Human Bias and Machine Learning: Evidence from Résumé Screening

Bo Cowgill*
Columbia University

March 26, 2016

Preliminary and incomplete.

Abstract

Should firms delegate hiring decisions to rules or algorithms? I study a field experiment in the application of machine learning to hiring inside a large firm. I find disagreement between algorithmic and status quo (human) evaluation on about 30% of job applicants. When these candidates are interviewed, the machine-favored candidates are 17% more likely to pass double-blind interview and receive a job offer offer. They are also 2σ-4σ more productive as employees. Tests of the usefulness of combining human and algorithmic judgement fare poorly for human judgement. The algorithm benefits candidates coming from an atypical career or educational backgrounds. I find evidence suggesting the candidates helped by the machine are also overlooked by other potential employers in the labor market. The machine candidates are about 12% less likely to show evidence of competing job offers during salary negotiations, and are 15% more likely to accept job offers when extended by the employer. Lastly, I find the machine candidates were evaluated as having similar general cognitive skills and human capital, but higher "intangible" characteristics such as leadership and cultural fit with the company. This results suggests there may be significant productivity benefits from automating white-collar subjective decisionmaking jobs.

1 Introduction

How large are the labor market inefficiencies introduced by human judgement? When should firms delegate decisions to rule-based algorithms? I examine these questions in the context of hiring decisions. Human evaluators may observe more abstract, contextual information about a

*The author thanks seminar participants at the 2015 Empirical Management Conference, the Institutions and Innovation Conference at Harvard Business School, the Tinbergen Institute, Universidad Carlos III de Madrid, NBER Economics of Digitization, the Kauffman Emerging Scholars Conference and Columbia Business School, as well as Jason Abaluck, Ajay Agrawal, Susan Athey, Laura Gee, Dan Gross, John Horton, Danielle Li, Christos Makridis, Stephan Meier, John Morgan and Paul Oyer.

job candidate. However, human evaluators may exhibit limited attention, idiosyncratic biases and other behavioral shortcomings in judgement. By contrast, algorithms may limit the opportunity for idiosyncratic biases and reduce variance. However, rule-based decisions may leave out critical, difficult-to-quantify contextual information necessary for optimal selection outcomes. In addition, these algorithms must be trained by humans – who may themselves embody biases.

I evaluate these questions using a series of field- and natural- experiments in selecting candidates for interviews for skilled technical positions. I start by exploiting the random assignment of job applicants to human screeners – showing wide variation in the average acceptance rates of different human screeners and in the weights assigned to different candidate characteristics. Applicants selected by more stringent screeners are *not* more likely to pass interviews and obtain a job offer.

Next I introduce a field experiment that randomly grants interviewing discretion to either a human or a simple machine learning algorithm that evaluates the text of candidate's resumes. The human and algorithm disagree on about 30% of candidates. Among those 30%, the machine's judgement is more willing to grant job offers to candidates without work experience or who come from rare educational backgrounds (for example, a school not attended by any other applicant in the firm's applicant pool).

The machine -preferred candidates pass the interview panel at about 9% higher rate than the human -selected ones, and accept job offers (when extended) offers at about a 11% higher rate. If the algorithm selects its "best" candidates in order to match the (smaller) quantity of human selected candidates, these rates are even higher. I show evidence that the machine candidates pass the interview panel in part because their worst (most negative) interview evaluation from the panel are more favorable. By contrast, the most positive evaluations from the panel are roughly similar to the human-picked candidates.

The machine's candidates are also less likely to negotiate the terms of their job offer – suggesting that they are being "missed" not only by the human evaluators at this company, but also by potentially competing employers. I also find early evidence that these candidates are also better employees once they begin their jobs (note: I currently have limited data about on-the-job performance because these outcomes are still being realized).

Limiting human discretion improves both the rate of false positives (candidates selected for interviews who fail) as well as false negatives (candidates who were denied an interview, but would have passed if selected). These benefits come exclusively through re-weighting information on the resume and not by introducing new information – or by constraining the message space available to describe candidates. For example, the mechanism in Hoffman et al. (2015) works by asking the candidate additional, human-designed survey questions for which the candidate's response is constrained by the human-designed multiple-choice survey options.

These results show evidence of productivity gains from IT adoption through a field experiment in human and machine judgement. The results suggest that the "missing data" that the computer lacks (for example, direct observations of personality traits, cultural fit with the company or other hard-to-codify variables) seem to be compensated for with a reduction of behavioral biases. Specifically, the machine can attend to a greater number of resume variables, can apply the same criteria to all candidates and evaluate candidates even if some information is "missing" – for example,

when a candidate lacks work experience or attended an unknown school. I also show experimental evidence for how a naive algorithm may help identify qualified candidates from traditionally underrepresented backgrounds – even if this was trained on historical human data containing the underrepresentation.

2 Empirical Setting

The setting from this study is a multinational organization offering products and services in several markets. In this paper, I examine the screening process for technical staff such as programmers, hardware or software engineers and software-oriented technical scientists and specialists. The sample in this paper is only for one such type of job for one particular geographic location where the firm has an office.

This organization wanted to hire as many candidates as were qualified. This firm has never able to satisfy its internal demand for workers in these roles. "Talent shortage" is a common complaint by employers regarding workers with technical skills. As such, the applicants in this paper were not competing against each other, but against the firm's hiring criteria. The economic problem of the firm was to hire qualified candidates for the job, and *not* to select the best candidates for a limited set of available openings.

The application process is typical for this industry and proceeded as follows. First, candidates apply to the company through a website. Some candidates were recruited or solicited; the applications in this study are only the unsolicited ones. Next, a human screener reviews the applications of the candidate. On average, the screener spends about two minutes per resume and makes a decision to interview the candidate or not. This paper includes a field experiment in replacing these decisions with an algorithm.

The next stage of screening is a set of bilateral interviews with a subset of the firm's incumbent workers. The first interview takes place over the phone. If this interview is successful, a series of in-person interviews are scheduled with workers from the firm. Each interview lasts about an hour on average. The interviews are mostly unstructured, with the interviewer deciding his or her own questions. The firm offers some guidance about interview content but doesn't strictly regulate the interview content (for example, by giving interviewers a script).

After the meetings, the employees who meet the candidate file a report containing impressions, a recommendation and the content of the interview discussion. These reports are first reviewed by a recruiting assistant – if the reports are positive enough, a larger committee reviews them and decides to extend an offer or not.

The candidate can then negotiate terms of the offer not. Typically, the firm engages in negotiation only in order to respond to competing job offers. The candidate eventually accepts or rejects the offer. Those who accept the offer begin working at the firm. At any time the candidate could withdraw his application if he or she accepts a job elsewhere or declines further interest. The job applications are for a particular geographic area which the candidate is aware of at the time of his/her application. The firm does not decline to pursue applications of qualified candidates on the belief that certain candidates "would never come here [even if we liked him/her]." For these jobs,

the firm believes it can offer reasonably competitive terms and does not terminate applications unless a) the candidate does not pass some aspect of screening, or b) the candidate withdraws interest.

The steps above are typical for this industry. For the analysis in this paper, I code an applicant as being interviewed if he/she passed the resume screen and was interviewed in any way (including the phone interview). I code candidates as passing the interview if they were subsequently extended a job offer.

Table 1 contains descriptive statistics and average success rates at the critical stages above. As described in the next section, the firm used a machine learning algorithm to rank candidates. Table 1 reports separate results for the "Top 1%-2%" – the subjects of the experiment in this study – and the remainder of applicants. Table 1 shows that the candidates above the machine's threshold are positively selected on a number of traits. They also tend to pass rounds of screening at much higher rates even without any intervention from the machine. One notable exception is the offer acceptance rate, which is lower for the candidate that the machine ranks highly. One possible explanation for this is that the algorithms' model is similar to the broader market's, and highly ranked candidates may attract competitive offers.

3 Design of Selection Algorithm

The firm was motivated to develop an algorithm in part because of its lower marginal costs of screening, and in part to address potential mistakes in the judgements of the human screeners. Attempting to address these problems, the firm first developed software to parse and transform digital resumes into structured data. The resulting software produced a matrix containing one row per job application with over one million covariates.

The covariates came in three broad types: One set of covariates was about the candidate's education such as institutions, degrees, majors and GPAs. The second set of covariates is about work experience including former employers and job titles. The last set of covariates contained natural language processing covariates, in particular counts of how often various unigrams and bigrams appeared in the resume. Within economics, this approach to codifying text is similar to Gentzkow and Shapiro (2010)'s codification of political speech. Demographic data are not included in the model and neither are names.

The resulting dataset contained over one million variables per job application. The firm compiled historical data from the last several years of applications, codifying each resume's covariates along with whether the candidate was ultimately offered a job or not. They used machine learning – in particular, LASSO (Tibshirani, 1996) and support vector machines (Vapnik, 1979; Cortes and Vapnik, 1995) – to weigh covariates in order to predict the success of the historical applications for this position. The firm's analysts utilized a standard set of machine learning techniques – regularization, cross-validation, separating training and test data – to select and weigh variables.[1] These techniques (and others) were ment to ensure that the weights were not overfit to the training data, and that the algorithm accurately predicted which candidates would succeed in new, non-training

[1] See Friedman et al. (2013) for a comprehensive overview of these techniques. Varian (2014) and Athey and Imbens (2015) have excellent surveys for economists.

samples.

A few observations about the algorithm. First, the algorithm introduced no new data into the decision-making process. In theory, all of the covariates described above can also be observed by human resume screeners. The human screeners could also view an extensive list of historical outcomes on candidates. In a sense, any comparisons between humans and this algorithm is inherently unfair to the machine. A human can quickly consult the Internet or a friend's advice to examine an unknown' school's reputation. The algorithm was given no method to consult outside sources or bring in new information that the human couldn't.

Second: The algorithm imposed no constraints on the message space the candidate could fill. Candidates were welcome to include any content to their resume they pleased. In practice, much of the content of resumes are deliberate attempts to persuade the reader of the candidate's quality.

These distinction is important in the context of the small, existing literature on job testing and "algorithmic" hiring. In most job testing settings, a job test introduces a combination new information and reweighed old information. In many cases, rule-based hiring not only involves a set of criteria, but also new, human-crated input variables. For example, the variables introduced in the screening algoritm studied by Hoffman et al. (2015) are responses to human-designed, multiple-choice survey instrument.

One could expect that new, human-curated information would account for a substantial amount of the benefit of the rules-based decision-making. How much benefit comes from the new information, rather than optimal weighting and "rules"? How much benefit comes from shrinking the message space to a set of human-designed multiple choice responses? How much benefit comes from the addition of human designed survey variables, separately from reanalyzing information screeners already have?

This paper provides some clues to these questions in part because the algorithm's input variables are the same (or less than) as the screener's. In both cases, the input is a text document with an enormous potential message space. The only human curation has been performed by the candidate – who acts adversarially to screening rather than in cooperation with it. The performance improvement comes entirely from reweighing pre-existing information.

Lastly: Although this algorithm is computationally sophisticated, it is econometrically naive. The programmers were not interested in interpreting the model causally. Similarly, the programmers ignored the two-stage, selected nature of the historical screening process. Candidates in the training data are first chosen for interviews and then need to pass the interviews. If historical screeners selected the wrong candidates for interviews, this would lead to biased estimates of the relationship between characteristics and success. In economics, these issues were raised in James (1979), but the programmers in this setting did not integrate these ideas into its algorithms.

They were also aware that the training method above would potentially amplify mistakes in previous screening decisions. For example: If the firm had wrongly screened out workers with electrical engineering degrees, the algorithm above may perpetuate this mistake. The programmers and their management made peace with this in part because the firm was satisfied with the existing hiring process.

4 Field Experiment

How does this algorithm perform compared to the status quo human screening process? Given that the algorithm was trained based on historical data, one might expect few disagreements between the human decision-making process and this algorithm. How often to the two signals disagree? When the signals differ, which is better?

To answer this question, the firm introduced a field experiment. In the experiment, a randomly selected subset of "machine approved" candidates were given interviews. The remaining set of machine-approved candidates were blindly examined through the status quo process – without knowing the machine's evaluation. The operationalized definition of "machine approved" was being in the top 1%-2% of candidates (according to the machine).

One random half were guaranteed interviews (treatment), and the other half (control) were left at the discretion of human screeners. One can examine the efficacy of using the machine's judgment in part by studying the outcomes of these candidates. What percentage of the control group was selected by the humans? How did the treatment group perform during the interview process? Did they perform better or worse than the human selected candidates from the same applicant pool, or from below the "top 1%"?

These research questions require the firm interview more candidates than normal – in order to examine potential disagreements between the algorithm and human. Thus the experiment required an expansion of the firm's interviewing capacity. The ≈1% threshold was selected in part because the firm's interviewing capacity could accommodate this amount of extra interviews without overly distracting employees from productive work. The top ≈1% comprised about 800 applicants over roughly one year. While this seems like a small number of candidates, this group comprised about 30% of the firm's hires from this applicant pool over the same time period.

The research questions also require that the treatment and control status of candidates remained blind to both the evaluators and the candidates themselves. Without blindness, evaluators impressions could become contaminated by their feelings towards the earlier screening method. In this setting, neither the candidates nor the interviewers were told about the existence of the experiment and were thus blind to the treatment/control status of the applicants.

In Table 2, Panel A I compare observable characteristics between treatment and control group in order to test for balance across a number of dimensions. In Panel B, I report the performance of treatment and control groups in the hiring process. The first result is substantial disagreement between machine and human judgement. The machine and humans agree on roughly 50% of candidates, and disagree on 30%. The remaining ≈20% withdrew their applications prior to the choice to interview.

Roughly 30% of candidates in the experiment were approved by the machine, but counterfactually disapproved by humans. The most common reason cited for the human rejection in this group is lack of qualifications. Separate regressions show that the machine appears more generous towards candidates without work experience and candidates coming from rare educational backgrounds.

Table 2, Panel B also shows that many of the candidates in the treatment group succeed in sub-

sequent rounds of interviews. The yield of candidates is about 8% higher in the treatment group. Table 2, Panel C assesses whether machine picked candidates are more likely to pass subsequent rounds of screening conditional on being picked. The conditional success rates are generally higher for the treatment group, but not statistically significant.

Table 3 examines these differences as regressions and adds controls, which tighten the standard errors on the differences. These results show that the average candidate in the treatment group is more likely to pass the interview than a candidate selected by a human screener from the same applicant pool. It also shows that the machine candidates are more likely to accept an offer.

In Panel C of Table 3, I examine marginal success rates of machine's candidates using instrumental variables. Here, I use the experiment as an instrument for which candidates are interviewed (or given an offer). The marginal candidate passes interviews in 37% of cases – about 17% more than the average success rate in the control group. The marginal candidate accepts a job offer extended about 87% of cases, which is about 15% higher than the average in the control group. Tests of statistical significance of theses differences are reported in the bottom of Panel C, Table 3.

In Table 5, I show that the machine candidates are are less likely to negotiate their offer terms.

In the above analysis, the machine was permitted to interview more candidates than the human. A separate question is whether the machine candidates would perform better if its capacity was constrained to equal the human's. In Table 4, I repeat the above exercise but limit the machine's quantity to match the human's. In this case, the results are sharper. The machine selected candidates improve upon the human passthrough rates.

5 Job Performance

The candidates who are hired go on to begin careers at their firm, where their career outcomes can be measured. At the time of this draft, most of these career outcomes have not been realized. For example, outcomes relating to promotion and retention often take at least one year to be realized. Future drafts of this paper may be able to study these outcomes. At present, I examine variables relating to technical productivity. The jobs in this paper involve developing software. As with many companies, this code is stored in a centralized repository (similar to http://github.com) that facilitates tracking programmer's contributions to the base of code.

This system permits reporting about each programmer's lines of code added and deleted. I use these as rudimentary productivity measures. The firm doesn't create performance incentives on these metrics, in part because it would encourage deliberately inefficient coding. The firm also uses a system of peer reviews for each new contribution of code.[2] These peer reviews cover both the logical structure and formatting or "readability" of the code as outlined in company guidelines.[3] These peer reviews and guidelines bring uniformity and quality requirements to the definitions of "lines of code" used in this study.

Despite the quality control protocols above, one may still worry about these outcome metrics.

[2]For a description of this process, see https://en.wikipedia.org/wiki/Code_review.

[3]See descriptions of these conventions at https://en.wikipedia.org/wiki/Coding_conventions and https://en.wikipedia.org/wiki/Programming_style.

Perhaps the firm would prefer fewer lines of elegant and efficient code. A great programmer should thus have fewer lines of code and perhaps delete code more often. As such, I examine both lines of code added and deleted in Table 6. These are adjusted to a per-day basis and standardized. The conclusions are qualitatively similar irrespective of using adding or deleting lines: The marginal candidate interviewed by the machine both adds and deletes more lines of code than those picked by humans from the same pool.

6 Combining Human and Machine Signals

In the "treatment" branch of the experiment, all machine-approved candidates were automatically given an interview. Before these candidates' were interviewed, they were shown to human screeners who were informed that the algorithm had suggested interviewing this candidate. The human screeners were next asked if they agreed with the machine's decision to interview. This is a similar setup to the control group, except that in the control group the machine's preference was blind.

After learning the machine's choice, the human screeners agreed on 85% of non-withdrawn applications (70% of total applications). By contrast, in the control group – where human screeners were asked for *independent* evaluations without knowing the machine's choice – the humans agreed on only 60% of non-withdrawn applications (50% of total applications).

This large difference suggests that the human screeners substantially change their minds after learning the machine's choice. The humans' propensity to agree with the algorithm speaks to how much the human screeners *themselves* place faith in their own private signals of quality. We observe this difference, even though the screeners were not told details of how the algorithm worked or about its performance.

After recording their agreement (or disagreement), the screeners were also asked to assess the treatment candidate on a 1-5 scale. In Table 7, I measure whether these human provided signals contain information using "horserace" regressions (Fair and Shiller, 1989).

I find that in isolation, the human evaluations contain some predictive information. That is, they can predict which candidates among the machine-selected candidates will successfully pass interviews. However, when both signals can be combined, nearly all weight should be placed on the machine's score of the candidates. Once the algorithm's ranking enters the regression, the human evaluation offers no additional predictive power.

Regarding candidates' acceptance of extended offers, I show in Panel B of Table 7 the human's assessment has no predictive power, even in isolation. The machine's ranking does.

References

Athey, Susan and Guido Imbens, "NBER Summer Institute 2015 Econometric Lectures: Lectures on Machine Learning," 2015.

Cortes, Corinna and Vladimir Vapnik, "Support-vector networks," *Machine learning*, 1995, *20* (3), 273–297.

Fair, Ray C and Robert J Shiller, "The informational content of ex ante forecasts," *The Review of Economics and Statistics*, 1989, pp. 325–331.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, NY: Springer-Verlag New York, 2013.

Gentzkow, Matthew and Jesse M Shapiro, "What Drives Media Slant? Evidence from US Daily Newspapers," *Econometrica*, 2010, *78* (1), 35–71.

Hoffman, Mitch, Lisa B Kahn, and Danielle Li, "Discretion in Hiring," 2015.

James, Heckman, "Sample Selection Bias as a Specification Error.," *Econometrica*, 1979.

Tibshirani, Robert, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, pp. 267–288.

Vapnik, Vladimir Naumovich, *Estimation of dependences based on empirical data [In Russian]* 1979. English Translation by Kotz, Samuel in 1982 by publisher Springer-Verlag New York.

Varian, Hal R, "Big data: New tricks for econometrics," *The Journal of Economic Perspectives*, 2014, pp. 3–27.

Tables and Figures

Table 1: Descriptive Statistics

Panel A: Characteristics

	Above Thresh	Below Thresh	Difference
Has Doctorate	0.280	0.066	0.214***
Ever Attended Elite School	0.576	0.211	0.365***
Ever Attended Top Tier School	0.315	0.339	-0.025**
Ever Attended Non-Selective School	0.035	0.122	-0.088***
Average Elite of all Schools Attended	0.478	0.158	0.320***
Referred	0.147	0.054	0.093***
No Work Experience (New Graduate)	0.189	0.182	0.007
Rare School	0.376	0.733	-0.357***

Panel B: Cumulative Acceptance Rates

	Above Thresh	Below Thresh	Difference
Interview	0.563	0.318	0.244***
Job Offer	0.112	0.007	0.105***
Accept Offer	0.080	0.006	0.074***

Panel C: Incremental Acceptance Rates

	Above Thresh	Below Thresh	Difference
Interview	0.563	0.318	0.244***
Job Offer	0.199	0.020	0.179***
Offer Accept	0.714	0.854	-0.140*

Notes: This table presents descriptive statistics of the sample of applicants. "Above the threshold" refers to candidates whom the machine estimated to be in the top 1-2% of applicants. "Below the threshold" candidates refer to the remaining candidates. Randomization in the experiment took place among candidates above the threshold.

Panel A contains applicant characteristics. In Panel A, "Above the threshold" includes the characteristics of both control and treatment candidates (Table 2 shows covariate balance between treatment and control).

Panel B shows cumulative acceptance rates (the rate of applicants who made it from the beginning to each stage). The second row of Panel B can be read to mean, "Of all applicants who applied, X% were extended an offer."

Panel C shows incremental acceptance rates (pass rates of applicants were accepted until the previous stage). The second row of Panel C can be read to mean, "Of all applicants who *were interviewed*, Y% were extended an offer."

In Panels B and C, I include only the "Above" candidates in the control group because these outcomes were affected by the experiment.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 2: Covariate Balance and Acceptance Rates (Basic Averages)

Panel A: Characteristics

	Treatment	Control	Difference
Has Doctorate	0.26	0.28	-0.02
Ever Attended Elite School	0.62	0.58	0.04
Ever Attended Top Tier School	0.28	0.31	-0.04
Ever Attended Non-Selective School	0.03	0.03	-0.00
Average Elite of all Schools Attended	0.52	0.48	0.04
Referred	0.13	0.15	-0.02
No Work Experience (New Graduate)	0.22	0.19	0.03
Rare School	0.36	0.38	-0.01

Panel B: Cumulative Acceptance Rates

	Treatment	Control	Difference
Interview	0.84	0.56	0.28***
Job Offer	0.21	0.11	0.10***
Accept Offer	0.17	0.08	0.09***

Panel C: Incremental Acceptance Rates

	Treatment	Control	Difference
Interview	0.84	0.56	0.28***
Job Offer	0.25	0.20	0.05
Offer Accept	0.80	0.71	0.09

Notes: The above comparisons are of raw means. See later tables for estimated differences of averages with controls, and for IV-based estimates of marginal effects.

Panel A presents covariate balance between treatment and control groups in the "Above the threshold" applicants.

Panel B shows cumulative acceptance rates (the rate of applicants who made it from the beginning to each stage). The second row of Panel B can be read to mean, "Of all applicants who applied, X% were extended an offer."

Panel C shows incremental acceptance rates (pass rates of applicants were accepted until the previous stage). The second row of Panel C can be read to mean, "Of all applicants who *were interviewed*, Y% were extended an offer."

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 3: Average Success Rates (Regressions w/ Controls)

Panel A: Cumulative Acceptance Rates

	Interview	Job Offer	Accept Offer
Treatment	0.28***	0.095***	0.082***
	(0.032)	(0.025)	(0.023)
R^2	0.12	0.13	0.13
Observations	770	770	770

Panel B: Incremental Acceptance Rates

	Interview	Job Offer	Accept Offer
Treatment	0.30***	0.078**	0.070
	(0.033)	(0.036)	(0.078)
R^2	0.19	0.22	0.25
Observations	770	544	124

Panel C: Marginal Success Rates using Instrumental Variables

	Job Offer	Accept Offer
Interview	0.33***	
	(0.085)	
Job Offer		0.82***
		(0.12)
F-stat (1st Stage)	2.33	25.0
Mean Outcome of Control	0.20	0.71
Difference	0.13	0.11
Observations	765	765

Notes: The above tables contain linear regressions of arriving at each stage (passing the previous) conditional on the treatment and controls. Controls include the month of the application, whether the applicant was referred and education and experience controls. Panel A shows cumulative acceptance rates (the rate of applicants who made it from the beginning to each stage). Panel B shows incremental acceptance rates (pass rates of applicants were accepted until the previous stage).

Panel C shows marginal passthrough rates of the machine's additional candidates. This differs from the average succes rates, as measured in Panel A and B and in earlier tables. In the average success rate numbers above, the "Treatment" includes both the candidates both methods selected as well as those that the machine selected but the human screeners would not have. In Panel C, I use the treatment/control status to isolate the success rates of the "marginal" candidate whom the machine liked but the human screeners would have rejected. I then compare these "marginal" success rates to the incremental success rate in the control group. The results are reported in the "Difference" row. The marginal candidates pass the interview and accept offers at a higher rate than the control group.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 4: Fixing Quantitites

Panel A: Cumulative Acceptance Rates

	Treatment	Control	Difference
Interview	0.534	0.563	-0.028
Job Offer	0.175	0.112	0.063***
Accept Offer	0.142	0.080	0.062***

Panel B: Incremental Acceptance Rates

	Treatment	Control	Difference
Interview	0.534	0.563	-0.028
Job Offer	0.327	0.199	0.128***
Accept Offer	0.812	0.714	0.097

Panel C: Incremental Acceptance Rates (Regression w/ Controls)

	Job Offer	Accept Offer
Treatment	0.14***	0.063
	(0.043)	(0.082)
R^2	0.24	0.28
Observations	422	111

Notes: The above comparisons are of raw means. The above analysis restricts the machine's discretion to match the quantity in the human-selected condition.

Panel A shows cumulative acceptance rates (the rate of applicants who made it from the beginning to each stage). The second row of Panel A can be read to mean, "Of all applicants who applied, X% were extended an offer."

Panel B shows incremental acceptance rates (pass rates of applicants were accepted until the previous stage). The second row of Panel B can be read to mean, "Of all applicants who *were interviewed*, Y% were extended an offer."

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 5: Negotiations

	Negotiated Offer	Negotiated Offer
Treatment	-0.13**	-0.12**
	(0.062)	(0.056)
Controls	No	Yes
P-value	0.037	0.041
R^2	0.051	0.12
Observations	124	124

Notes: As described in Section 2, candidates extended a job offer sometimes request improved offer terms. In the regressions above, each observation is a candidate who was extended an offer. The dependent variable is whether the candidate received an updated job offer reflecting negotiations. The outcome variable thus reflects whether the candidate was successful in improving the terms (even in minor ways), and *not* whether the candidate requested changed terms. Success in improving the terms typically happens only if the candidate can persuade the firm that he/she has competing offers. It is possible that some candidates requested changes and were denied. However, the firm's managers report that they generally try to update the offer in some way in response to a request for different terms if there are competing offers.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 6: Job Performance

	Lines of Code (Added)	Lines of Code (Deleted)
Tenure at Firm (Days)	0.37**	0.14***
	(0.15)	(0.040)
F-stat (1st Stage)	13.5	13.5
Mean Outcome of Control	-0.088	-0.12
Difference	0.46***	0.26***
Observations	761	761

Notes: This table measures the on-the-job productivity of candidates in both groups. The regressions above use instrumental variables. Each observation is a candidate. The endogenous variable is tenure length at the firm measured in days. This is zero for non-hired candidates and positive for candidates who were hired and started work. The instrumental variable is the experiment, which affects tenure by altering who is hired at all.

The outcome variable is lines of code added (and deleted). See a discussion of this variable in Section 6. The resulting coefficient on tenure can be interpreted as lines of code added per additional day of work. I normalize the outcome variable using the mean and standard deviation of everyone hired through the experiment and compare this coefficient on the average lines of code per day submitted in the human-selected control group. The results of this test are reported in the "Difference" row.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 7: Combination of Algorithmic + Human Evaluation

Panel A: Passing Interviews

	Job Offer	Job Offer	Job Offer	Job Offer
Human would-have Interviewed (binary)	0.10*		0.094*	
	(0.056)		(0.056)	
Human Evaluation (from 1-5)		0.11***		0.061
		(0.040)		(0.040)
Algorithm Score			0.81***	0.77***
			(0.15)	(0.16)
R^2	0.11	0.12	0.21	0.21
Observations	333	333	333	333

Panel B: Accepting Offers

	Accept Offer	Accept Offer	Accept Offer	Accept Offer
Human would-have Interviewed (binary)	0.089		0.054	
	(0.23)		(0.23)	
Human Evaluation (from 1-5)		0.095		0.043
		(0.065)		(0.067)
Algorithm Score			0.49***	0.47***
			(0.15)	(0.16)
R^2	0.020	0.037	0.10	0.11
Observations	82	82	82	82

Notes: This table contains "horserace" regressions (Fair and Shiller, 1989) predicting candidate outcomes from machine and human assessments. See Section 6 for discussion.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.