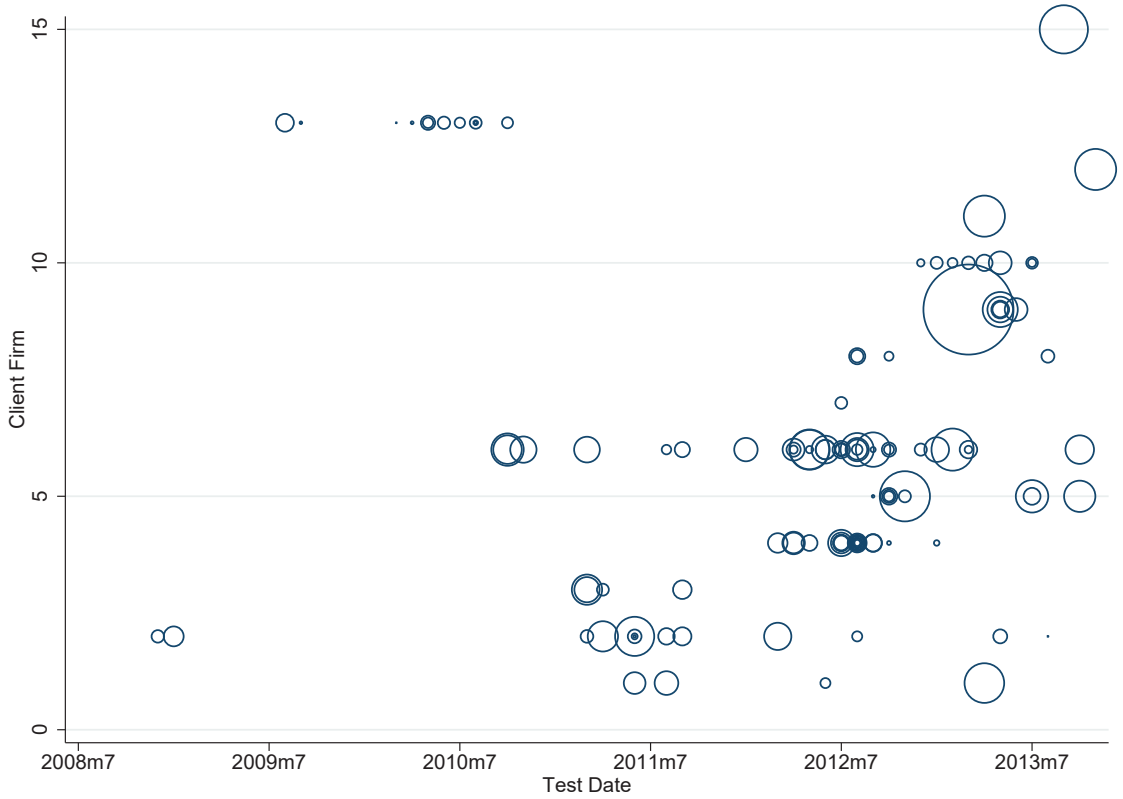
Notes: Locations are lined up on the y-axis, grouped by client firm. Dots indicate that the location hired in a given month, while a gap means no hires were made that month. Filled circles refer to periods after testing is adopted, using our definition (the modal hire was tested), while hollow circles refer to periods before testing. Dates are restricted to a 3 year window around testing adoption, covering 89% of hires. All dots are hollow for Firm 14 because it does not have a location meeting our definition of testing.

Appendix Figure A3: Event Study of Duration Outcomes, Balanced Panel



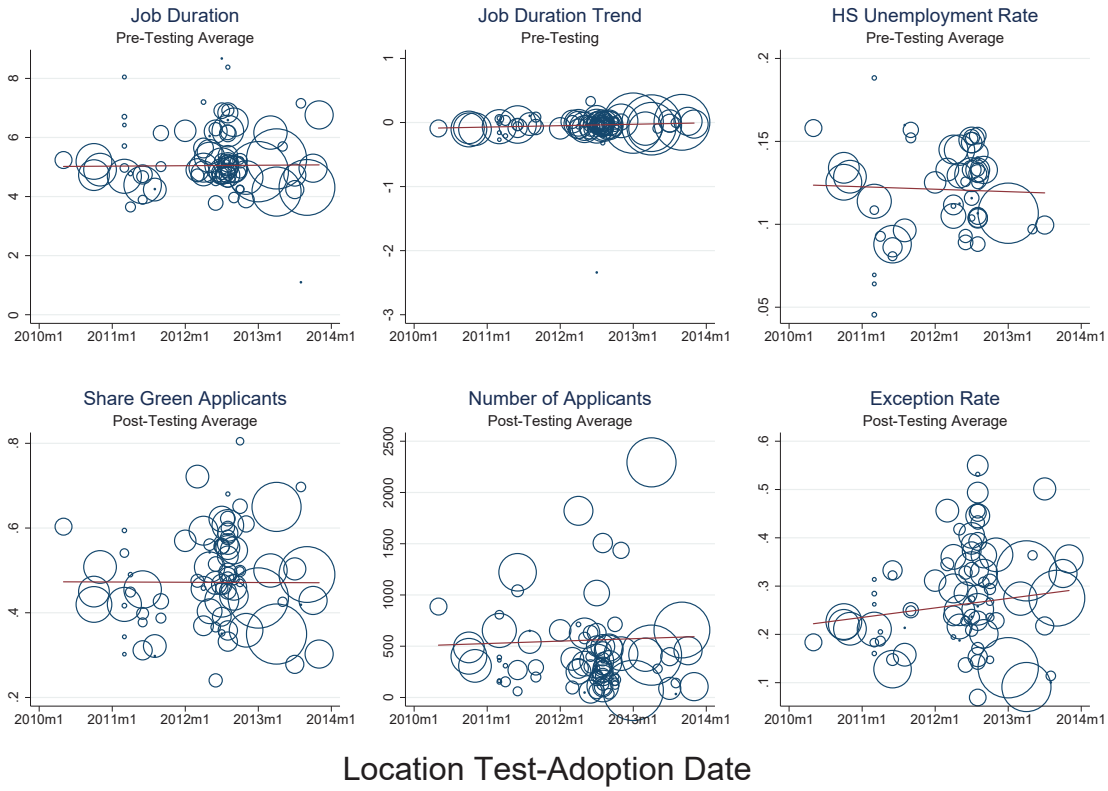
Notes: See notes to Figure II of the main text. The sample is restricted to locations with observations in each quarter from 4 lags before testing to 4 leads after (indicated with vertical dashed lines). The graph window is restricted to 10 quarters before and after testing. Dashed lines indicate the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

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



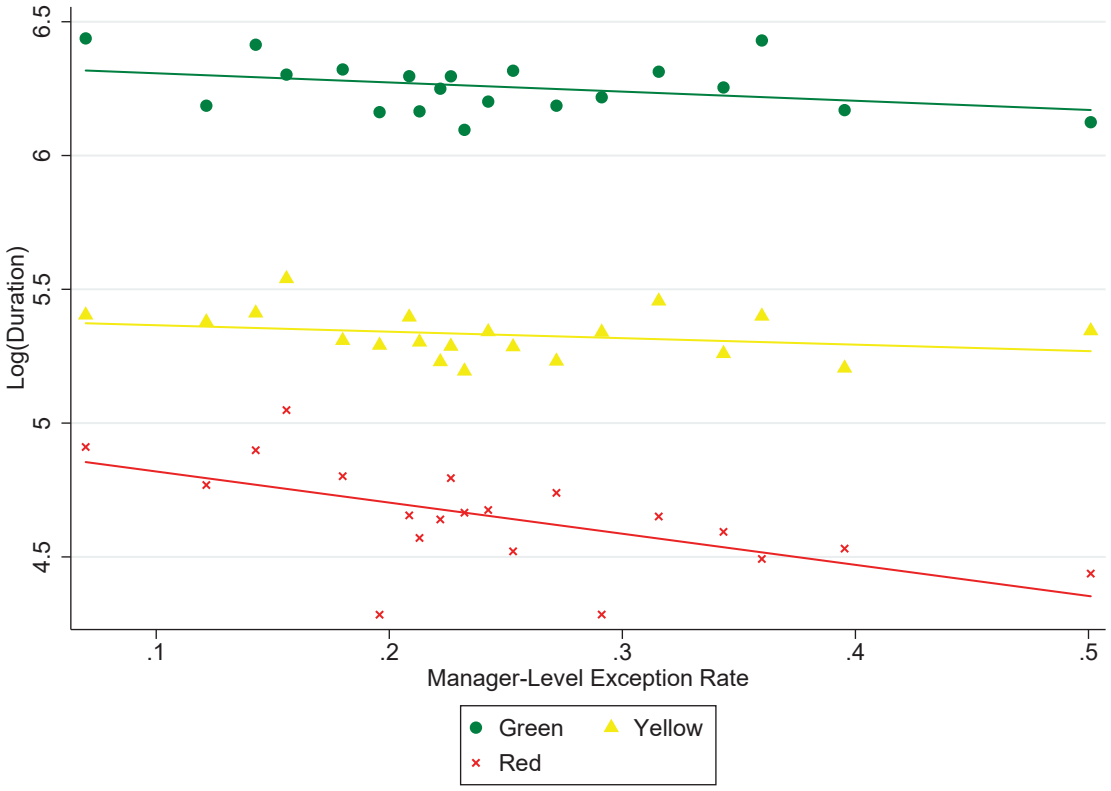
Notes: Figure A4 plots location-specific assigned testing dates on the x-axis, organized by client firm on the y-axis. Circles are weighted by location size, as defined by the number of workers currently employed in our data on July, 2013. As noted in Figure A2, Firm 14 does not appear on the graph because it does not have a location that meets our definition of testing.

Appendix Figure A5: Location Observables and Date of Testing Adoption



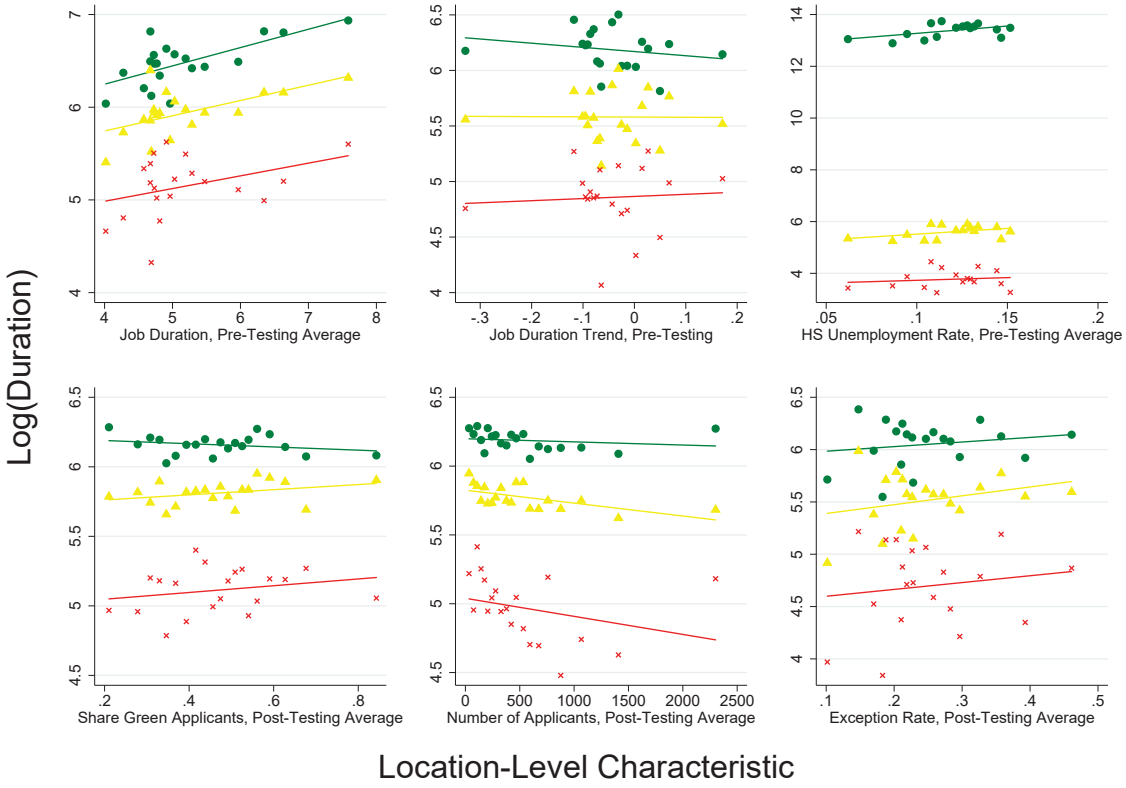
Notes: Figure A5 plots the relationship between various location-level variables (y-axis) and date of test adoption (x-axis). Circles and fitted lines are weighted by location size. In the top left panel, pre-testing durations are obtained from a censored normal regression of log durations on an exhaustive set of location fixed effects estimated on the pre-testing sample (and with no constant/intercept term included). The top middle panel plots location-specific time trends estimated from a censored normal regression of log durations on location fixed effects and location-specific time trends in the pre-testing sample (and with no constant/intercept term included). The remaining variables are raw averages at the location-level either pre- (top right) or post- (bottom panels) testing.

Appendix Figure A6: Manager-Level Exception Rates and the Color Score-Job Duration Relationship



Notes: This graph shows the relationship between color score and job duration for 20 equally sized manager-level exception rate bins. Specifically, we estimate censored normal regressions of log duration on 20 exhaustive indicators for exception rate bin and location, hire month, and position fixed effects (and with no constant/intercept term included), separately by color score. We plot the coefficients on the exception rate bins as well as the line of best fit.

Appendix Figure A7: Location Observables and The Color Score-Job Duration Relationship



Notes: This graph shows the relationship between color score and job duration for 20 equally sized bins based on the location-level characteristic specified on the x-axis. Specifically, we estimate censored normal regressions of log duration on 20 exhaustive indicators for the location characteristic bin and hire month and position fixed effects (and with no constant/intercept term included), separately by color score. (We exclude location fixed effects from these regressions because they are collinear with the location characteristics.) We plot the coefficients on the bins as well as the best linear fit.

Appendix Table A1: Robustness for Results on the Impact of Testing

	Dependent Variable: Log(Duration)					
	(1)	(2)	(3)	(4)	(5)	(6)
Impact of Testing						
Post-Testing	0.368***	0.316***	0.316***	0.296**	0.261**	0.516**
	(0.120)	*(0.121)*	*(0.119)*	*(0.150)*	*(0.117)*	*(0.245)*
Differential Impact of Testing by Exception Rates						
Post-Testing	0.366***	0.321***	0.320***	0.299**	0.294***	0.495***
	(0.119)	*(0.120)*	*(0.119)*	*(0.151)*	*(0.110)*	*(0.161)*
*Exception Rate*Post-Testing*	-0.142**	-0.151**	-0.146**	-0.127*	-0.178***	-0.444**
	(0.0652)	*(0.0696)*	*(0.0687)*	*(0.0663)*	*(0.0556)*	*(0.202)*
N	265,648	265,648	265,648	216,676	96,273	83,910
Base Controls	X	X	X	X	X	X
Testing Definition:						
Modal Worker Tested	X			X	X	X
Any Worker Tested		X				
Individual Worker Tested			X			
Location Restrictions:						
Observed Both Pre/Post Testing				X	X	
Observed in Balanced 4 Quarter Window					X	
Client Had No Pre-Sample Testing						X

*** p<0.01, ** p<0.05, * p<0.1

Notes: This table reports censored normal regressions with standard errors clustered at the location level. The top panel provides estimates of the impact of testing on log durations (see Table II), while the bottom panel estimates the differential impact of testing by location-level exception rates (see Table IV). Column 1 reproduces baseline specifications from the preceding tables. Column 2 defines the test adoption date as the first time a hire is observed with a test score at a location. Column 3 defines test adoption as whether the individual hire has a test score. Column 4 restricts to the 83 locations that are observed both before and after testing. Column 5 further restricts to locations that are observed in each of the four quarters prior and post testing. Column 6 restricts to locations that likely did not have job testing before partnering with our data firm. Base controls include location, hire month, and position fixed effects.

Appendix Table A2: Robustness to Alternative Exception Rates

Dependent Variable: Log(Duration)				
	(1)	(2)	(3)	(4)
# Exceptions Relative to Random				
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.359***	0.232***
			(0.119)	*(0.0588)*
*Exception Rate*Post-Testing*	-0.0730**	-0.0635**	-0.157**	-0.125**
	(0.0327)	*(0.0258)*	*(0.0713)*	*(0.0556)*
Exception Score Relative to Max Score				
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.356***	0.230***
			(0.120)	*(0.0656)*
*Exception Rate*Post-Testing*	-0.0237	-0.0707***	-0.190**	0.0420
	(0.0261)	*(0.0190)*	*(0.0933)*	*(0.0676)*
Exception Score Relative to Random Score				
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.353***	0.222***
			(0.117)	*(0.0609)*
*Exception Rate*Post-Testing*	-0.0585	-0.0149	-0.155**	-0.0999**
	(0.0364)	*(0.0241)*	*(0.0763)*	*(0.0498)*
N	91,319	91,319	265,648	265,648
Base Controls	X	X	X	X
Full Controls		X		X

*** p<0.01, ** p<0.05, * p<0.1

Notes: Columns 1 and 2 estimate the post-testing correlation between manager-level exception rates and log duration (see Table III). Columns 3 and 4 estimate the differential impact of testing by location-level exception rates (see Table IV). The top panel defines the exception rate as the number of order violations divided by the number of order violations under random hiring. The next panels use an exception score (1 point for yellow and 2 points for green hires) divided by the maximum possible score (middle panel) or the score under random hiring (bottom panel). Base controls include location, hire month, and position fixed effects. Full controls add client-by-year effects, local unemployment rates, and location-specific time trends (and applicant pool controls in columns 1 and 2).

B Theory Appendix

B.1 Preliminaries

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

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

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

Define $U_t \equiv E[U_i|t_i, s_i, b_i]|t_i$. Based on standard projection formulas for the Gaussian distribution, we know that $U_t \sim N((1-k)\mu_t, \Sigma)$, where $\Sigma = (1-k)^2\sigma^2 + k^2\sigma_b^2$ and $\sigma^2 = \frac{\sigma_a^4}{\sigma_a^2+\sigma_\epsilon^2}$.

Let Φ and ϕ denote the standard normal cdf and pdf, respectively. Therefore, the hire probability and $\underline{u}$ are pinned down by equation B1:

$$W = p_G(1 - \Phi(z_G)) + (1 - p_G)(1 - \Phi(z_Y)) \tag{B1}$$

where $z_t = \frac{\underline{u}-(1-k)\mu_t}{\sqrt{\Sigma}}$.

The firm's payoff under Discretion is $E[a|Hire]$, i.e., the expected quality conditional on being hired. By using the properties of the bivariate normal distribution, this can be expressed as follows, where $\lambda(\cdot)$ is the inverse Mills ratio of the standard normal.

$$\begin{aligned} W * E[a|Hire] = {} & p_G(1 - \Phi(z_G)) \left[\mu_G + \frac{(1-k)\sigma^2}{\sqrt{\Sigma}}\lambda(z_G)\right] \\ & (1 - p_G)(1 - \Phi(z_Y)) \left[\mu_Y + \frac{(1-k)\sigma^2}{\sqrt{\Sigma}}\lambda(z_Y)\right] \end{aligned} \tag{B2}$$

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 Propositions

Propositions 1 and 2, formalized below, provide intuition for our empirical analysis. Proposition 1 states that the exception rate (the probability that a Y is hired above a G applicant) is increasing in both the precision of a manager's private information and his or her bias. Proposition 2 says that the quality of hired workers, $E[a|Hire]$, is decreasing in manager bias. It also shows that absent bias, the quality of hires is increasing in the precision of a manger's private information.

Proposition 1 *The exception rate is increasing in managerial bias, k, as well as weakly increasing in 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.

Recall from above that U_t is normally distributed with mean $(1-k)\mu_t$ and variance $\Sigma = (1-k)^2\sigma^2 + k^2\sigma_b^2$. 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. Under this rescaling, 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 $\tilde{U_Y}$: $\Delta_U = \frac{(1-k)(\mu_G - \mu_Y)}{\sqrt{\Sigma}}$.

One can show (with some algebra) that $\frac{\partial \Delta U}{\partial k} = \frac{-k(\mu_G - \mu_Y)\sigma_b^2}{\Sigma^{3/2}}$, which is clearly negative for $k \neq 0$. 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, one can 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 for $k < 1$ (and $\frac{\partial \Delta U}{\partial \sigma_\epsilon^2} = 0$ for $k = 1$). 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. ■

Proposition 2 *Holding constant information, the quality of hires is decreasing in managerial bias, k. When $k = 0$, the expected quality of hires for a given manager, $E[a|Hire]$, is increasing in the precision of the manager's private information, $1/\sigma_\epsilon^2$.*

Proof The expected quality of hires, $E[a|Hire]$, is given in equation B2, above, and is a function of manager type. With some messy algebra, we obtain that:

$$\frac{\partial E\left[a|Hire\right]}{\partial k} = -\frac{k\sigma_b^2}{W\Sigma^{3/2}}\left[A\sigma^2 + \frac{k^2\sigma_b^2 B\left(\mu_G - \mu_Y\right)^2}{A\Sigma}\right] \tag{B3}$$

and

$$\frac{\partial E\left[a|Hire\right]}{\partial \sigma^2} = \frac{1-k}{2W\Sigma^{3/2}}\left[A\left((1-k)^2\sigma^2 + 2k^2\sigma_b^2\right) - \frac{(1-k)^2 k^2\sigma_b^2 B\left(\mu_G - \mu_Y\right)^2}{A\Sigma}\right] \tag{B4}$$

where $A \equiv p_G\phi\left(\tilde{z}_G\right) + \left(1-p_G\right)\phi\left(\tilde{z}_Y\right)$ and $B \equiv p_G\left(1-p_G\right)\phi\left(\tilde{z}_G\right)\phi\left(\tilde{z}_Y\right)$.

First, note that the derivative wrt k is negative; expected quality of hires is strictly decreasing in bias. Second, note that when setting $k=0$, the derivative wrt σ^2 is positive. Also, recall that σ^2 moves in the same direction as $1/\sigma_\epsilon^2$. Therefore, expected quality of hires is strictly increasing in the precision of private information, when the manager is unbiased.

We next provide intuition for these results by summarizing the logic for why these inequalities should at least weakly hold.

Since the family of normal distributions is Blackwell-ordered by precision, Blackwell's Theorem tells us that an increase in the precision of information must weakly increase the manager's utility. For $k=0$, the manager maximizes $U = E[a|Hire]$, which is a function of the precision of the manager's private information. Therefore, absent bias, the expected quality of hires is increasing in manager information.

For k, consider two managers, with bias k^H and k^L, respectively, where $k^L < k^H$. Each manager chooses a group of hires that maximizes $(1-k)E[a|Hire]+kE[b|Hire]$. Let $\left(a^H, b^H\right)$ denote the realized expectation of a and b, conditional on being hired by the manager with high bias, and let $\left(a^L, b^L\right)$ denote the same for the manager with low bias. Because $\left(a^H, b^H\right)$ and $\left(a^L, b^L\right)$ are chosen optimally, we have the following incentive compatibility (IC) constraints:

$$\begin{aligned}\left(1-k^H\right)a^H + k^H b^H &\geq \left(1-k^H\right)a^L + k^H b^L \\ \left(1-k^L\right)a^L + k^L b^L &\geq \left(1-k^L\right)a^H + k^L b^H\end{aligned}$$

We would like to prove that $a^H \leq a^L$. Suppose to the contrary that $a^H > a^L$.

First, note that it cannot be that $a^H > a^L$ *and* $b^H > b^L$ because the choice, $\left(a^L, b^L\right)$, would violate the IC of the low-bias manager.

Second, consider where $a^H > a^L$ and $b^H \leq b^L$. In this circumstance, the high-bias manager is choosing candidates with higher a and lower b than the low-bias manager. We can rearrange and sum the IC constraints to show that they imply the following:

$$\left(k^H - k^L\right)\left(\left(b^H - b^L\right) - \left(a^H - a^L\right)\right) \geq 0$$

However, this expression is false because, by assumption, we have $k^H > k^L$, $a^H > a^L$, and $b^H \leq b^L$.

Therefore, by contradiction, we have shown that $a^H \leq a^L$. ■

Discussion on Propositions 1 and 2. From Propositions 1 and 2, we observe that if high-exception managers achieve worse outcomes than low-exception managers, this must be because high-exception managers are biased or mistaken.

Formally, consider two managers, manager 1 and manager 2. The two managers have type (k_i, h_i), where k_i is bias and h_i is the precision (i.e., inverse variance) of each manager's private information, for $i \in \{1, 2\}$. The two managers have exception rates, R_i, and quality of hires, a_i, for $i \in \{1, 2\}$. We claim that if $R_1 > R_2$ and $a_1 < a_2$, then it must be that $k_1 > 0$.

To see this, suppose to the contrary that $k_1 = 0$. Then by Proposition 1, it must be that $h_1 > h_2$. That is, if manager 1 is weakly less biased than manager 2 but still has more exceptions, manager 1 must have more precise private information. Now consider a third manager with bias, $k_3 = 0$, precision of private information $h_3 = h_2$, and outcomes a_3. That is, manager 3 has no bias and the same information as manager 2. By Proposition 2, it must be that $a_1 > a_3 > a_2$.[20] But this is a contradiction.

Having discussed how Propositions 1 and 2 help frame our empirical work, we now present Proposition 3. Proposition 3 illustrates the fundamental tradeoff firms face when allocating authority: managers have private information, but they are also biased. Greater bias pushes the firm to prefer No Discretion, while better information tends to push 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 for any level of bias, there is a precision of private information small enough that firms prefer No Discretion. Uninformed managers would at best follow test recommendations and, at worst deviate because they are mistaken or biased. Finally, part 3 states that, for any fixed information precision threshold, there exists an accompanying bias threshold such that if managerial information is greater and bias is smaller, firms prefer to grant discretion. Put simply, Discretion beats out No Discretion when a manager has very precise information, but only if the manager is not too biased.

Proposition 3 *We 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 $k > k'$.*

[20] Manager 1 should have better outcomes than manager 3 because they both have no bias but manager 1 has better information. Manager 3 should have weakly better outcomes than manager 2 because they have the same information but manager 3 is unbiased (so therefore weakly less biased than 2).

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 precision 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 and low bias, worker quality is higher under Discretion than No Discretion.*

We next prove each item of Proposition 3:

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 $k = 0$, under No Discretion, the quality of hires remains equal to μ_G. Write $\rho = 1/\sigma_\epsilon^2$ for the precision of the manager's information, and write $\mu(k, \rho)$ for average quality of hire under Discretion for a manager with bias k and precision ρ. Consider a precision ρ' with $0 < \rho' < \rho$. By Blackwell's Theorem, it must be that $\mu_G \leq \mu(0, \rho')$. That is, an unbiased manager with $\rho' > 0$ will do better than hiring by the test score alone. Further, since quality of hire is increasing in precision (Proposition 2), we know that $\mu(0, \rho') < \mu(0, \rho)$. By transitivity, $\mu_G < \mu(0, \rho)$, and the firm would do better under Discretion.

Thus, Discretion is better than No Discretion for $k = 0$ and the opposite is true for $k = 1$. Proposition 2 shows that the firm's payoff is strictly decreasing in k. There must therefore be a single cutpoint, k', where, below that point, the firm's payoff for Discretion is larger 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 Fix bias $k > 0$. One can show that as $1/\sigma_\epsilon^2$ approaches 0 and therefore σ^2 approaches 0, the expected quality of hires (equation (B2)) is strictly less than μ_G. To see this, note that equation (B2) can be rearranged to obtain:

$$W * E[a|Hire] = p_G(1-\Phi(z_G))\mu_G + (1-p_G)(1-\Phi(z_Y))\mu_Y + \frac{(1-k)\sigma^2}{\sqrt{\Sigma}}\left[p_G\phi(z_G) + (1-p_G)\phi(z_Y)\right] \quad \text{(B5)}$$

As σ^2 approaches 0, the second term vanishes. This term is bounded a way from μ_G since the gap $z_G - z_Y$ approaches $\frac{(1-k)(\mu_G-\mu_Y)}{k\sigma_b}$, which is bounded away from ∞.

Thus, as precision of private information tends towards ∞, the firm will prefer No Discretion (which has a payoff of μ_G) to Discretion.

We also point out that the firm's payoff under Discretion, expressed above in equation (B2), is clearly continuous in σ (which is continuous in $1/\sigma_\epsilon^2$).

Thus, as the manager tends towards 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 precision 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 and low bias, worker quality is higher under Discretion than No Discretion.*

Proof 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). Since the firm's payoff under Discretion is continuous in both k and σ_ϵ^2 (see Equation (B2) above), Δ must also be continuous in these variables. By Proposition 2, the payoff of Discretion is strictly decreasing in the bias k, and so $\Delta(\sigma_\epsilon^2, k)$ is strictly decreasing in k. Moreover, when $k = 0$, Discretion is strictly preferable to No Discretion for $\sigma_\epsilon^2 < \infty$ (see proof to Proposition 3 part 1), so $\Delta(\sigma_\epsilon^2, k) > 0$ for $\sigma_\epsilon^2 < \infty$. Finally, Proposition 2 shows that when $k = 0$, the firm's payoff under Discretion is increasing in $\frac{1}{\sigma_\epsilon^2}$, i.e., $\Delta(\sigma_\epsilon^2, 0)$ is decreasing in σ_ϵ^2.

Fix any $\overline{\rho} \in (0,\infty)$ and let $\overline{\sigma_\epsilon^2} = 1/\overline{\rho}$. We seek to show that there exists k'' such that $\Delta(\sigma_\epsilon^2, k) > 0$ for all $\sigma_\epsilon^2 < \overline{\sigma_\epsilon^2}$ and all $k < k''$.

Let $y = \Delta(\overline{\sigma_\epsilon^2}, 0) > 0$. This is the difference in the firm's payoff for Discretion compared to No Discretion for an unbiased manager, with some minimal amount of information, $\overline{\sigma_\epsilon^2}$. We know $y > 0$ because for an unbiased manager, Discretion strictly improves upon No Discretion (see part 1 of Proposition 3).

Since $\Delta(\sigma_\epsilon^2, 0)$ is decreasing in σ_ϵ^2, it holds that $\Delta(\sigma_\epsilon^2, 0) > y$ for all $\sigma_\epsilon^2 < \overline{\sigma_\epsilon^2}$. Therefore, it suffices to show that there exists k'' such that $\Delta(\sigma_\epsilon^2, 0) - \Delta(\sigma_\epsilon^2, k) < y$ for all $k < k''$ and

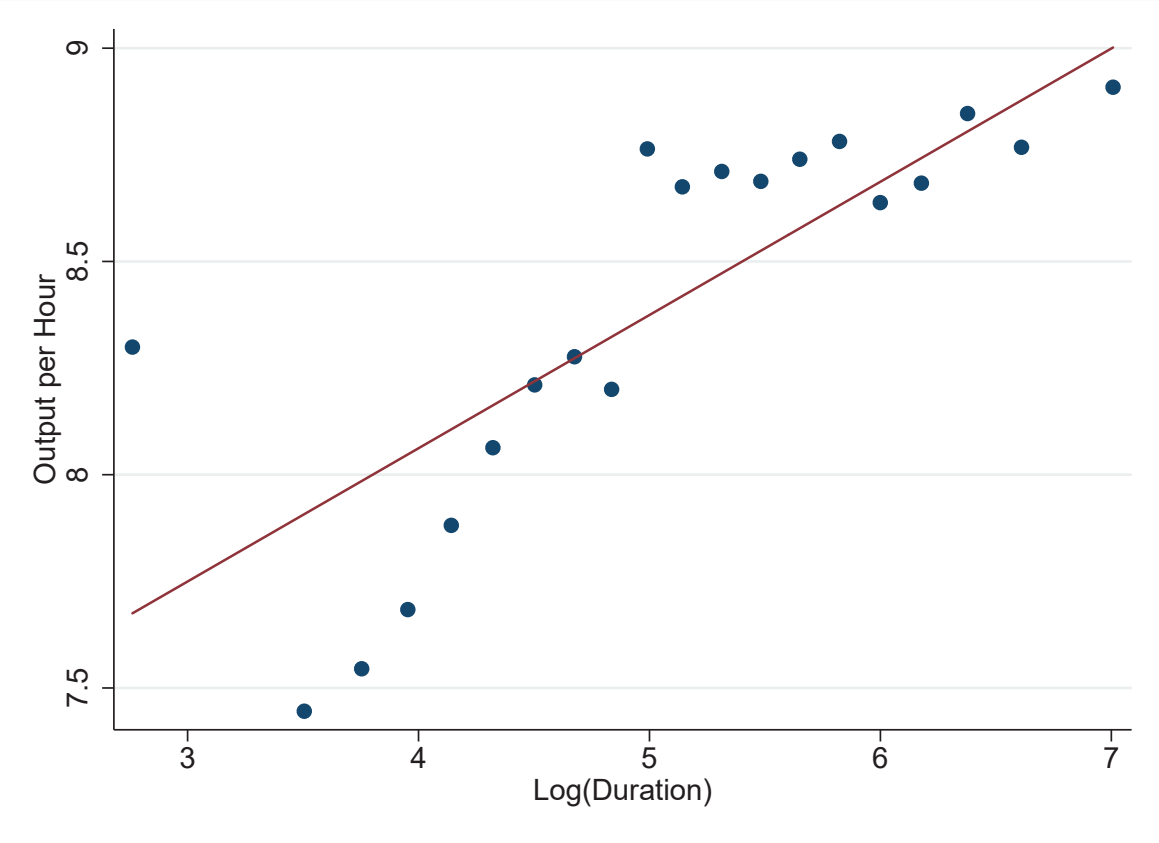
$\sigma_\epsilon^2 < \overline{\sigma_\epsilon^2}$. That is, for bias less than k'' and info more precise than the minimal case $\overline{\sigma_\epsilon^2}$, we want to show that there is some small enough bias, such that removing this bias improves the firm's payoff by only a small amount, relative to the case of no bias and minimal information, y. If this is true, then the bias amount is not enough to make the firm prefer No Discretion, because $y > 0$.

Let $d(k) = \max_{\sigma_\epsilon^2 \in [0, \overline{\sigma_\epsilon^2}]} \Delta(\sigma_\epsilon^2, 0) - \Delta(\sigma_\epsilon^2, k)$. We know $d(k)$ exists because $\Delta()$ is continuous wrt σ_ϵ^2 and the interval over which we take the maximum is compact. In words, $d(k)$ finds the largest possible improvement from eliminating bias, for bias k, across all values of information in the range.

It now suffices to show that there exists k'' such that $d(k) \leq y$ for all $k < k''$. This holds because $d(0) = 0$ (by definition) and $y > 0$, and because $d(k)$ is continuous in k (since Δ is). ■

C Supplemental Tables and Figures

Appendix Figure C1: Output per hour and Job Durations



Notes: Figure C1 plots average output per hour within 20 evenly sized bins, based on log(duration). It controls for location fixed effects to account for differences in average output per hour across locations. We use "binscatter" in Stata.

Appendix Table C1: Testing and job durations
Additional outcomes

	>3 Months (Mean=0.62; SD=0.49)		>6 Months (Mean=0.46; SD=0.50)		>12 Months (Mean=0.32; SD=0.47)	
	(1)	(2)	(3)	(4)	(5)	(6)
Introduction of Testing						
Post-Testing	0.0427*	0.0259	0.0919**	0.0597***	0.106***	0.0750***
	(0.0220)	*(0.0200)*	*(0.0371)*	*(0.0228)*	*(0.0369)*	*(0.0198)*
N	256,641	256,641	243,580	243,580	217,514	217,514
Post-Testing Correlations						
Manager Exception Rate	-0.0261***	-0.0171**	-0.0158**	-0.0101*	-0.00471	-0.0127**
	(0.00940)	*(0.00780)*	*(0.00638)*	*(0.00602)*	*(0.00496)*	*(0.00483)*
N	82,365	82,365	71,388	71,388	56,436	56,436
Differential Impact of Testing by Location-Level Exception Rate						
Post-Testing	0.0420*	0.0258	0.0912**	0.0581***	0.101***	0.0738***
	(0.0215)	*(0.0187)*	*(0.0369)*	*(0.0218)*	*(0.0350)*	*(0.0191)*
*Location Exception Rate*Post-Testing*	-0.0271*	-0.0372**	-0.0297	-0.0318*	-0.0548***	-0.0248
	(0.0160)	*(0.0173)*	*(0.0206)*	*(0.0179)*	*(0.0196)*	*(0.0172)*
N	256,641	256,641	243,580	243,580	217,514	217,514
Base Controls	X	X	X	X	X	X
Full Controls		X		X		X

*** p<0.01, ** p<0.05, * p<0.1

Notes: See notes to Tables II, III, and IV of the main text. The dependent variables are the probability that a worker survives 3, 6, or 12 months, respectively, among those who are not right-censored, i.e., those hired at least that many months before the data end date for each of the 15 firms. We use OLS regressions. Base controls include location, hire month, and position fixed effects. Full controls add client-by-year effects, local unemployment rates, and location-specific time trends. Full controls in the middle panel also include applicant pool characteristics. The top panel provides estimates of the impact of testing on job duration outcomes. The middle panel estimates the post-testing correlation between job duration and manager-level exception rates. The bottom panel estimates the differential impact of testing by location-level exception rates.

Appendix Table C2: Job Duration of Workers, by Length of Time in Applicant Pool

	Dependent Variable: Log(Duration)					
	(1)	(2)	(3)	(4)	(5)	(6)
	Green Workers		**Yellow Workers**		**Red Workers**	
Waited 1 Month	0.00545	-0.0276	-0.0271	-0.0139	-0.0338	-0.0449
	(0.0281)	*(0.0263)*	*(0.0320)*	*(0.0242)*	*(0.0622)*	*(0.0752)*
Waited 2 Months	-0.0352	-0.0714	-0.0204	-0.0542	0.00713	0.0467
	(0.0586)	*(0.0632)*	*(0.0647)*	*(0.0663)*	*(0.144)*	*(0.174)*
Waited 3 Months	0.00486	-0.0941	0.112	0.120	0.0338	0.0493
	(0.0673)	*(0.0851)*	*(0.0855)*	*(0.0867)*	*(0.220)*	*(0.242)*
N	47,809	47,809	24,496	24,496	4,098	4,098
Base Controls	X	X	X	X	X	X
Initial Applicant Pool FEs		X		X		X

*** p<0.01, ** p<0.05, * p<0.1

Notes: Regressions are restricted to the post-testing sample, adjust for censoring, and cluster standard errors at the location level. Each panel compares applicants who started working in the month they applied (omitted category) to those who started 1, 2, or 3 months later, separately by color. Panels restrict to applicant pools (location-recruiter-initial application month) with variation in wait time, and further restrict to locations and pools with at least 10 and 5 observations, respectively. Base controls are location, hire month, and position type fixed effects. Initial applicant pool fixed effects are defined by the manager-location-month for the pool when candidates first applied.

Appendix Table C3: Exception Rates and Duration Outcomes Applicant Pools with at Least as Many Green Applicants as Total Hires

	Dependent Variable: Log(Duration)			
	(1)	(2)	(3)	(4)
	Post-Testing Sample		Introduction of Testing	
Post-Testing			0.323***	0.262***
			(0.117)	*(0.0609)*
*Exception Rate*Post-Testing*	-0.112***	-0.111***	-0.243**	-0.116
	(0.0355)	*(0.0303)*	*(0.0949)*	*(0.0774)*
N	76,425	76,425	250,754	250,754
Base Controls	X	X	X	X
Full Controls		X		X

*** p<0.01, ** p<0.05, * p<0.1

Notes: Columns 1 and 2 estimate the post-testing correlation between manager-level exception rates and log duration (see Table III). Columns 3 and 4 estimate the differential impact of testing by location-level exception rates (see Table IV). Columns 1 and 2 include only hires from applicant pools with at least as many green applicants as total hires, in the post-testing sample. Columns 3 and 4 add all pre-testing observations. Base controls include location, hire month, and position fixed effects. Full controls add client-by-year effects, local unemployment rates, and location-specific time trends (and applicant pool controls in columns 1 and 2). In order to identify these controls, we must further restrict this subsample to locations that hire in at least 2 months in the post-testing period (all but 0.2% of observations).

References

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

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