Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening

Bo Cowgill[*]

Columbia University

August 29, 2018

Abstract

Where should better learning technology improve decisions? I develop a formal model of decision-making in which better learning technology is complementary with experimentation. Noisy, inconsistent decision-making by humans introduces quasi-experimental variation into training datasets, which complements learning. The model makes heterogeneous predictions about when machine learning algorithms can improve human biases. These algorithms will can remove human biases exhibited in historical training data, but only if the human training decisions are sufficiently noisy; otherwise the algorithms will codify or exacerbate existing biases. I then test these predictions in a field experiment hiring workers for white-collar jobs. The introduction of machine learning technology yields candidates that are a) +14% more likely to pass interviews and receive a job offer, b) +18% more likely to accept job offers when extended, and c) 0.2σ-0.4σ more productive once hired as employees. They are also 12% less likely to show evidence of competing job offers during salary negotiations. These results were driven by candidates who were evaluated in a noisy, biased way in historical data used for training. These candidates are broadly non-traditional, particularly candidates who graduated from non-elite colleges, who lack job referrals, who lack prior experience, whose credentials are atypical and who have strong non-cognitive soft-skills.

1 Introduction

Where should better learning technology improve decisions? Many observers suggest that the better use of data in decisions will result in more empirically grounded, less-biased outcomes.

[*]The author thanks seminar participants at NBER Summer Institute (Labor), NBER Economics of Digitization, the Institutions and Innovation Conference at Harvard Business School, the Tinbergen Institute, Universidad Carlos III de Madrid, the Kauffman Emerging Scholars Conference, Columbia Business School, the Center for Advanced Study in the Behavioral Sciences (CASBS) at Stanford, the Summer Institute for Competitive Strategy at Berkeley, the Wharton People and Organizations Conference, the NYU/Stern Creativity and Innovation Seminar and the University of Chicago Advances in Field Experiments Conference, as well as Jason Abaluck, Ajay Agrawal, Susan Athey, Thomas Barrios, Wouter Diesen, Laura Gee, Dan Gross, Linnea Gandhi, John Horton, Daniel Kahneman, Danielle Li, Christos Makridis, Stephan Meier, John Morgan, Harikesh Nair, Paul Oyer, Olivier Sibony, Tim Simcoe and Jann Spiess.

However, predictive algorithms trained using historical data could codify and amplify historical bias. Scholars concerned about algorithmic bias have pointed to a number of troubling examples from judicial decision-making (Angwin et al., 2016), to hiring (Datta et al., 2015; Lambrecht and Tucker, 2016) to targeted advertising (Sweeney, 2013). Policymakers ranging from German chancellor Angela Merkel[1] to the US Equal Employment Opportunity Commission[2] have reacted with public statements and policy guidance. The European Union has adopted sweeping regulations targeting algorithmic bias.[3]

Despite these worries, counterfactual comparisons to other decision-making methods are rare. Where they exist, machine judgement appears to less biased than human judgement, even when trained on historical data (Kleinberg et al., 2017; this paper). How can algorithms trained on biased historical data ultimately decrease bias, rather than prolong it? Where in the economy will machine learning and data create better decision-making and resulting productivity benefits?

I develop a formal model of the effects of improving learning technology on decision-making and apply the model to machine learning, algorithmic bias and their connections to human decision-making. The key feature of this model is that improved learning technology is complementary with greater experimentation. In the model, human decision-makers generate a historical dataset containing biased decisions, which can arise either from taste-based discrimination, or poorly-calibrated statistical discrimination that updates slowly.[4]

The human decisions, however, are not only biased but also noisy and inconsistent. Noise in human decisions plays a critical role in how effectively new learning technology can utilize the historical record. The noisiness of human decision making provides experimental variation that is complementary for with learning technology. Noisy human judgements creates experimental variation in the training data that facilitates de-biasing, rather than codification of pre-existing bias. Without noise, new learning technology has no information about counterfactuals decisions and their outcomes.

Noisy human judgements reduce the sample selection problem driving algorithmic bias. With sufficient noise, superior learning technology can overcome not only sample selection bias, but also biases in how outcomes are graded in the training sample. Depending on the level of noise, an algorithm can either replicate historical human bias or completely correct it. The requirements for completely eliminating bias are extreme, and a more plausible scenario is that this type of algorithm will reduce, if not fully eliminate, bias.

The model suggests that settings most ripe for productivity enhancements from supervised machine learning are those in which human judges exhibit both bias and inconsistency. I then test the predictions of this model in a field experiment for hiring for white-collar, team-based jobs (computer programmers). Before the experiment, a machine learning algorithm is trained to screen candidates based on historical data about who the target firm hires and rejects. In the experiment,

[1] In October 2016, German chancellor Angela Merkel told an audience that "Algorithms, when they are not transparent, can lead to a distortion of our perception." `https://www.theguardian.com/world/2016/oct/27/angela-merkel-internet-search-engines-are-distorting-our-perception`

[2] In October 2016, the US EEOC held a symposium on the implications of "Big Data" for Equal Employment Opportunity law. `https://www.eeoc.gov/eeoc/newsroom/release/10-13-16.cfm`

[3] See the EU General Data Protection Regulation `https://www.eugdpr.org/`, adopted in 2016 and enforceable as of May 25, 2018.

[4] In my empirical section about resume screening, I show evidence suggesting the bias is statistical.

selection decisions by this algorithm and those by experienced human screeners (the status quo at the firm) are blindly evaluated in interviews and on-the-job performance (for those who are hired).

The field experiment yields four main results.

First, I find that the machine candidates outperform human screeners on nearly all measures of productivity. I find that the marginal candidate picked by the machine (but not by the human) is +14% more likely to pass a double-blind face-to-face interview with incumbent workers and receive a job offer offer, compared to candidates who the machine and human both select. These "marginal machine candidates" are also +18% more likely to accept job offers when extended by the employer, and 12% less likely to show evidence of competing job offers during salary negotiations. They are 0.2σ-0.4σ more productive once hired as employees. The increased quality of hires is achieved while increasing the volume of employees ultimately hired.

Second, I find that tests of *combining* human and algorithmic judgement fare poorly for human judgement. Regressions of interview performance and job acceptance on both human and machine assessments puts nearly all weight on the machine signal. I found no sub-group of candidates for whom human judgement is more efficient. When human screeners are informed of the machine's judgment, they elect to defer to the machine.

Why do humans perform so poorly, even as they have real-world experience and strong incentives? I find that that underperformance is mostly driven by poor calibration on a relatively small number of variables. The machine's ability to score a resume on more variables is not directly responsible. Between 70% and 90% of the productivity benefit from the algorithm's can be recovered from a low-dimensional OLS approximation.

However, I show that recovering the optimal parameters of the simpler model is not straightforward. The parameters may be impossible to recover without first creating the higher-dimensional model, and then approximating it with a simpler model. Thus, the machine's advantage in processing higher number of variables than a human can may be indirectly useful. They may help the machine learn a small number of optimal weights, even if most of the variables are ultimately can have effectively zero weight in evaluations.

Third, I find that heterogeneity in the results are driven by candidates who are broadly non-traditional. This includes candidates without a job referral, graduates from non-elite colleges, candidates with no prior work experience, candidates who did *not* have work experience from competitors. Results are also particularly strong for candidates with doctorates and/or mathematics degrees.

I examine heterogeneous effects based on the predicted likelihood that a candidate will be selected by a human screeners. The strongest effects on ultimate hiring outcomes come through candidates who were disproportionatley *unlikely* to be selected by humans. I also examine the residual of this model. Given a selection-on-observables assumption, I interpret high residual candidates as being judged noisily. I find that in the experiment, high-residual, high-noise candidates benefit much more than low-noise, small-residual candidates.[5]

When I examine interactions between bias and noise, I find that nearly all the effects come from

[5] In addition, the characteristics listed above (non-referred, non-elite) each exhibit large average residuals on their own.

high-residual candidates who with a low probability of being admitted by a human. These results provide empirical support for my theoretical model about where algorithms will have their greatest impact.

Lastly, I find that the machine advantage comes partly from selecting candidates with superior non-cognitive soft-skills such as leadership and cultural fit, and *not* from finding candidates with better cognitive skills. The computer's advantage appear precisely the soft dimensions of employee performance which some prior literature suggests that humans – and not machines – have innately superior judgement. Given the findings of psychologists about the noisiness and bias of assessments of cultural fit, this is also consistent with the theoretical model.

In the final section of the paper, I compare heterogeneous treatment effects to the "weights" inside the algorithm's model. Observers often call for an algorithm's code and numerical weights to be published as a way of evaluating the impact of algorithms. However, this could be highly misleading. Even if an algorithm (say) penalizes inexperienced candidates, it might help such candidates if the counterfactual method is worse. My comparisons show that the weights are not only different magnitudes as the treatment effects – they are also often not even the same sign.

While my setting has inherent limitations, these results show evidence of productivity gains from IT adoption . Limiting or eliminating human discretion through this form of digitization 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 – not by introducing new information (such as human-designed job-tests or survey questions) or by constraining the message space for representing candidates.

Section 2 outlines a theoretical framework for comparing human and algorithmic judgment. Section 3 discusses the empirical setting and experimental design, and section 4 describes the econometric framework for evaluating the experiment. Section 5 contains results. Section 6 concludes with discussion of some reasons labor markets may reward "soft skills" even if they can be effectively automated, and the effect of integrating machine learning into production processes.

2 Theoretical Framework

In this section, I develop a simple formal model whose goal is to help identify which settings algorithmic decision-making will improve decision-making. Although the setting is motivated by hiring, it can be applied to decision-making more generally.

This theory model builds upon an early on the economic effects of machine learning in decision-making. This theory model is related to Agrawal, Gans and Goldfarb (2017). In addition Hoffman, Kahn and Li (2016) contains a theoretical model of decisionmaking by humans and algorithms.

2.1 Setup

2.1.1 Players

The model features two players. The first is a human decision-maker, who is employed to review resumes and select candidates for a job test or interview. The second is a machine learning engineer, who takes historical data from the human's decisions and creates a predictive model of test outcomes based on the candidates' observable characteristics. This model will later be deployed on new candidates drawn from the same distribution.

Job candidates in this framework are not strategic players. Candidates come in two types, $\theta = 1$ and $\theta = 0$ in equal proportion.[6] Both have probabilities p_0 and p_1 of passing the test. Type 1 is more likely to pass ($p_1 > p_0$). After testing is completed, each tested candidate has an outcome y, a binary variable representing whether the candidate was accepted or not.

2.1.2 Utilities

The screener is paid a utility reward of $r \geq 0$ for a candidate who passes the test. We can think of r as the payoffs to correct decisions. In addition, the agent exhibits taste-based bias in favor of Type 0, and receives a taste-based payoff $b \geq 0$ for choosing Type 0.

The human judge also receives utility comes from random net utility shocks $\eta \sim F$ for picking Type 1. Suppose F is continuous, symmetric and has continuous and infinite support. F could be a normal distribution (which may be plausible based on the central limit theorem), but can assume other shapes as well.

These utility shocks add random noise and inconsistency to the screener's judgement. This formulation of noise – a utility function featuring a random component – is used in other models and settings, beginning as as early as (Marschak, 1959) and in more recent discrete choice research. The mean of F is zero – if there are average non-zero payoffs to picking either type, this would be included in the bias term b.

The noise shocks are motivated by the extensive psychology and behavioral economics literature, showing the influence of random extraneous factors in human decision-making. For example, the noise shocks may come from exogenous factors such as weather (Schwarz and Clore, 1983; Rind, 1996; Hirshleifer and Shumway, 2003; Busse et al., 2015), sports victories (Edmans et al., 2007; Card and Dahl, 2011), stock prices (Cowgill and Zitzewitz, 2008; Engelberg and Parsons, 2016) or other sources of environmental variance that affect decision-makers' mindset or mood, but are unrelated to the focal decision.[7]

At a recent NBER conference on AI and decision-making, Economics Nobel Laureate and psychologist Daniel Kahneman stated "We have too much emphasis on bias and not enough emphasis on random noise [...] most of the errors people make are better viewed as random noise [rather

[6] Proportions of candidates are not a critical part of this theory, and the conclusions of the paper do not depend on a particular proportion. For simplicity, I have used equal proportions.

[7] Note that these exogenous factors may alter the payoffs for picking both Type 0 and Type 1 candidates; F is the distribution of the *net* payoff for picking Type 1.

than bias]" (Kahneman, 2017). Kahneman has a longer popular article about the cost of noise in decision-making (Kahneman et al., 2016) and is writing a popular book about the prevalence and costs of noisy decision-making. [8]

The engineers are also strategic players in the model. The firm's management can create incentives for ML engineers to build the hiring algorithm in a variety of ways. This paper will not develop a theoretically optional incentive scheme for the ML engineers, but will instead examine how can be reduced under a particular incentive scheme (a loss function). I will examine a set of algorithms arising when engineers are asked to predict y (passing the test) from θ by approximating $E[y|\theta]$.

Because may not be the theoretically optimal incentive scheme for ML engineers, and these results should be interpreted as a lower-bound. Other incentive schemes may have better performance. However, the approach above is used in many real world settings, and requires only historical data that real-world practitioners may plausibly utilize without additional data-gathering.

ML engineers will be induced to approximate $E[y|\theta]$ under a variety of incentive schemes, for example if they are rewarded or penalized based on a symmetric loss function for y. A wide variety of ML algorithms can be used to predict $E[y|\theta]$.

The ML engineer's job in the model similar to a research econometrician's. However, the ML engineer in this setup is not required to produce a model of the human screening process that can be interpreted in light of economic theory of human decision-making. The ML engineer is simply required to is to predict outcomes for candidates in a way that's useful for his firm's selection process.

In the sections that follow, I will show conditions under which algorithms trained in the above manner and used in decision-making will reduce bias. As I mentioned, there are several ways that the algorithm above can be improved to further decrease bias. An emerging literature in computer science (Friedler and Wilson, eds, 2018) develops these improvements, although does not discuss the usefulness of noise in decreasing bias, and there are few empirical evaluations of how well these methods work compared to counterfactual methods.

2.1.3 Sequence

In the first part of the game, the screeners make choices. The historical record of this data generating process is recorded into a *training dataset*. This training dataset is then given to algorithm developers who are tasked with creating an algorithm to rank candidates for interviews. Then the data is over to engineers, who estimate a prediction model. The model is then put into production.

Note that in this setup, the human labeling process is both a) the source of training data for machine learning, as well as b) the counterfactual benchmark against which the machine learning is assessed.

[8] https://bit.ly/2o6wRG2

2.1.4 Information

Screeners in the model are able to see the θ variable (1 or 0) and the η noise realizations. The ML engineers can also see the θ variable for types, and whether they were eventually hired. However, the ML engineers do not know do not know the values p_1, p_0, q, b or the η realizations. The data-generating process does not encode the source of "noise," and thus it cannot be exploited for econometric identification in a statistical model by these engineers.

The ML engineers thus face a limited ability to infer information about candidate quality from the choice to interview. Candidates may be interviewed because of taste-based bias, because of the rewards of performance, or because of a random shock.

This paper will study an algorithm in which knowing why candidates were interviewed – or whether they were interviewed at all – is not necessary. I will assume that all the ML engineers can see is θ and an outcome variable y for each candidate. y will equal 1 if the candidate was tested and passed, and equals zero otherwise.

2.1.5 Limitations

In the next section, I analyze the equilibrium behavior for the setup above. First, I will mention a few limitations of the setup. Although the setup above may apply to many real-world settings, there are a few limitations of the model worth discussing.

First: Although human screeners are able to observe and react to the η realizations, they do not recognize them as noise and thus do not learn from the experimentation they induce. This assumption naturally fits settings featuring taste-based discrimination, as I modeled above. In Section 2.4.1, I discuss alternative microfoundations for the model, including statistical-discrimination. From the perspective of this theory, the most important feature of the screeners' bias that it is stubborn and is *not* self-correcting through learning. Insofar as agents are statistical discriminators, the experimentation is not deliberate and they do not learn from the exogenous variation generated by the noise.

Second: The human screeners and machine learning engineers do not strategically interact in the above model. For example, the human screeners do not attempt to avoid job displacement by feeding the algorithm deliberately sabotaged training data. This may happen if the screeners' direct, immediate costs and rewards from picking candidates outweigh the possible effects of displacement costs in the future of automation (perhaps because of present-bias).

In addition, there no role for "unobservables" in this model besides noise. In other words, the only variables privately observed by the human decision-maker (and not in the training data) are noise realizations η. These noise realizations are not predictive of the candidate's underlying quality, and serve only to facilitate accidental experimentation and exploration of the candidate space. By contrast, in other models (Hoffman et al., 2016), humans are able to see predictive variables that the ML algorithm cannot, and this can be the source of comparative advantage for the humans depending on how predictive the variable is.

For the theory in this paper to apply, the noise realizations η must be truly random – uncorrelated with other observed or unobserved variables as well as the final productivity outcome. If these conditions are violated, the algorithm may nonetheless have a positive effect on reducing bias. However, this would have to come about through a different mechanism than outlined in the proofs below.

2.2 Equilibrium Choices

2.2.1 Screener's Choices

A risk-neutral human screener will make the "right" decision (Type 1) if $rp_1 + \eta > rp_0 + b$. In other words, the screener makes the right decision if the random utility shocks are enough to offset the taste-based bias (b) favoring Type 0.

Let $\underline{\eta} = r(p_0 - p_1) + b$ be the minimum η necessary to offset the bias, given the other rewards involved. Such an η (or greater) happens with probability of $\Pr(\eta > r(p_0 - p_1) + b) = 1 - F(r(p_0 - p_1) + b)) = q$.

Because this paper is motivated by settings in which the training data are biased, we will restrict attention to the set of distributions F for which $q \in [0, \frac{1}{2}]$. In other words, there will be variation in how often the screener chooses the right decision, but she does never makes the right decision in a majority of cases.

The probability q of picking the right candidate changes as a function of the other parameters of this model. The partial derivatives of q are the basis for Proposition 1 and the comparative statics of the human screener selecting Type 1.

Proposition 1. *The screener's probability of picking Type 1 candidates (q), is decreasing in b, increasing in r, increasing in the quality difference in Type 1 and Type 0 ($p_1 - p_0$) and increasing in the variance of F.* Proof*: See Appendix A.1.*

Proposition 1 makes four statements that can be interpreted as follows. First: As the bias b is greater, the shock necessary to offset this bias must be larger. If F is held constant, these will be more rare.

Second: As the reward for successful decisions r increases, the human screener is equally (or more) likely to make the right decision to pick Type 1. This is because the rewards benefit from picking Type 1 will increasingly outweigh his/her taste-based bias. The ηs necessary to offset this bias are smaller and more common.

Third: Proposition 1 states that as the difference between Type 1 and Type 0 ($p_1 - p_0$) is larger, the screener is more likely to choose Type 1 despite her bias. This is because the taste-based bias against Type 1 is offset by a greater possibility of earning the reward r. The minimum η necessary for the Type 1 candidate to be hired is thus smaller and more probable.

Finally, q can be higher or lower depending on the characteristics of F, the random utility shocks function with mean of zero. For any b and r, I will refer to the *default* decision as the type the screener would choose without any noise. Given this default, F is "noisier" if increases the probability mass necessary to flip the decision from the default. This is similar to the screener "trembling" (Selten, 1975) and picking a different type than she would without noise.

Where Type 0 is the default, a *default* F will place greater probability mass above $\underline{\eta}$. This corresponds to a greater η realizations above $\underline{\eta}$ favoring Type 1 candidates. In these situations, q is increasing in the level of noise in F. For a continuous, symmetric distribution such as the normal distribution, greater variance in F places is noisier no matter what r and b, since it increases the probability of a η that flips the decision.

2.2.2 ML Engineer's Choices

As previously discuss in Section 2.1.2, this paper examines a set of algorithms in which is the engineer asked to predict y (passing the test) from θ by approximating $E[y|\theta]$. For Type 0 candidates, this converges to $(1-q)p_0$. For Type 1 candidates, this convergees to qp_1.

The ML engineers then use the algorithm to pick the type with a higher $E[Y|\theta]$. It then implements this decision consistently, without any noise. I will now compare the performance of the algorithm's selected candidate to that of the human decision process.

2.3 Effects of Shift from Human Screener to Algorithm

Proposition 2. *If screeners exhibit bias but zero noise, the algorithm will perfectly codify the humans' historical bias. The algorithm's perfomance will precisely equal that of the biased screeners and exhibit high goodness-of-fit measures on historical human decision data.* Proof*: See Appendix A.2.*

Proposition 2 formalizes a notion of algorithmic bias. In the setting above – featuring biased screeners $b > 0$ with no noise – there is no difference in the decision outcomes. The candidates approved (or rejected) by the humans would face the same outcomes in the machine learning algorithm.

The intuition behind Proposition 2 is machine learning cannot learn to improve upon the existing historical process without a source of variation and outcomes. Without a source of clean variation – exposing alternative outcomes under different choices – the algorithms will simply repeat what has happened in the past rather than improve upon it.

Because the model will perfectly replicate historical bias, it will exhibit strong goodness-of-fit measures on the training data. The problems with this algorithm will not be apparent from cross-validation, or from additional observations from the data generating process.

There are therefore no decision-making benefits to using the algorithm. However it is possible that the decision-maker receives other benefits, such as lower costs. Using an algorithm to make a decision may be cheaper than employing a human to make the same decision.

Proposition 3. *If screeners exhibit zero bias but non-zero amounts of noise, the algorithm will improve upon the performance of the screeners by removing noise. The amount of performance improvement is increasing in the amount of noise and the quality difference between Type 1 and Type 0 candidates.* Proof*: See Appendix A.3.*

Proposition 3 shows that performance improvements from the algorithm can partly come from improving consistency. Even when human decisions are not biased, noise may be a source of their poor performance. Although noise is useful in some settings for learning – which is the main theme of this paper – the noise harms performance if the decision process is already free of bias.

Proposition 4. *If biased screeners are NOT sufficiently noisy, the algorithm will codify the humans bias. The reduction in noise will actually make outcomes worse.* Proof*: See Appendix A.4.*

Proposition 4 describes a setting in which screeners are biased and noisy. This generates some observations about Type 1's superior productivity – but not enough for the algorithm to correct for the bias. In the proof for Proposition 4 in Appendix A.4, I formalize the threshold level of noise below which the algorithm is biased.

Beneath this threshold, the algorithm ends up codifying the bias, similarly to in Proposition 2 (which featured bias, but no noise). However, the adoption of machine learning actually worsens decisions in the setting of Proposition 4 (whereas it simply made no difference in the setting of Proposition 2). In a biased human regime, any amount of noise actually helps the right candidates gain employment.

The adoption of the machine learning removes this noise by implementing the decision consistently. Without sufficient experimentation in the underlying human process, this algorithm cannot correct the bias. The reduction in noise in this setting actually makes outcomes worse than if we trusted the biased, slightly noisy humans.

Proposition 5. *If screeners are biased and sufficiently noisy, the algorithm will reduce the humans' bias.* Proof*: See Appendix A.5.*

Proposition 5 shows the value of noise for debasing – one of the main results of the paper. If the level of noise is above the threshold in the previous Proposition 4, then the resulting algorithm will feature lower bias than the original screeners' data. This is because the random variation in the human process has acted as an randomized controlled trial – randomly exposing the learning algorithm to Type 1's quality, so that this productivity can be fully incorporated into the algorithm.

In this sense, experimentation and machine learning are compliments. The greater experimentation, the greater ability the machine learning to remove bias. However, this experimentation does not need to be deliberate. Random, accidental noise in decision-making is enough to induce the debiasing if the noise is a large enough influence on decision-making.

Proposition 6. *If the algorithms' human data contains non-zero bias then, "algorithmic bias" cannot be reduced to zero unless the humans in the training data were perfectly noisy (ie, picking at random).* Proof*: See Appendix A.6.*

Even if screeners are sufficiently noisy to reduce bias (as in Proposition 5), the algorithm's predictions still underestimate the advantage of Type 1 above Type 0.

In particular, the algorithm predicts a y of qp_1 for Type 1 and $(1-q)p_0$ for Type 0. The algorithm's implicit quality ratio of Type 1 over Type 0 is $qp_1/(1-q)p_0$. This is less than the quality ratio of Type 1 over Type 0 (p_1/p_0) – unless noise is maximized by increasing the variance of F until $q = 1/2$. This would make the training data perfectly representative (ie, humans were picking workers at random). Despite the reduction in bias, the algorithm will remain handicapped and exhibit some bias because of its training on biased training data.

Picking at random is extremely unlikely to appear in any real-world setting, since the purpose of most hiring is to select workers who are better than average and thus under-sample sections of the applicant pool perceived to be weaker. A complete removal of bias therefore appears infeasible from training datasets from real-world observations, particularly observations of agents who are *not* optimizing labels for *ex-post* learning.

It is possible for an algorithm to achieve a total elimination of bias without using perfectly representative training data. This may happen if a procedure manages to "guess" the a totally unbiased algorithm from some other heuristic. Some of the algorithmic innovations suggested by the emerging fairness literature may achieve this. However, in order to achieve certainty that this is algorithm is unbiased, one would need a perfectly representative training dataset (i.e., one where the screeners were picking at random).

Proposition 7. *The minimum amount of noise necessary for the machine learning to reduce bias is a* decreasing *function of the amount of bias.* Proof*: See Appendix A.7.*

Proposition 7 means that if the screeners display a large amount of bias, only a small amount of noise is necessary for the machines to correct the bias. Similarly if screeners display a small amount of bias, then high amounts of noise are necessary for the algorithm to correct the bias.

The intuition behind Proposition 7 is as follows: Suppose that screeners were highly biased against Type 1 workers, this would conceal the large productivity differences between Type 1 and Type 0 candidates. The machine learning algorithm would need to see only a few realizations – a small amount of noise – in order to reduce the bias. Because each "experiment" on Type 1 workers shows so much greater productivity, few such experiments would be necessary for the algorithm to learn the improvement.

By contrast, if the bias against Type 1 is small – large amounts of noise would be necessary for the algorithm to learn its way out of it. This is because each "experiment" yields a smaller average productivity gain. As a result, the algorithm requires more observations in order to understand the gains from picking Type 1 candidates.

A recent paper by Azevedo et al. (2018) makes a similar point about A/B testing. A company whose innovation policy is focused on large productivity innovations will need only small test of each experiment. If the experiments produce large effects, they will be detectable in small sample sizes.

Proposition 8. *In settings featuring bias sufficiently high noise, the algorithm's improvement in bias will be positive and increasing in the level of noise and bias. However, metrics of goodness-of-fit on the training data (and on additional observations from the data-generating process) have an upper bound that is low compared to settings with lower noise and/or lower bias.* Proof*: See Appendix A.8.*

The proof in Appendix A.8 comprares the algorithm's goodness-of-fit metrics on the training data in the setting of Proposition 5 (where debiasing happens) to Propositions 2 and Propositions 4, which codify bias. In the setting that facilitates debiasing, goodness-of-fit measures are not only low relative to the others, but also in absolute numbers (compared to values commonly seen in practice).

The implication of Proposition 8 is: If engineers avoid settings where models exhibit poor goodness-of-fit on the training data (and future samples), they will avoid the settings where machine learning has the greatest potential to reduce bias.

Proposition 9. *The "coefficient" or "weight" the machine learning algorithm places on $\theta = 1$ when ranking candidates does not equal the treatment effect of using the algorithm rather than human discretion for $\theta = 1$ candidates.* Proof*: See Appendix A.9.*

Proposition 9 discusses how observers should interpret the coefficients and/or weights of the machine learning algorithm. It shows that these weights may be highly misleading about the impact of the algorithm. For example: It's possible for an algorithm that places negative weight on $\theta = 1$ when ranking candidates could nonetheless have a strong positive benefit for $\theta = 1$ candidates and their selection outcomes. This would happen if the human penalized these characteristics even more than the algorithm did.

The internal weights of these algorithms are completely unrelated to which candidates benefit from the algorithm compared to a status quo alternative. The latter comparison requires a comparison to a counterfactual method of selecting candidates.

2.4 Extensions

2.4.1 Other Microfoundations for Noise and Bias

In the setup above, I model bias as taste-based discrimination, and noise coming from utility shocks within the same screener over time. However, both the noise and bias in the model can arise from different microfoundations. These do not affect conclusions of the model. I show these alternative microfoundations formally in Appendix A.10

The formulation above models the bias against Type 1 candidates as "taste-based" (Becker et al., 1957), meaning that screeners receive direct negative payoffs for selecting one type of worker. A taste-based discriminator may conscious of his/her taste-based bias (as would a self-declared racist) or unconscious (as would someone who feels worse hiring a minority, but can't say why). Either way, taste-based discrimination comes directly from the utility function.

Biased outcomes can also arise from statistical discrimination (Phelps, 1972; Arrow et al., 1973). Screeners exhibiting statistical discrimination (and no other type of bias) experience no direct utility preferences for attributes such as gender or race. "Statistical discrimination" refers to the process of making educated guesses about an unobservable candidate characteristic, such as which applicants perform well as employees. If applicants performance is (on average) even slightly correlated with observable characteristics such as gender or race, employers may be tempted to use these variables as imperfect proxies for unobservable abilities. If worker quality became easily observable, screeners exhibiting statistical discrimination would be indifferent between races or genders.

The framework in this paper can be reformulated so that the bias comes from statistical discrimination. This simply requires one additional provision: That the "educated guesses" are wrong, and are slow to update. Again, the psychology and behavioral economics literature provides ample examples of decision-makers having wrong, over-precise prior beliefs that are slow to update.

Similarly, the noise variable η can also have alternative microfoundations. The formulation beginning in Section 2 proposes that η represents time-varying noise shocks within a single screener (or set of screeners). However, η can also represent noise coming from between-screener variation. If a firm employs multiple screeners and randomly assigns applications to screeners, then noise can arise from idiosyncrasies in the each screener's tastes.

The judgment and decision-making literature contains many examples of this between-screener variation as a source of noise.[9] This literature uses "noise" to refer to within-screener and between-screener random variations interchangeably. Kahneman et al. (2016) simply writes, "We call the chance variability of judgments *noise*. It is an invisible tax on the bottom line of many companies." Similarly, the empirical economics literature has often exploited this source of random variation for causal identification.[10]

2.4.2 Additional Bias: How Outcomes are Codified

Until now, the model in this paper has featured selection bias in which lower-quality candidate joins the training data because bias. This is a realistic portrayal of many fields, where performance is accurately measured for workers in the field, but entry into the field may contain bias. For example: In jobs in finance, sales and some manual labor industries, performance can be measured objectively and accurately for workers in these jobs. However, entry into these labor markets may feature unjust discrimination.

In other settings, bias may also appear within the training data in the way outcomes are evaluated for workers who have successfully entered. For example: Suppose that every positive outcome by a Type 1 candidate is scored at only 90% as valuable as those by Type 0. In this extension, I will evaluate the model's impact when $\theta = 1$ candidates are affected both types of bias.

Let $\delta \in [0, 1]$ represent the discount that Type 1's victories are given in the training data. High δs represent strong bias in the way Type 1's outcomes are evaluated. If $\delta = 0.9$, then Type 1's victories are *codified* as only 10% as valuable as Type 0's even if they are equally valuable in an objective sense. This could happen if (say) the test evaluators were biased against Type 1, and subtracted points unfairly.[11]

In Appendix B, I provide microfoundations for δ and update the propositions above to incorporate both types of bias. Again, noise is useful for debasing in many settings (Appendix Proposition 15). The introduction of the second type of bias actually increases the usefulness of noise. However, the existence of the second type of bias also creates limitations. For a threshold level of δ,

[9] For example, this literature has shown extensive between-screener variation in valuing stocks (Slovic, 1969), evaluating real-estate (Adair et al., 1996), sentencing criminals (Anderson et al., 1999), evaluating job performance (Taylor and Wilsted, 1974), auditing financies (Colbert, 1988), examining patents (Cockburn et al., 2002) and estimating task-completion times Grimstad and Jørgensen (2007).

[10] For example, assignment of criminal cases to judges (Kling, 2006), patents applications to patent examiners (Sampat and Williams, 2014; Farre-Mensa et al., 2017), foster care cases to foster care workers Doyle Jr et al. (2007); Doyle Jr (2008), disability insurance applications to examiners Maestas et al. (2013), bankruptcy judges to individual debtors (Dobbie and Song, 2015) and corporations (Chang and Schoar, 2013) and job seekers to placement agencies Autor and Houseman (2010).

[11] As with the earlier bias in hiring (b), the evaluation bias here (δ) could itself be the result of tastes or statistical inferences about the underlying quality of work.

the algorithm under this procedure will not decrease bias and can only entrench it (Appendix Proposition 11) no matter how much noise in selection.

These conclusions assume that evaluations could be biased (δ), but these evaluations are not themselves noisy (in the same way that selection decisions were). Future research will add a parameter for noisy post-hire evaluations.

3 Empirical Setting

The job openings in this paper are technical staff such as programmers, hardware engineers and software-oriented technical scientists and specialists. Workers in this industry are involved in multi-person teams that design and implement technical products.

Successful contributions in this environment requires workers to collaborate with colleagues. In a typical project, a new product can be conceptualized as several interacting technical "modules" that function together as a coherent product. Each team member is tasked with designing and implementing a module, and ensuring that the technology of his or her module cooperates with others'. The team discusses as a group to achieve consensus on the macro-level segmentation of the product into "modules" and the assignment of various modules to teammates. Frequently circumstances arise that require these workers to switch module assignments. For example, some modules may take unexpectedly long and need to be subdivided. This resembles Deming's 2015 "trading tasks."

The internal promotion process in this market often involves peer feedback and subjective performance reviews. In fact, the incentives for positive subjective reviews from workplace peers are so strong that a number of scholars and journalists have expressed concern that these systems encourage "influence activities" (Milgrom and Roberts, 1988; Milgrom, 1988, Gubler et al., 2013) – that is, the system encourages social skills rather than programming. Eichenwald's 2012 journalistic account of Microsoft's promotion system[12] says that "[E]very employee has to impress not only his or her boss but bosses from other teams as well. And that means schmoozing and brown-nosing as many supervisors as possible."

Work in this industry thus involves substantial amounts of coordination, negotiation, persuasion and social perceptiveness – which the four skills in the the O*NET database in used by Deming (2015) to label jobs requiring social skills. This is especially true if one considers the behaviors necessary to be promoted. Consistent with this account, the occupations corresponding to this work rank above the median in the O*NET database.[13]

Separately from the underlying job details, the *hiring process itself* in white-collar work often re-

[12] http://www.vanityfair.com/business/2012/08/microsoft-lost-mojo-steve-ballmer

[13] The exact categorization of these jobs in the O*NET database requires some interpretation. Based on title alone, the most similar occupation is "Software Developers, Applications." On these four measure, is at the median for three and slightly below for the fourth (social perceptiveness). However, the job description in O*NET for this occupation leaves out the design and coordination aspects of the job. These aspects are better captured in the occupations labeled "Computer and Information Systems Managers" and "Information Technology Project Managers," both of which rate highly on all four measures of social skills. Like the jobs in this paper, the "management" expressed in these latter O*NET occupations does not necessarily involve direct command authority over subordinates and often refer to managing processes through coordination.

quires substantial coordination, negotiation, persuasiveness and social perception. Job candidates are often evaluated by a panel of interviewers who have differing needs and opinions, and whose feelings must be distilled into actionable decisions. For example: While a firm may be hiring for a role in one division, they may find another candidate who is better-matched for in a related division. Who has prioritized access to the candidate? Can a new position be created that combines both divisions? If so, what is the career path in this hybrid position, and what happens to the previous openings – does the hybrid job replace either or both earlier requisitions? Settling these questions may require discussion, persuasion and trading favors between divisions and/or members of the hiring panel.

In making an interviewing decision, a screener must put himself or herself into all of the shoes of many potentially affected parties – both those involved in the final job placement, as well as those involved in the hiring process. In addition, the screener must put himself or herself in the mind of the candidate: Will the candidate already have another job offer that he/she will like more? Will the candidate want the job after learning more details? How will the candidate react to peculiarities of pay, coworkers and procedures?

For these reasons, the O*NET occupation "human resource specialists" *also* ranks highly on all four O*NET measures of social skills (coordination, negotiation, persuasion and social perceptiveness). The actions automated in this paper are the decisions to to interview (or reject) candidates. This is only one part of the full hiring process. However, the initial decision is tied to the later outcomes through incentives: In this industry, HR specialists are awarded substantial performance incentives for selecting a candidate who passes screening. For this reason, the screeners (and their automated replacements) must be able to anticipate the social aspects of later screening and performance outcomes.

A few details inform the econometric specifications in this experiment. In this talent market, firms commonly desire as many qualified workers as it can recruit. Firms often do not have a quota of openings for these roles; insofar as they do they are never filled. "Talent shortage" is a common complaint by employers regarding workers with technical skills. The economic problem of the firms is to identify and select well-matched candidates, and *not* to select the best candidates for a limited set of openings. Applicants are thus not competing against each other, but against the hiring criteria.

The application process for jobs in this market proceeds as follows. First, candidates apply to the company through a website.[14] Next, a human screener reviews the applications of the candidate. This paper includes a field experiment in replacing these decisions with an algorithm.

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

After the meetings, the employees who met the candidate communicate the content of the interview discussion, impressions and a recommendation. During the course of this experiment,

[14] Some candidates are also recruited or solicited; the applications in this study are only the unsolicited ones.

the firm also asked interviewers to complete a survey about the candidate evaluating his or her general aptitute, cultural fit and leadership ability. With the input from this group, the employer decides to make an offer.

Next, the candidate can then negotiate terms of the offer not. Typically, employers in this market 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 any time the candidate could withdraw his application if he or she accepts a job elsewhere or declines further interest.

The setting from this study is a single firm with several products and services. The sample in this paper is only for one job opening, and for one geographic location where there is an office.[15] The hiring company 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 employer in this paper believes it can offer reasonably competitive terms; it does not terminate applications unless a) the candidate fails some aspect of screening, or b) the candidate withdraws interest.

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 14 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 14 reports separate results for candidates more than 10% likely to be offered a job – the subjects of the experiment in this study – and the remainder of applicants.

Table 14 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.

3.1 Selection Algorithm

Firms offering products and consulting in HR analytics have exploded in recent years, as a result of several trends. On the supply side of applications, several factors have caused an increase in application volumes for posted jobs throughout the economy. First, the digitization of job applications has lowered the marginal cost of applying. Second, the Great Recession caused a greater number of applicants to be looking for work. On the demand side, recent information technology improvements have enabled firms to handle the volume of online applications. Firms are motivated to adopt these algorithms in part of the volume/costs, and also because of the address potential mistakes in the judgements of human screeners.

How common is the use of algorithms for screening? The public appears to believe it is already very common. The author conducted a survey of ≈3,000 US Internet users, asking "Do you believe

[15] In this industry, candidates are typically aware of the geographic requirements upon applying.

that most large corporations in the US use computer algorithms to sort through job applications?"[16]

About two-thirds (67.5%) answered "yes."[17] Younger and more wealthy respondents were more likely to answer affirmatively, as were those in urban and suburban areas.

A 2012 *Wall Street Journal* article[18] estimates that the proportion of large companies using resume-filtering technology as "in the high 90% range," and claims "it would be very rare to find a Fortune 500 company without [this technology]."[19] The coverage of this technology is sometimes negative. The aforementioned WSJ article suggests that someone applying for a statistician job could be rejected for using the term "numeric modeler" (rather than statistician). However, the counterfactual human decisions mostly left unstudied. Recruiters' attention is necessarily limited, and human screeners are also capable of mistakes which may be more egregious than the above example. One contribution of this paper is to use exogenous variation to observe counterfactual outcomes.

The technology in this paper uses standard text-mining and machine learning techniques that are common in this industry. The first step of the process is broadly described in a 2011 LifeHacker article[20] about resume-filtering technology:[21] "[First, y]our resume is run through a parser, which removes the styling from the resume and breaks the text down into recognized words or phrases. [Second, t]he parser then sorts that content into different categories: Education, contact info, skills, and work experience."

In this setting, the predictor variables fall into four types.[22] The first set of covariates was about the candidate's education such as institutions, degrees, majors, awards and GPAs. The second set of covariates is about work experience including former employers and job titles. The third contains self-reported skill keywords that appear in the resume.

The final set of covariates were about the other keywords used in in the resume text. The keywords on the resumes were first merged together based on common linguistic stems (for example, "swimmer" and "swimming" were counted towards the "swim" stem). Then, resume covariates were created to represent how many times each stem was used on each resume.[23]

Although many of these keywords do not directly describe an educational or career accomplishment, they nonetheless have some predictive power over outcomes. For example: Resumes often use adjectives and verbs to describe the candidate's experience in ways that may indicate his or her cultural fit or leadership style. For example: Verbs such as "served" and "directed" may indicate distinct leadership styles that may fit into some companies' better than others. Such verbs would be represented in the linguistic covariates – each resume would be coded by the number of times

[16] The phrasing of this question may include both "pure" algorithmic screening techniques such as the one studied in this paper, as well as "hybrid" methods, where a human designs a multiple-choice survey instrument, and responses are weighted and aggregated by formula. An example of the latter is studied in Hoffman, Kahn and Li (2016).

[17] Responses were reweighed to match the demographics of the American Community Survey. Without the reweighing, 65% answered yes.

[18] http://www.wsj.com/articles/SB10001424052970204624204577178941034941330, accessed June 16, 2016.

[19] As with the earlier survey, this may include technological applications that differen than the one in this paper.

[20] http://lifehacker.com/5866630/how-can-i-make-sure-my-resume-gets-past-resume-robots-and-into-a-humans-hand

[21] Within economics, this approach to codifying text is similar to Gentzkow and Shapiro (2010)'s codification of political speech.

[22] Demographic data are generally not included in these models and neither are names.

[23] The same procedure was used for two-word phrases on the resumes.

it used “serve” and “direct” (along with any other word appearing in the training corpus). If the machine learning algorithm discovered a correlation between one of these words and outcomes, it would be kept in the model.

For each resume, there were millions of such linguistic variables. Most were excluded by the variable selection process described below. The training data for this algorithm contained historical resumes from previous four years of applications for this position. The final training data dataset contained over one million explanatory variables per job application and several hundred thousand successful (and unsuccessful) job applications.

The algorithms used in this experiment machine learning methods – in particular, LASSO (Tibshirani, 1996) and support vector machines (Vapnik, 1979; Cortes and Vapnik, 1995) – to weigh covariates in order to predict success of the historical applications for this position. Applications were coded as successful if the candidate was extended an offer. A standard set of machine learning techniques – regularization, cross-validation, separating training and test data – were used to select and weigh variables.[24] 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 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: This modeling approach imposes no constraints on the job applicant’s message space. The candidate can fill the content of her resume with whatever words she chooses. The candidate’s experience was unchanged by the algorithm and his/her actions were not required to be different than the status quo human process. As with a spoken conversation with a hiring manager, this algorithm did not impose constraints on what mix of information, persuasion, framing and presentation a candidate could use in her presentation of self.

By contrast, other “automated” job screening interventions drastically limit the candidate’s message space. For example, the variables introduced in the screening algoritm studied by Hoffman, Kahn and Li (2016) are responses to human-designed, multiple-choice survey instrument which are given weights by an algorithm. Hoffman, Kahn and Li (2016) provide convincing evidence that these tests can be very valuable to the employer. However they speak to a different research question than the subject of this paper for several reasons.

In these surveys, the survey questions are designed by humans. Additional information is available to the algorithm exists because a human – not a machine – decided to solicit this information from the candidate. A large part of the benefit an algorithm in this context may come from the fact that an experienced organizational psychologist knew to add a particular question to the survey. The success or failure of these applications may owe more to insightful human survey design than

[24] See Friedman et al. (2013) for a comprehensive overview of these techniques. Athey and Imbens (2015) has an excellent surveys for economists.

machine social skills.

The multiple-choice format vastly constrains the candidate's message space. This contrasts with normal human interaction. The communication style evolved by humans over millions of years of evolution is not constrained by multiple choice answers. This feature destroys the analogy to human social skills and makes the work of the computer much easier. In addition: The constrained format of the responses are, again, *also* designed by a human survey designer, like the questions themselves. The benefit of reduced message space should also be attributed to deliberate human design.

Third: The counterfactual human recruiters in these studies do not have the benefit of additional information nor the simplified message space. The "treatment" is a combination of new information, reduced message space and reweighing of information – of which only the latter was provided entirely by a machine. By contrast, the application in this paper provides a much cleaner comparison of human and machine judgment based on common inputs. For both sides of the experiment, 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 from digitization in this context comes entirely from reweighing information that humans are able to see.

Lastly: Although the algorithm in this paper is computationally sophisticated, it is econometrically naive. The designers were not interested in interpreting the model causally. Similarly, the algorithm designers 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 Heckman (1979), but the programmers in this setting did not integrate these ideas into its algorithms.

4 Potential Outcomes Framework for Screening Experiment

How does one the effectiveness of one screening method (such as machine learning) compare to another (such as human evaluation)? As Oyer and Schaefer (2011) discuss, field experiments varying hiring criteria are relatively rare ("What manager, after all, would allow an academic economist to experiment with the firms screening, interviewing or hiring decisions?"). In this section, I lay out some simple econometrics for causal inferences about hiring criteria. I present a stylized potential outcomes framework similar to Neyman (1923/1990); Rubin (1974, 2005), and apply this framework to my setting to obtain causal estimates.

Many firms screen job candidates using a test such as a job interview or a skills assessment. Candidates in these settings face multiple stages of screening: They must be selected for a test, and then pass the test in order to become employed.[25] Because testing is expensive, the firm must target testing to candidates most likely to pass. The econometric setup below helps measure the effects of changing criteria for testing. This can be used by firms and researchers to study tradeoffs

[25] In many settings, remaining employed or earning promotions or raises requires a similar process.

between the quantity and average yield of testing criteria.[26] This procedure takes the test as given, and evaluates various testing criteria to determine the optimal screen.[27]

For the exposition below, I will use generic testing language wherever possible and use hiring examples for clarification. The method can be applied in non-hiring settings featuring similar testing and selection tradeoffs.[28]

First, I introduce some notation. Each observation is "candidate" for testing, indexed by i. From the employer's perspective, the relevant counterfactuals for i are 1) hiring candidate i, 2) hiring someone else or 3) leaving the position unfilled. Because this paper is about employers' selection strategy, the statistics below will focus on the potential outcomes from the employer's perspective. I compare the firm's outcomes from one screening criteria vs another.[29]

Each candidate applying to the employer has a true, underlying "type" of $\theta_i \in \{0, 1\}$, representing whether i can pass the test if administered. The potential outcomes for any candidate are $Y_i = 1$ (passed the test) or $Y_i = 0$ (did not pass the test, possibly because the test was not given). Because this empirical strategy is oriented around the firm's strategy, candidates outcomes are coded as zero for candidates are rejected or work elsewhere.[30]

θ represents a generic measure of match quality from the employer's perspective. It may reflect both vertical and horizontal measures of quality. The tests in question may evaluate a candidate in a highly firm-specific manner (Jovanovic, 1979). Y reflects the performance of the candidate on a single firm's private evaluation, which may not necessarily be correlated with the wider labor market's assessment.[31]

For each candidate i, the econometrician observes either $Y_i|T = 1$ (whether the test was passed if it occurred) or $Y_i|T = 0$ (whether the test was passed if it didn't occur, which is zero). The missing or unobserved variable is how an untested candidate would have performed on the test, if it had been given. No assumptions about the distribution of θ are required.

[26] For example, some firms may want to use a criteria that maximizes average test yields, conditional on a fixed budget of N tests. Alternatively, other firms may prefer to relax the N budget, and instead maximize the sum of total test yield – possibly at the expense of the average yield. In either case, the procedure below helps quantify the tradeoffs between testing criteria.

[27] I do not address whether the test itself is optimal.

[28] For example, doctors may want to administer costly tests – but only to patients who are likely to have a particular illness (Abaluck et al., 2016). Alternatively, police may want to spend investigative resources to evaluate ("test") allegations of criminality, but only in cases likely to uncover actual crime. College admissions officers may want to offer interviews, but only for applicants likely to pass (or likely to accept offers if extended). Venture capitalists may want to interview companies, but only those most likely to succeed. As I discuss later, the test itself is a form of "criteria" for employment. The procedure described here can be iteratively applied up and down the production function to select an optional hiring and/or promotion criteria (rather than testing criteria) from among several discrete alternatives.

[29] One could also look at the effects of varying screening criteria on candidates' eventual outcomes, including outcomes later in life. This paper is focused on the firms' outcomes, i.e. whether the firm hires more productive workers when using one screening mechanism vs another.

[30] Binary outcomes is used to simplify exposition. In addition, many empirical outcomes in this paper are binary, such as whether candidates passed an interview or were extended a job offer (or not). Non-binary outcomes can be used as well, which I show in a later section.

[31] It is possible that the candidate applied and/or took another test through a different employer, possibly with a different outcome. These outcomes are not used in this procedure for two reasons. First, firms typically cannot access data about evaluations by other companies. Second: Even if they could, the other firm's evaluation may not be correlated with the focal firm's.

Suppose we want to compare the effects of adopting a new testing criteria, called Criteria B, against a status quo testing criteria called Criteria A.[32] For any given candidate, $A_i = 1$ means that Criteria A suggests testing candidate i and $A_i = 0$ means Criteria A suggests *not* testing i (and similarly for $B = 1$ and $B = 0$). I will refer to $A = 1$ candidates as "A candidates" and $B = 1$ candidates as "B candidates." I'll refer to $A = 1$ & $B = 0$ candidates as "$A \setminus B$ candidates," and $A = 1$ & $B = 1$ as "$A \cap B$ candidates." The Venn diagram in Figure 1 may provide a useful visualization.

For many candidates, Criterion A and B will agree. As such, the most informative observations in the data for comparing A and B are where they disagree. If the researcher's data contains A and B labels for all candidates, it would suffice to test randomly selected candidates in $B \setminus A$ and $A \setminus B$ and compare the outcomes. Candidates who are rejected (or accepted) by both methods are irrelevant for determining which strategy is better.[33]

However, often researchers do not know the full extent of disagreement between A and B. The act of selecting a B candidate (say, by scheduling an interview), may pre-empt evaluation by A, making the candidate unavailable for A's assessment by removing him from the candidate pool. I propose a strategy for addressing this problem below using an instrument (such as a field experiment) for causal inference.

The framework proceeds in two steps. First, I estimate the test success rate of $B \setminus A$ candidates – that is, candidates who would be hired *if and only if* Criteria B were being used and who would be rejected if A were used.

Next, I will then compare the above estimate to the success rate of $A \cap B$ candidates (candidates that both criteria approve), for all A candidates and for $A \setminus B$ candidates (ones that A approves and B doesn't). Then I will compare these test rates to make an inferences the effects of using A vs B.

To estimate $E[Y|T = 1, A = 0, B = 1]$ (outcomes of candidates who would be rejected by Criteria A, but tested by Criteria B) one cannot simply test all B candidates or a random sample of them. Some of the B candidates are also A candidates. The econometrician needs an instrument, Z_i, for decisions to test that is uncorrelated with A_i. Because the status quo selects only A candidates, the effect of the instrument is to select candidates who would otherwise not be tested.

For exposition, suppose the instrument Z_i is a binary variable at the candidate level. It varies randomly between one and zero with probability 0.5, for all candidates for whom $B_i = 1$. The instrument must affect who is interviewed – for exposition, assume that firm tests all candidates for whom $Z_i = 1$, irrespective of A_i.[34]

In order to measure the marginal yield of Criteria B, we need variation in Z_i within $B_i = 1$.[35] The instrument Z_i within $B_i = 1$ is "local" in that that it only varies for candidates approved by Criteria

[32] Criterion A and B can be a "black box" – I will not be relying on the details of how either criteria are constructed as part of the empirical strategy. In this paper, A is human discretion and B is machine learning. However, A could also be "the CEO's opinion" and B could be "the Director of HR's opinion." One Criteria could be "the status quo," which may represent the combination of methods currently used in a given firm.

[33] Unless there is a SUTVA-violating interaction between candidates in testing outcomes.

[34] These characteristics are true for this paper, but these assumptions can be relaxed to be more general.

[35] Additional random variation in Z_i beyond $B = 1$ is not problematic, but isn't necessary for identifying $E[Y|T = 1, A = 0, B = 1]$. Z_i can be constant everywhere $B = 0$.

B. Z_i identifies a local average testing yield for Criteria B.

We can now think of all candidates as being in one of four types: a) "Always tested" – these are candidates for whom $T_i = 1$ irrespective of whether Criterion A or B are used ($A_i = B_i = 1$), b) "Never tested," for which $T_i = 0$ irrespective of Criteria A or B ($A_i = B_i = 0$). The instrument does not effect whether these two groups are treated. Next, we have c) "Z-compliers," who are tested only if $Z_i = 1$, and d) "Z-defiers," who are tested only if $Z_i = 0$.

Identification of this "local average testing yield" requires the typical five IV conditions. I outline each condition in theory in Appendix C, with some interpretation of these assumptions in a hiring setting. In the following section (4.1), I show that each condition is met for my empirical setting.

Under these assumptions, we can estimate the average yield of $A = 0$ & $B = 1$ candidates as:

$$E[Y|T = 1, A = 0, B = 1] = \frac{E[Y_i|Z_i = 1, B_i = 1] - E[Y_i|Z_i = 0, B_i = 1]}{E[T_i|Z_i = 1, B_i = 1] - E[T_i|Z_i = 0, B_i = 1]} \tag{1}$$

The value above can be estimated through two-stage least-squares in a procedure akin to instrumental variables (Angrist et al., 1996). The outcome "caused" by the test is the *revelation* of θs for the tested candidates, so that the firm can act on the revealed information by extending offers. Importantly, the test itself does not cause θ to change for any candidate.

The resulting estimand is a "marginal success rate" of the candidates tested by B but not A. This estimand has units of "*new successful tests* over *new administered tests*."[36]

Next, I show how the IV conditions are met in my empirical setting. Then, I show how to extend this framework further into the production function to measure the effects on other downstream outcomes beyond early stage testing acceptance.

4.1 Application to my Empirical Setting via Field Experiment

In my empirical setting, all incoming applications (about 40K candidates) were scored and ranked by the algorithm. Candidates with an estimated probability of 10% (or greater) of getting a job offer were flagged as "machine approved."[37] This group comprised about 800 applicants over roughly one year. [38]

The field experiment worked by generating a random binary variable Z for all machine-picked candidates (one or zero with 50% probability). Candidates who draw a one are automatically

[36] β_{2SLS} is the ratio of the "reduced form" coefficient to the "first stage" coefficient. In this setup, the "reduced form" comes from a regression of Y on Z, and the "first stage" comes from a regression of T on Z. Applied in this setting, the numerator measures new successful tests caused by the instrument, and the denominator estimates new administered tests caused by the instrument. The ratio is thus the marginal success rate – new successful tests per new tests taken.

[37] The threshold of 10% was chosen in this experiment for capacity reasons. The experiment required the firm to spend more resources on interviewing in order to examine counterfactual outcomes in disagreements between the algorithm and human. Thus the experiment required an expansion of the firm's interviewing capacity. The ≈10% 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.

[38] 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.

given an interview. Those who drew a zero – along with the non-machine approved candidates – are blindly left to be judged by the status quo human process.

The human evaluators were thus given access to a random half of machine-approved candidates, so that they could be independently evaluated along with those rejected by the screening algorithm. Importantly, the humans were not told how the machine evaluated each candidate – they were not told about the existence of the machine screening and had no choice than to evaluate the candidates independently.

The random binary variable Z acts as an instrument for interviewing that can be used with the potential outcomes framework above. Candidates selected for an interview (from either method) were sent blindly into an interview process. Neither the interviewers nor the candidates were told about the experiment or which candidates (if anyone) came from which selection process.

The IV conditions in Appendix C are met as follows:

1. **SUTVA**: SUTVA would be violated if the treatment group's outcomes interact with the control group's. This would be problematic if firms had an inelastic quota of hiring slots. In my empirical setting and many others, the employer's policy is to make an offer to anyone who passes the test. "Passing" depends on performance on the test relative to an objective standard, and not by a relative comparison between candidates on a "curve."[39]

2. **Ignorable assignment of Z**. Covariate balance tests in Table 2 appear to validate the randomization.

3. **Exclusion restriction**, or $Y(Z,T) = Y(Z',T), \forall Z, Z', T$. In my empirical setting, the instrument is a randomized binary variable Z_i. This variable was hidden from subsequent screeners. Graders of the test did not know which candidates were approved (or disapproved) by Criteria A or B, or which candidates (if any) were affected by an instrument. The existence of the experiment and instrument were never disclosed to test graders or candidates – the evaluation by interviewers was double-blind.

4. **Inclusion restriction**. The instrument must have a non-zero effect on who is tested. In my empirical setting, this is clearly met. The experiment strongly affected who was interviewed. B candidates were +30% more likely to be interviewed if when $Z_i = 1$.

5. **Monotonicity**. The instrument here was used to guarantee certain candidates an interview, and not to deny anyone an interview (or make one less likely to be interviewed).

The econometric setup above does not require that the two methods test the same *quantity* of candidates. This is a useful feature that makes the approach more generic: Many changes in testing or hiring policy may involve tradeoffs between the quantity and quality of examined candidates.

In my empirical setting, the machine learning algorithm identified 800 candidates, and the human screeners identified a larger number (XXX). It's possible that the higher success rate is the

[39] This policy is common in many industries where hiring constraints are not binding – for example, when there are few qualified workers, or workers who are interested in joining the firm, compared to openings. As Lazear et al. (2016) discuss, much classical economic theory does not model employers face an inelastic quota of "slots." Instead it models employers featuring a continuous production function where tradeoffs are feasible between worker quantity, quality and cost.

result of extending offers to fewer, higher quality people. To address this, I will compare the outcomes of machine-only candidates not only to the average human-only candidate, but also to the average candidate selected by both mechanisms (of which there were much fewer). Then, I will fix the quantities of interviews available to both mechanisms to measure differences in yield, conditional on an identical "budget" of interviews.

4.2 Offer accepts, on-the-job productivity and other "downstream" outcomes

In some cases, firms may care not only about downstream outcomes after the job test or interview. For example: They may care about who accepts extended job offeres, or who performs well as an employee after testing and hiring. It's possible that a new interviewing criteria identifies candidates who pass, but do *not* accept offers (perhaps because many other firms have simultaneously recruited these candidates). The framework above can be extended to measure how these outcomes are affected by changes to screening.

For these empirical questions, a research can use a different Y (the outcome variable measuring test success). Suppose that $Y_i' = 1$ if the candidate was tested, passed *and* accepted the offer. This differs from the original Y, which only measures if the test was passed. Using this new variable, the same 2SLS procedure can be used to measure the effects of changing Criteria A to B on offer-acceptance or other downstream outcomes. Such a change would estimate a local average testing yield whose units are *new accepted offers / new tests*, rather than *new tests passed (offers extended) / new tests*.

In some cases, a researcher may want to estimate the offer acceptance rate, whose units are "offer accepts" / "offer extends." The same procedure can be used for this estimation as well, with an additional modification. In addition changing Y to Y', the researcher would also have to change the endogenous variable T to T' (where $T' = 1$ refers to being extended an offer). In this setup, the instrument Z_i is an instrument for receiving an offer rather than being tested. This can potentially be the same instrument as previously used. The resulting 2SLS coefficient would deliver an estimand whose units are "offer accepts" / "offer extends" for the marginal candidate.

Accepting offers is one of many "downstream" outcomes that researchers may care about. We may also care about how downstream outcomes such as productivity and retention once on the job, as well as the characteristics of productivity (innovativeness, efficiency, effort, etc). This would requiring using an outcome variable Y' representing "total output at the firm" (assuming this can be measured), whose value is zero for those who aren't hired. T' would represent being hired, and Z_i would need to instrument for T (being hired). This procedure would estimate the change in downstream output under the new selection scheme.[40]

We can think of these extensions as a form of imperfect compliance with the instrument. As the econometrician studies outcomes at increasingly downstream stages, the results become increasingly "local," and conditional on the selection process up to that stage. For example, results about accepted job offers may be conditional on the process process for testing, interviewing, persuasion, compensation and bargaining with candidates in the setting being studied. The net effectiveness of A vs B ultimately depends on how these early criteria interact with downstream assessments.

[40] In some cases, such as the setting in this paper, it could make more sense to study output per day of work.

4.3 Comparison of this Method to others in Literature

In this section, I contrast the approach above to those used by other fields studying personell assessment. My experiment allows me to compare my experimental estimates to those obtained by other methods – including methods advocated by government policymakers – and thus quantify the bias in these alternatives (for my setting). I also evaluate the assumptions behind these methodologies empirically. Additional discussion is in Appendix E.

Other studies have recently made causal inferences about the effects of hiring policies. In particular, Autor and Scarborough (2008), Hoffman, Kahn and Li (2016) and Horton (2013). Although these studies have not laid out a potential outcomes framework, they can be re-interpreted in light of the above.

5 Results

In Table 2 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%. 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 subsequent 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.

5.1 Job Performance

The candidates who are hired go on to begin careers at their firm, where their career outcomes can be measured. 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. Later, I use these variables as surrogate outcomes (Prentice, 1989; Begg and Leung, 2000) for subjective performance reviews and promotion using the Athey et al. (2016) framework.

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.[41] These peer reviews cover both the logical structure, formatting and readability of the code as outlined in company guidelines.[42] 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. 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 7. 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.

5.2 Cultural Fit and Leadership Skills

During the sample period of the experiment, the employer in this experiment began asking interviewers for additional quantitative feedback about candidates. The additional questions asked interviewers to assess the candidate separately on multiple dimensions. In particular, they asked interviewers for an assessment of the candidate's "general aptitude," "cultural fit" and "leadership ability." The interviewers were permitted to assess on a 1-5 scale. These questions were introduced to the interviewers gradually and orthogonally to the experiment.

Because of the gradual introduction, do not have assessments for all of the candidates in the experiment. In order to expand the sample size, I combine the variation from the experiment with regression discontinuity around the 10% threshold. For the regression discontinuity, I use the Imbens and Kalyanaraman (2011) bandwidth. The machine picked candidates aren't different

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

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

from the human picked ones in general aptitude, but are more highly rated in soft dimensions such as cultural fit and leadership.

5.3 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 13, 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 13 the human's assessment has no predictive power, even in isolation. The machine's ranking does.

6 Conclusion

The idea that a computer could have better "social skills" than a human may sound counterintuitive. Autor's 2015 discussion of the future of automation mentions social skills as the type of work that can't easily automated because we humans our don't consciously "know the rules" and thus cannot program a machine to perform this work. However, Autor (2015) specifically mentions machine learning – the intervention examined in this paper – as one of two methods for which these tasks could one day be automated.

One possible caveat to the strong performance of the algorithm in this paper is that the algorithm had to be "trained" to model historical human decisions. Without the historical human decision, the algorithm's performance would not have been possible. However, the algorithm could have been easily trained using bandit methods (rather than with historical data). All the firm would have wasted was some possibly bad interviews. This might have been a more efficient, less biased way to train the algorithm if one were starting from scratch.

The results of this study should not be over-interpreted. Despite the positive performance of machines in the selection task, computers are currently inept at many tasks requiring emotional intelligence such as therapy, sales or customer support. As cited by Deming (2015), computer scientists have yet to create a chat robot that can successfully pose as a human (the "Turing Test.")

However, even the results above should be viewed as a lower bound of quality. The statistical modeling approach in this paper is "naive," underutilizing decades-old techniques from other disciplines (such as Heckman's 1979 sample-selection methods or Robbins et al.'s 1952 multi-armed bandit techniques[43]) that could plausibly improve performance. More generally, computational power and input data are increasing over time, and firms continue to invest in new ways to automate social skills. At the time of this writing, several technology companies (Apple, Facebook, Yahoo! Google and others) recently announced large investments in "chatbot" digital assistants with whom users can converse as one would with other humans.

Future research (including a future version of this paper) should better examine where subjective decisionmaking tasks fit into the labor market. One source of data about jobs that require subjective decision making is from the US Labor Department's O*NET database used by Autor, Deming and others. O*NET categorizes occupations based on the tasks, skills and type of work involved. These occupations can be linked to current and historical employment and wage data.

One such category in O*NET contains occupations involving "Judging the Qualities of Things, Services, or People." This category contains professions with heavy subjective decisionmaking requirements. According to O*NET, the occupation requiring the most subjective assessment is *Clinical Psychologist*. Human resource and screening professionals rank near the top. The occupation requiring the least such is *Model* (as in fashion, art or photography).[44] This category may be a good source of data for future research about how automating subjective judgements may impact the labor market and production processes.

[43] Related methods in machine learning are called "active learning." For a survey, see Settles, 2010.

[44] A separate O*NET category, "Evaluating Information to Determine Compliance with Standards," involves occupations requiring more "objective" judgments.

References

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh, "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care," *American Economic Review*, 2016, *106* (12), 3730–64.

Adair, Alastair, Norman Hutchison, Bryan MacGregor, Stanley McGreal, and Nanda Nanthakumaran, "An analysis of valuation variation in the UK commercial property market: Hager and Lord revisited," *Journal of Property Valuation and Investment*, 1996, *14* (5), 34–47.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb, "Exploring the Impact of Artificial Intelligence: Prediction versus Judgment," 2017.

Anderson, James M, Jeffrey R Kling, and Kate Stith, "Measuring interjudge sentencing disparity: Before and after the federal sentencing guidelines," *The Journal of Law and Economics*, 1999, *42* (S1), 271–308.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 1996, *91* (434), 444–455.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks," *ProPublica, May*, 2016, *23.*

Arrow, Kenneth et al., "The theory of discrimination," *Discrimination in labor markets*, 1973, *3* (10), 3–33.

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

_ , Raj Chetty, Guido Imbens, and Hyunseung Kang, "Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index," *arXiv preprint arXiv:1603.09326*, 2016.

Autor, David H, "Why are there still so many jobs? The History and Future of Workplace Automation," *The Journal of Economic Perspectives*, 2015, *29* (3), 3–30.

_ and David Scarborough, "Does job testing harm minority workers? Evidence from retail establishments," *The Quarterly Journal of Economics*, 2008, pp. 219–277.

_ and Susan N Houseman, "Do Temporary-Help Jobs Improve Labor Market Outcomes for Low-Skilled Workers? Evidence from" Work First"," *American Economic Journal: Applied Economics*, 2010, pp. 96–128.

Azevedo, Eduardo M, Deng Alex, Jose Montiel Olea, Justin M Rao, and E Glen Weyl, "A/b testing," 2018.

Becker, Gary Stanley et al., "Economics of Discrimination," 1957.

Begg, Colin B and Denis HY Leung, "On the use of surrogate end points in randomized trials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2000, *163* (1), 15–28.

Busse, Meghan R, Devin G Pope, Jaren C Pope, and Jorge Silva-Risso, "The psychological effect of weather on car purchases," *The Quarterly Journal of Economics*, 2015, *130* (1), 371–414.

Card, David and Gordon B Dahl, "Family violence and football: The effect of unexpected emotional cues on violent behavior," *The Quarterly Journal of Economics*, 2011, *126* (1), 103–143.

Chang, Tom and Antoinette Schoar, "Judge specific differences in Chapter 11 and firm outcomes," *Unpublished working paper, National Bureau of Economic Research Cambridge*, 2013.

Cockburn, Iain M, Samuel Kortum, and Scott Stern, "Are all patent examiners equal? The impact of examiner characteristics," Technical Report, National Bureau of Economic Research 2002.

Colbert, Janet L, "Inherent risk: An investigation of auditors' judgments," *Accounting, Organizations and society*, 1988, *13* (2), 111–121.

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

Cowgill, Bo and Eric Zitzewitz, "Mood Swings at Work: Stock Price Movements, Effort and Decision Making," 2008.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta, "Automated experiments on ad privacy settings," *Proceedings on Privacy Enhancing Technologies*, 2015, *2015* (1), 92–112.

Deming, David J, "The growing importance of social skills in the labor market," Technical Report, National Bureau of Economic Research 2015.

Dobbie, Will and Jae Song, "Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection," *The American Economic Review*, 2015, *105* (3), 1272–1311.

Edmans, Alex, Diego Garcia, and Øyvind Norli, "Sports sentiment and stock returns," *The Journal of Finance*, 2007, *62* (4), 1967–1998.

Eichenwald, Kurt, "Microsoft's Lost Decade," *Vanity Fair*, 2012.

Engelberg, Joseph and Christopher A Parsons, "Worrying about the stock market: Evidence from hospital admissions," *The Journal of Finance*, 2016.

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

Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist, "What is a Patent Worth? Evidence from the US Patent Lottery," Technical Report, National Bureau of Economic Research 2017.

Friedler, Sorelle A. and Christo Wilson, eds, *Conference on Fairness, Accountability and Transparency, 23-24 February 2018*, Vol. 81 of *Proceedings of Machine Learning Research* PMLR 2018.

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.

Grimstad, Stein and Magne Jørgensen, "Inconsistency of expert judgment-based estimates of software development effort," *Journal of Systems and Software*, 2007, *80* (11), 1770–1777.

Gubler, Timothy, Ian Larkin, and Lamar Pierce, "The dirty laundry of employee award programs: Evidence from the field," *Harvard Business School NOM Unit Working Paper*, 2013, (13-069).

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

Hirshleifer, David and Tyler Shumway, "Good day sunshine: Stock returns and the weather," *The Journal of Finance*, 2003, *58* (3), 1009–1032.

Hoffman, Mitch, Lisa B Kahn, and Danielle Li, "Discretion in Hiring," 2016.

Horton, John J, "The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment," *Forthcoming, Journal of Labor Economics*, 2013.

Imbens, Guido and Karthik Kalyanaraman, "Optimal bandwidth choice for the regression discontinuity estimator," *The Review of economic studies*, 2011, p. rdr043.

Jovanovic, Boyan, "Job matching and the theory of turnover," *The Journal of Political Economy*, 1979, pp. 972–990.

Jr, Joseph J Doyle, "Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care," *Journal of political Economy*, 2008, *116* (4), 746–770.

_ **et al.**, "Child Protection and Child Outcomes: Measuring the Effects of Foster Care," *American Economic Review*, 2007, *97* (5), 1583–1610.

Kahneman, Daniel, "Remarks by Daniel Kahneman," *NBER Economics of AI Workshop*, 2017.

_ **, M Rosenfield, Linnea Gandhi, and Tom Blaser**, "Noise: How to overcome the high, hidden cost of inconsistent decision making," *Harvard Business Review*, 2016, *10*, 38–46.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," Technical Report, National Bureau of Economic Research 2017.

Kling, Jeffrey R, "Incarceration length, employment, and earnings," *The American economic review*, 2006, *96* (3), 863–876.

Lambrecht, Anja and Catherine E Tucker, "Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads," 2016.

Lazear, Edward P, Kathryn L Shaw, and Christopher T Stanton, "Who Gets Hired? The Importance of Finding an Open Slot," Technical Report, National Bureau of Economic Research 2016.

Maestas, Nicole, Kathleen J Mullen, and Alexander Strand, "Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt," *The American Economic Review*, 2013, *103* (5), 1797–1829.

Marschak, Jacob, "Binary choice constraints and random utility indicators," Technical Report, YALE UNIV NEW HAVEN CT COWLES FOUNDATION FOR RESEARCH IN ECONOMICS 1959.

Milgrom, Paul and John Roberts, "An economic approach to influence activities in organizations," *American Journal of sociology*, 1988, pp. S154–S179.

Milgrom, Paul R, "Employment contracts, influence activities, and efficient organization design," *The Journal of Political Economy*, 1988, pp. 42–60.

Neyman, Jerzy S, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.," *Statistical Science*, 1923/1990, *5* (4), 465–472.

Oyer, Paul and Scott Schaefer, "Personnel Economics: Hiring and Incentives," *Handbook of Labor Economics*, 2011, *4*, 1769–1823.

Phelps, Edmund S, "The statistical theory of racism and sexism," *The american economic review*, 1972, pp. 659–661.

Prentice, Ross L, "Surrogate endpoints in clinical trials: definition and operational criteria," *Statistics in medicine*, 1989, *8* (4), 431–440.

Rind, Bruce, "Effect of beliefs about weather conditions on tipping," *Journal of Applied Social Psychology*, 1996, *26* (2), 137–147.

Robbins, Herbert et al., "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, 1952, *58* (5), 527–535.

Rubin, Donald B, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, 1974, *66* (5), 688.

Rubin, Donald B., "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association*, 2005, *100* (469), 322–331.

Sampat, Bhaven and Heidi L Williams, "How do patents affect follow-on innovation? Evidence from the human genome," *available at http://economics.mit.edu/files/9778*, 2014.

Schwarz, Norbert and Gerald L Clore, "Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states.," *Journal of personality and social psychology*, 1983, *45* (3), 513.

Selten, Reinhard, "Reexamination of the perfectness concept for equilibrium points in extensive games," *International journal of game theory*, 1975, *4* (1), 25–55.

Settles, Burr, "Active learning literature survey," *University of Wisconsin, Madison*, 2010, *52* (55-66), 11.

Slovic, Paul, "Analyzing the expert judge: A descriptive study of a stockbroker's decision process.," *Journal of Applied Psychology*, 1969, *53* (4), 255.

Sweeney, Latanya, "Discrimination in online ad delivery," *Queue*, 2013, *11* (3), 10.

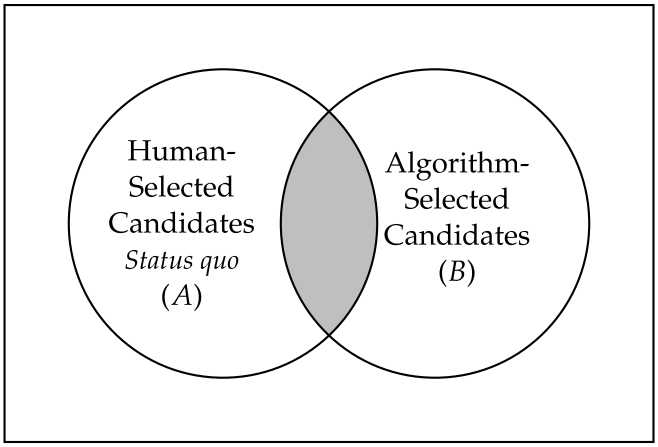
Taylor, Robert L and William D Wilsted, "Capturing judgment policies: A field study of performance appraisal," *Academy of Management Journal*, 1974, *17* (3), 440–449.

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.

Tables and Figures

Figure 1: Caption for figure below.



Notes: The above is a useful visualization of the empirical setting. The left circle A represents candidates selected by the status quo. The right circle B represents candidates selected by the machine learning. The shaded area $A \cap B$ represents candidates accepted by both. The goal of the empirical analysis is to compare the average outcomes between the unshaded areas of A and B ($A \setminus B$ vs $B \setminus A$).

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	0.513	0.164	0.349***
Referred	0.147	0.054	0.093***
No Work Experience (New Graduate)	0.189	0.182	0.007
Rare School	0.387	0.735	-0.348***

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 have a greater than 10% probability of receiving an offer. "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	0.56	0.51	0.04
Referred	0.13	0.15	-0.02
No Work Experience (New Graduate)	0.22	0.19	0.03
Rare School	0.41	0.39	0.02

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**	0.10
	(0.033)	(0.036)	(0.077)
R^2	0.19	0.22	0.27
Observations	770	544	124

Panel C: Marginal Success Rates using Instrumental Variables

	Job Offer	Accept Offer
Interview	0.34***	
	(0.046)	
Job Offer		0.91***
		(0.030)
F-stat (1st Stage)	69.7	177.6
Mean Outcome of Control	0.20	0.71
Difference	0.14***	0.2***
Mean Outcome of Population	0.026	0.84
Difference	0.31***	0.07***
Observations	38242	38242

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	0.13***	0.18**
	(0.043)	(0.088)
R^2	0.021	0.082
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 3, 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)	0.38**	0.14***
	(0.15)	(0.041)
F-stat (1st Stage)	11.5	11.5
Mean Outcome of Control	-0.088	-0.12
Difference	0.47***	0.26***
Observations	770	770

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 7. 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: Leadership, Cultural Fit and General Aptitude

	Cultural Fit	Leadership	General Aptitude
Interviewed (inst w/ Treatment & RD)	1.89***	1.80***	1.02**
	(0.090)	(0.093)	(0.44)
F-stat (1st Stage)	455.2	387.3	96.6
Mean Outcome of Control	1.52	1.60	1.28
Difference	0.37***	0.2**	-.26
Observations	54328	54328	54328

Notes: This table presents results on the interview evaluations of the marginal candidate preferred by the algorithm, but rejected by a human. The coefficient estimated in the first row is the average assessment in each dimension of these marginal candidates. In the "Difference" row, I report the difference between these marginal machine-picked candidates against the average candidate in the experiment.

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

Table 8: Combination of Algorithmic + Human Evaluation

Panel A: Passing Interviews

	Job Offer	Job Offer	Job Offer	Job Offer	Job Offer	Job Offer
Human would Interview	0.14**					0.099
	(0.063)					(0.061)
Human Score		0.031***			0.020*	
		(0.012)			(0.012)	
Algorithm would Interview			0.25***			0.24***
			(0.051)			(0.051)
Algorithm Score				0.13***	0.12***	
				(0.024)	(0.025)	
R^2	0.18	0.18	0.22	0.26	0.27	0.22
Observations	333	333	333	333	333	333

Panel B: Accepting Offers

	Accept Offer	Accept Offer	Accept Offer	Accept Offer	Accept Offer	Accept Offer
Human would Interview	0.041					0.031
	(0.21)					(0.22)
Human Score		0.040*			0.027	
		(0.023)			(0.021)	
Algorithm would Interview			0.040			0.097
			(0.16)			(0.16)
Algorithm Score				0.11***	0.10***	
				(0.029)	(0.030)	
R^2	0.21	0.24	0.20	0.31	0.34	0.21
Observations	82	82	82	82	82	82

Notes: This table contains "horserace" regressions (Fair and Shiller, 1989) predicting candidate outcomes from machine and human assessments. See Section 5.3 for discussion.

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

Table 9: Heterogeneous Treatment Effects: Education Degree and Quality

Panel A: Extended Offer

	All	Doctorate	Elite School	Non-Elite School
Treatment	0.15***	0.34***	0.14***	0.20***
	(0.042)	(0.086)	(0.054)	(0.073)
Adj. R^2	0.026	0.14	0.021	0.053

Panel B: Interviewed

	All	Doctorate	Elite School	Non-Elite School
Treatment	0.36***	0.49***	0.31***	0.43***
	(0.033)	(0.074)	(0.041)	(0.059)
Adj. R^2	0.22	0.32	0.19	0.24

Panel C: Marginal Pass Rate

	All	Doctorate	Elite School	Non-Elite School
Machine-only Success Rate	0.40***	0.70***	0.45***	0.46***
	(0.11)	(0.19)	(0.17)	(0.17)
F-stat	120.7	44.1	57.8	53.6
Machine + Human Rate	0.20	0.10	0.22	0.13
Difference	0.2*	0.59***	0.23	0.33**
Human Only Rate	0.023	0.027	0.033	0.017
Difference	0.38***	0.67***	0.42**	0.44***

Notes:

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

Table 10: Heterogeneous Treatment Effects: Education Topic

Panel A: Extended Offer

	All	CS	Other Sci/Eng	EE	Bus/Econ	Math	Hum/SocSci
Treatment	0.15***	0.15***	0.11	0.064	-0.014	0.38***	-0.097
	(0.042)	(0.045)	(0.069)	(0.11)	(0.22)	(0.11)	(0.19)
Adj. R^2	0.026	0.026	0.011	-0.014	-0.059	0.16	-0.030

Panel B: Interviewed

	All	CS	Other Sci/Eng	EE	Bus/Econ	Math	Hum/SocSci
Treatment	0.36***	0.36***	0.32***	0.37***	0.29**	0.38***	0.45***
	(0.033)	(0.036)	(0.055)	(0.10)	(0.13)	(0.085)	(0.16)
Adj. R^2	0.22	0.22	0.19	0.22	0.042	0.21	0.30

Panel C: Marginal Pass Rate

	All	CS	Other Sci/Eng	EE	Bus/Econ	Math	Hum/SocSci
Machine-only Success Rate	0.40***	0.41***	0.35*	0.17	-0.050	1.00***	-0.21
	(0.11)	(0.12)	(0.21)	(0.29)	(0.74)	(0.33)	(0.45)
F-stat	120.7	97.6	35.0	13.8	5.01	20.4	8.46
Machine + Human Rate	0.20	0.20	0.19	0.16	0.27	0.13	0.28
Difference	0.2*	0.21*	0.15	0.01	-0.32	0.87***	-0.49
Human Only Rate	0.023	0.022	0.020	0.029	0.028	0.049	0.048
Difference	0.38***	0.39***	0.33	0.14	-0.08	0.95***	-0.26

Notes:

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

Table 11: Heterogeneous Treatment Effects: Background and Experience

Panel A: Extended Offer

	All	Referred	Not Referred	No Experience	Experience
Treatment	0.15***	0.12	0.15***	0.36***	0.094*
	(0.042)	(0.12)	(0.044)	(0.089)	(0.051)
Adj. R^2	0.026	-0.00068	0.029	0.14	0.0086

Panel B: Interviewed

	All	Referred	Not Referred	No Experience	Experience
Treatment	0.36***	0.16**	0.40***	0.43***	0.35***
	(0.033)	(0.065)	(0.037)	(0.077)	(0.040)
Adj. R^2	0.22	0.067	0.25	0.28	0.21

Panel C: Marginal Pass Rate

	All	Referred	Not Referred	No Experience	Experience
Machine-only Success Rate	0.40***	0.76	0.38***	0.84***	0.27*
	(0.11)	(0.77)	(0.11)	(0.23)	(0.14)
F-stat	120.7	5.73	119.8	30.8	76.9
Machine + Human Rate	0.20	0.24	0.19	0.18	0.20
Difference	0.2*	0.52	0.19*	0.66***	0.07
Human Only Rate	0.023	0.041	0.021	0.033	0.020
Difference	0.38***	0.72	0.36***	0.81***	0.25*

Notes:

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

Table 12: Heterogeneous Treatment Effects: Prob Selection by Human Screeners

Panel A: Extended Offer

	All	$\hat{p} \uparrow$	$\hat{p} \downarrow$	$\sigma_{est} \uparrow$	$\sigma_{est} \downarrow$
Treatment	0.15***	0.094	0.15**	0.20***	0.031
	(0.042)	(0.060)	(0.061)	(0.065)	(0.035)
Adj. R^2	0.026	0.0057	0.036	0.034	-0.00092

Panel B: Interviewed

	All	$\hat{p} \uparrow$	$\hat{p} \downarrow$	$\sigma_{est} \uparrow$	$\sigma_{est} \downarrow$
Treatment	0.36***	0.25***	0.47***	0.32***	0.41***
	(0.033)	(0.043)	(0.048)	(0.046)	(0.048)
Adj. R^2	0.22	0.16	0.25	0.20	0.24

Panel C: Marginal Pass Rate

	All	$\hat{p} \uparrow$	$\hat{p} \downarrow$	$\sigma_{est} \uparrow$	$\sigma_{est} \downarrow$
Machine-only Success Rate	0.40***	0.38	0.33**	0.61***	0.076
	(0.11)	(0.23)	(0.13)	(0.19)	(0.084)
F-stat	120.7	33.0	97.5	47.9	74.0
Machine + Human Rate	0.20	0.28	0.12	0.34	0.049
Difference	0.2*	0.1	0.21*	0.27	0.03
Human Only Rate	0.023	0.054	0.011	0.066	0.013
Difference	0.38***	0.33	0.31**	0.55***	0.06

Notes:

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

Table 13: Transparency vs. Treatment Effects

	Weight in Scoring Algo.	Treatment Effect
Referred	0.031***	0.12
	(0.0031)	(0.12)
Non-Elite School	-0.038***	0.20***
	(0.0016)	(0.070)
No Experience	0.0024	0.36***
	(0.0019)	(0.091)
Statistically Non-Traditional	-0.077***	0.15***
	(0.0022)	(0.056)

Notes:

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

Appendices

A Proofs of Main Propositions

A.1 Proof of Proposition 1

Note that $q = 1 - F(x)$, which is decreasing in the argument x. In our setting, $x = r(p_0 - p_1) + b$, which is increasing in b and so $1 - F(x)$ is decreasing in b. Note that $p_1 > p_0$, so x is also decreasing in $r \implies 1 - F(x)$ is increasing in r. Also: x is decreasing in $p_1 - p_0$, so $1 - F(x)$ is increasing in $p_1 - p_0$. Finally, if the variance of F is large, then $F(x)$ is smaller for any given x given the earlier restrictions on F. As a result, $1 - F(x)$ is larger (increasing).

A.2 Proof of Proposition 2

If screeners are biased but not noisy, then $b > 0$ and the variance of F is zero. As a result, the probability that the human will pick the correct (Type 1) candidate is $q = 0$, because $\underline{\eta}$ will have to be positive to offset the bias. If F has mean and variance of zero, η will never be positive.

The machine learning engineer will receive a training dataset consisting of half Type 0, half Type 1s. The Type 1 candidates will all have an outcome y equal to zero. The Type 0 candidates will have outcomes y that are 1 with probability p_0, and 0 with probability $1 - p_0$. Thus the machine predictions of $E[Y|\theta = 0]$ will converge to p_0, and $E[Y|\theta = 1]$ will converge to zero.

The engineers will then pick the Type with the greatest estimated $E[Y|\theta]$, and will implement the decision consistently. Because $p_0 > 0$, the machine will prefer the Type 0 candidates. The original, underlying human process will yield p_0 successful candidates. The machine learning algorithm will yield the same output, creating zero difference in the quality of decison-making.

A.3 Proof of Proposition 3

If screeners exhibit zero bias then $b = 0$, they prefer Type 1 but pick Type 0 only if a noise realization η is sufficiently low. This will happen with probability $1 - q$. If $b = 0$, noisier Fs increase the probability of switching the decision-maker away from his/her default. In this setting, the default is to pick Type 1 (the better candidate), and so noisier Fs will decrease q (as discussed in Proposition 1).

The machine learning engineer will receive a training dataset consisting of half Type 0, half Type 1s. The Type 1 candidates will have outcomes y that are 1 with probability qp_1, and 0 with the remaining probability. The Type 0 candidates will have outcomes y that are 1 with probability $(1 - q)p_0$, and 0 with the remaining probability.

Thus the machine predictions of $E[Y|\theta = 1]$ will converge to qp_1, and $E[Y|\theta = 0]$ will converge to $(1 - q)p_0$. The engineers will then pick the Type with the greatest estimated $E[Y|\theta]$, and will implement the decision consistently. Because $qp_1 > (1 - q)p_0$, the machine will prefer the Type 1 candidates.

The original, underlying human process will yield $qp_1 + (1 - q)p_0$ successful candidates. The machine learning algorithm will yield p_1 successful candidates, for a output difference in output of $(p_1 - p_0)(1 - q)$. Note that this is increasing in the quality difference between Type 1 and Type 0 candidates $(p_1 - p_0)$ and in the amount of noise. Because $b = 0$, as the variance of F goes up, q goes down.

A.4 Proof of Proposition 4

If screeners are biased, then $b > 0$ and greater variance in F increases q, the probability that Type 1 will be selected. The machine's predictions will converge to $E[Y|\theta = 0] = (1 - q)p_0$ for Type 0 and $E[Y|\theta = 0] = qp_1$ for Type 1.

The ML engineers will select Type 1 if $qp_1 > (1 - q)p_0$, and will otherwise select Type 0. It will select Type 0 if $q < p_0/(p_1 + p_0)$. For a given r, b, p_1 and p_0, this happens only if the noise function F does not have sufficiently large variance.

If the variance in F is not sufficiently large, the algorithm will implement Type 0 (the wrong choice) consistently. Performance under this algorithm will yield p_0 candidates. This is worse for the performance than if the original human process had been used, which would yield $qp_1 + (1 - q)p_0$.

A.5 Proof of Proposition 5

If screeners are biased, then $b > 0$ and greater variance in F increases q, the probability that Type 1 will be selected. The machine's predictions will converge to $E[Y|\theta = 0] = (1 - q)p_0$ for Type 0 and $E[Y|\theta = 0] = qp_1$ for Type 1.

The ML engineers will select Type 1 if $qp_1 > (1 - q)p_0$, and will otherwise select Type 0. It will select Type 1 if $q > p_0/(p_1 + p_0)$. For a given r, b, p_1 and p_0, this happens only if the noise function F has sufficiently large variance.

A.6 Proof of Proposition 6

Even if screeners are biased and sufficiently noisy to reduce bias ($q > p_0/(p_1 + p_0)$, see Proposition 5), the algorithm's predictions still underestimate the advantage of Type 1 above Type 0. In particular, the algorithm predicts a y of qp_1 for Type 1 and $(1 - q)p_0$ for Type 0. The algorithm's implicit quality ratio of Type 1 over Type 0 is $qp_1/(1 - q)p_0$. This is less than the quality ratio of Type 1 over Type 0 (p_1/p_0) – unless noise is maximized by increasing the variance of F until $q = 1/2$.

A.7 Proof of Proposition 7

In the presence of bias, the algorithm can reduce the bias if $q > p_0/(p_1 + p_0) = \underline{q}$. Recall from Proposition 1 that q is decreasing in b and increasing in the variance of F. To achieve any arbitrary $q > \underline{q}$ (for a fixed r, p_1 and p_0), either i) the variance of F can stay fixed and bias b can go down, or ii) the bias b can stay fixed and the variance can increase.

A.8 Proof of Proposition 8

In the presence of bias, the algorithm can reduce the bias if $q > p_0/(p_1 + p_0) = \underline{q}$. This proof will examine the false positive and false negative rates in the training data of this setting. In this setting, Proposition 5 shows that the machine learning algorithm will select Type 1. Proposition 7 shows that as the variance of F increases, the algorithm will be able to reduce smaller and smaller biases.

I will study two particular measures of goodness-of-fit: Precision and recall, as these are commonly used in the machine learning literature. In this setting, recall measures: "If a candidate is selected by a machine, what's the probability that a human screener would have picked it as well?" Precision measures, "If the candidate is selected by the human, what's the probability that the machine would have selected it?" Precision and recall can vary between zero and one, with higher values corresponding to higher goodness-of-fit.

In our setting, recall is q (the probability that the human would pick a machine approved candidate of Type 1). Precision is $q/2$. Note that these move in the same direction as a function of the primitives (p_1, p_0, b and r) as discussed in Proposition 1. Both precision and recall are increasing in the amount of noise.

However, recall that in our setup, $q \in [0, 1/2]$. This means that precision cannot go above 0.5 and cannot go above 0.25. These are relatively low benchmarks. By comparison, both precision and recall on the training data are 1 from Proposition 2, where screeners exhibit bias but no noise.

A.9 Proof of Proposition 9

The "coefficient" or "weight" the machine learning algorithm places on feature $\theta = 0$ is $E[Y|\theta = 1] - E[Y|\theta = 0]$. This is the difference in the algorithm's expected score for Type 1 and Type 0 candidates. The treatment effect of the machine learning algorithm on Type 1 candidates equals the change in the probability of Type 1 candidates being selected between the human decision-makers and algorithm.

I will show that these two quantities are not generically equal or even the same sign. I will present two counterexamples:

- Suppose screeners are biased, but NOT sufficiently noisy. As a result the algorithm codifies bias rather than reduces it ($q < p_0/(p_1 + p_0)$). The coefficient or weight in the machine learning algorithm will equal $qp_1 - (1 - q)p_0$.

  In the human regime, Type 1's probability of being selected is $q < 1$, and using the algorithm the probability is 0. The treatment effect is equal to $-q$, which is not generally equal to the coefficient or weight featured in the algorithm ($qp_1 - (1-q)p_0$) and not even the same sign.

- Now suppose screeners are biased, and sufficiently noisy. As a result the algorithm improves bias ($q > p_0/(p_1+p_0)$). The coefficient or weight in the machine learning algorithm will again equal $qp_1 - (1-q)p_0$.

  In the human regime this value is $q < 1$, and using the algorithm the probability is 1. The treatment effect is equal to $1-q$, which is not generally equal to the coefficient or weight featured in the algorithm ($qp_1 - (1-q)p_0$).

A.10 Alternative Microfoundations to Noise and Bias

B Propositions with Additional Bias

B.1 Microfoundations of δ

As discussed in Section 2.4.2, bias may come both in choices to hire as well as how to evaluate candidates who have been hired. In this model, b represents the amount of bias in hiring, and δ represents how much Type 1 candidates are discriminated against in evaluation.

Like b, δ can have several microfoundations. One possibility, presented above, is that taste-based discrimination is responsible for δ. In this microfoundation, an evaluator experiences direct utility gains for underreporting Type 1's success.

However, δ can also represent statistical discrimination by an evaluator. An evaluator may not be able to directly measure a worker's productivity. This is a realistic assumption in many settings, where a worker's output cannot be fully monitored.

For example: A salesperson may be hired to invest in relationships with clients. Her manager may not be able to observe all aspects of relationship-building, and the outcome of this activity may take years to realize in firm revenue. The manager may therefore may use a worker's observable characteristics to make educated guesses about the quality of the work that's hard to directly observe. If these inferences are wrong and slow to update, this will lead to a $\delta \in [0,1]$ based on statistical discrimination.

B.2 Additional Propositions

Proposition 10. *The screener's probability of picking Type 1 candidates (q), is decreasing in b, increasing in r, increasing in the quality difference in Type 1 and Type 0 ($p_1 - p_0$), decreasing in δ and increasing in the variance of F.*

Proof. A risk-neutral human screener will make the "right" decision (Type 1) if $rp_1(1-\delta) + \eta > rp_0 + b$. In other words, the screener makes the right decision if the random utility shocks are enough to offset the taste-based bias (b) favoring Type 0.

Let $\underline{\eta} = r(p_0 - p_1(1-\delta)) + b$ be the minimum η necessary to offset the bias, given the other rewards involved. Such an η (or greater) happens with probability of $\Pr(\eta > r(p_0 - p_1(1-\delta)) + b) = 1 - F(r(p_0 - p_1(1-\delta)) + b)) = q$.

Note that $q = 1 - F(x)$, which is decreasing in the argument x. In our setting, $x = r(p_0 - p_1(1-\delta)) + b$, which is increasing in b and so $1 - F(x)$ is decreasing in b. Note that $p_1 > p_0$, so x is also decreasing in $r \implies 1 - F(x)$ is increasing in r. Also: x is decreasing in $p_1 - p_0$, so $1 - F(x)$ is increasing in $p_1 - p_0$. x is increasing in δ, so $1 - F(x)$ is decreasing in δ. Note that the addition of δ to the model will make q smaller than in the original Proposition 1, unless $\delta = 0$. Finally, if the variance of F is large, then $F(x)$ is smaller for any given x given the earlier restrictions on F. As a result, $1 - F(x)$ is larger (increasing). □

Proposition 11. *If bias in the evaluation δ is above a threshold, the machine learning approach will entrench bias irrespective of the amount of noise. This threshold value is decreasing in the quality of Type 1.*

Thus the machine predictions of $E[Y|\theta = 1]$ will converge to $qp_1(1-\delta)$, and $E[Y|\theta = 0]$ will converge to $(1-q)p_0$. The algorithm picks Type 1 if $qp_1(1-\delta) > (1-q)p_0$, which happens only if $\delta < (p_1 + p_0)/p_1$, which is decreasing in p_1, increasing in p_0 and decreasing in the quality difference $p_1 - p_0$.

Proposition 12. *If screeners exhibit bias but zero noise, the algorithm will perfectly codify the humans' historical bias. The algorithm's perfomance will precisely equal that of the biased screeners.*

Proof. If screeners are biased but not noisy, then $b > 0$ and the variance of F is zero. In other words, the screener makes the right decision if the random utility shocks are enough to offset the taste-based bias (b) against hiring Type 1 and against scoring Type 1. As a result, the probability that the human will pick the correct (Type 1) candidate is $q = 0$, because $\underline{\eta}$ will have to be positive to offset the bias. If F has mean and variance of zero, η will never be positive. This is essentially unchanged from Proposition 12.

The machine learning engineer will receive a training dataset consisting of half Type 0, half Type 1s. The Type 1 candidates will all have an outcome y equal to zero. The Type 0 candidates will have outcomes y that are 1 with probability p_0, and 0 with probability $1 - p_0$. Thus the machine predictions of $E[Y|\theta = 0]$ will converge to p_0, and $E[Y|\theta = 1]$ will converge to zero.

The engineers will then pick the Type with the greatest estimated $E[Y|\theta]$, and will implement the decision consistently. Because $p_0 > 0$, the machine will prefer the Type 0 candidates. The original, underlying human process will yield p_0 successful candidates. The machine learning algorithm will yield the same output, creating zero difference in the quality of decison-making. □

Proposition 13. *If screeners exhibit i. zero bias in hiring ($b = 0$), ii. zero or sufficiently low bias in scoring hired workers ($\delta < (p_1 - p_0)/p_1$), but iii. non-zero amounts of noise, the algorithm will improve upon the performance of the screeners by removing noise. The amount of performance improvement is increasing in the amount of noise and the quality difference between Type 1 and Type 0 candidates.*

If screeners exhibit zero bias then $b = 0$, they prefer Type 1 but pick Type 0 only if a noise realization η is sufficiently low. This will happen with probability $1 - q$. If $b = 0$, noisier Fs increase the probability of switching the decision-maker away from his/her default. In this setting, the default is to pick Type 1 (the better candidate), and so noisier Fs will decrease q (as discussed in Proposition 1).

The machine learning engineer will receive a training dataset consisting of half Type 0, half Type 1s. The Type 1 candidates will have outcomes y that are 1 with probability $qp_1(1-\delta)$, and 0 with the remaining probability. The Type 0 candidates will have outcomes y that are 1 with probability $(1-q)p_0$, and 0 with the remaining probability.

Thus the machine predictions of $E[Y|\theta = 1]$ will converge to $qp_1(1-\delta)$, and $E[Y|\theta = 0]$ will converge to $(1-q)p_0$. The engineers will then pick the Type with the greatest estimated $E[Y|\theta]$, and will implement the decision consistently. Because $qp_1(1-\delta) > (1-q)p_0$, the machine will prefer the Type 1 candidates.

The original, underlying human process will yield $qp_1(1-\delta) + (1-q)p_0$ successful candidates. The machine learning algorithm will yield $p_1(1-\delta)$ successful candidates, for a output difference in output of $(p_1(1-\delta) - p_0)(1-q)$. Note that this is decreasing in δ, increasing in the quality difference between Type 1 and Type 0 candidates $(p_1 - p_0)$, and in the amount of noise. Because $b = 0$ and $(\delta < (p_1 - p_0)/p_1)$, as the variance of F goes up, q goes down.

Proposition 14. *If screeners and evaluators are biased ($b > 0$ and $\delta > 0$) are NOT sufficiently noisy, the algorithm will codify bias. The reduction in noise will actually make outcomes worse.*

If screeners are biased, then $b > 0$ and greater variance in F increases q, the probability that Type 1 will be selected. The machine's predictions will converge to $E[Y|\theta = 0] = (1-q)p_0$ for Type 0 and $E[Y|\theta = 0] = qp_1(1-\delta)$ for Type 1.

The ML engineers will select Type 1 if $qp_1(1-\delta) > (1-q)p_0$, and will otherwise select Type 0. It will select Type 0 if $q < p_0/(p_1(1-\delta) + p_0)$. For a given r, b, p_1, δ and p_0, this happens only if the noise function F does not have sufficiently large variance.

If the variance in F is not sufficiently large, the algorithm will implement Type 0 (the wrong choice) consistently. Performance under this algorithm will yield p_0 candidates. This is worse for the performance than if the original human process had been used, which would yield $qp_1(1-\delta) + (1-q)p_0$.

Proposition 15. *If screeners are biased and sufficiently noisy, the algorithm will reduce the humans' bias.*

If screeners are biased, then $b > 0$ and greater variance in F increases q, the probability that Type 1 will be selected. The machine's predictions will converge to $E[Y|\theta = 0] = (1-q)p_0$ for Type 0 and $E[Y|\theta = 0] = qp_1(1-\delta)$ for Type 1.

The ML engineers will select Type 1 if $qp_1(1-\delta) > (1-q)p_0$, and will otherwise select Type 0. It will select Type 1 if $q > p_0/(p_1(1-\delta) + p_0)$. For a given r, b, p_1, δ and p_0, this happens only if the noise function F has sufficiently large variance.

Proposition 16. *If the algorithms' human data contains non-zero bias then, "algorithmic bias" cannot be reduced to zero unless the humans in the training data were perfectly noisy (ie, picking at random).*

Even if screeners are sufficiently noisy to reduce bias (as in Proposition 15), the algorithm's predictions still underestimate the advantage of Type 1 above Type 0 – unless the training data were perfectly representative (ie, humans were picking workers at random). Despite the reduction in bias, the algorithm will remain handicapped and exhibit some bias because of its training on biased training data.

Picking at random is extremely unlikely to appear in any real-world setting, since the purpose of most hiring is to select workers who are better than average and thus under-sample sections of the applicant pool perceived to be weaker. A complete removal of bias therefore appears infeasible from training datasets from real-world observations, particularly observations of agents who are *not* optimizing labels for *ex-post* learning.

Even if screeners are biased and sufficiently noisy to reduce bias ($q > p_0/(p_1(1-\delta)+p_0)$, see Proposition 15), the algorithm's predictions still underestimate the advantage of Type 1 above Type 0. In particular, the algorithm predicts a y of qp_1 for Type 1 and $(1-q)p_0$ for Type 0. The algorithm's implicit quality ratio of Type 1 over Type 0 is $qp_1/(1-q)p_0$. This is less than the quality ratio of Type 1 over Type 0 (p_1/p_0) – unless noise is maximized by increasing the variance of F until $q = 1/2$.

Proposition 17. *The minimum amount of noise necessary for the machine learning to reduce bias is a* decreasing *function of the amount of bias.*

Proposition 17 means that if the screeners display a large amount of bias, only a small amount of noise is necessary for the machines to correct the bias. Similarly if screeners display a small amount of bias, then high amounts of noise are necessary for the algorithm to correct the bias.

The intuition behind Proposition 17 is as follows: Suppose that screeners were highly biased against Type 1 workers, this would conceal the large productivity differences between Type 1 and Type 0 candidates. The machine learning algorithm would need to see only a few realizations – a small amount of noise – in order to reduce the bias. Because each "experiment" on Type 1 workers shows so much greater productivity, few such experiments would be necessary for the algorithm to learn the improvement.

By contrast, if the bias against Type 1 is small – large amounts of noise would be necessary for the algorithm to learn its way out of it. This is because each "experiment" yields a smaller average productivity gain. As a result, the algorithm requires more observations in order to understand the gains from picking Type 1 candidates.

In the presence of bias, the algorithm can reduce the bias if $q > p_0/(p_1+p_0) = \underline{q}$. Recall from Proposition 10 that q is decreasing in b and increasing in the variance of F. To achieve any arbitrary $q > \underline{q}$ (for a fixed r, p_1 and p_0), either i) the variance of F can stay fixed and bias b can go down, or ii) the bias b can stay fixed and the variance can increase.

Proposition 18. *In settings featuring bias sufficiently high noise, the algorithm's improvement in bias will be positive and increasing in the level of noise and bias. However, metrics of goodness-of-fit on the training data (and on additional observations from the data-generating process) are* decreasing *in the amount of noise and bias.*

The implication of Proposition 18 is: If engineers avoid settings where models exhibit poor goodness-of-fit on the training data (and future samples), they will avoid the settings where machine learning has the greatest potential to reduce bias.

Proposition 19. *The "coefficient" or "weight" the machine learning algorithm places on $\theta = 1$ when ranking candidates does not equal the treatment effect of using the algorithm rather than human discretion for $\theta = 1$ candidates.*

Proposition 19 discusses how observers should interpret the coefficients and/or weights of the machine learning algorithm. It shows that these weights may be highly misleading about the impact of the algorithm. For example: It's possible for an algorithm that places negative weight on $\theta = 1$ when ranking candidates could nonetheless have a strong positive benefit for $\theta = 1$ candidates and their selection outcomes. This would happen if the human penalized these characteristics even more than the algorithm did.

The internal weights of these algorithms are completely unrelated to which candidates benefit from the algorithm compared to a status quo alternative. The latter comparison requires a comparison to a counterfactual.

The "coefficient" or "weight" the machine learning algorithm places on feature $\theta = 0$ is $E[Y|\theta = 1] - E[Y|\theta = 0]$. This is the difference in the algorithm's expected score for Type 1 and Type 0 candidates. The treatment effect of the machine learning algorithm on Type 1 candidates equals the change in the probability of Type 1 candidates being selected between the human decision-makers and algorithm.

I will show that these two quantities are not generically equal or even the same sign. I will present two counterexamples:

- Suppose screeners are biased, but NOT sufficiently noisy. As a result the algorithm codifies bias rather than reduces it ($q < p_0/(p_1 + p_0)$). The coefficient or weight in the machine learning algorithm will equal $qp_1 - (1 - q)p_0$.

  In the human regime, Type 1's probability of being selected is $q < 1$, and using the algorithm the probability is 0. The treatment effect is equal to $-q$, which is not generally equal to the coefficient or weight featured in the algorithm ($qp_1 - (1 - q)p_0$) and not even the same sign.

- Now suppose screeners are biased, and sufficiently noisy. As a result the algorithm improves bias ($q > p_0/(p_1 + p_0)$). The coefficient or weight in the machine learning algorithm will again equal $qp_1 - (1 - q)p_0$.

  In the human regime this value is $q < 1$, and using the algorithm the probability is 1. The treatment effect is equal to $1 - q$, which is not generally equal to the coefficient or weight featured in the algorithm ($qp_1 - (1 - q)p_0$).

C IV Assumptions in a Candidate-Level Hiring Field Experiment

1. **SUTVA**: Candidate i's outcome depends only upon his treatment status, and not anyone else's. This permits us to write $T_i(Z) = D_i(Z_i)$ and $Y_i(Z_i, D(Z)) = Y_i(Z_i, D_i(Z_i))$.

   In a testing setup, this assumption might be problematic if candidates are graded on a "curve" or relative ranking, rather than against an absolute standard.[45] It would also be problematic if the firm (or candidates) in question were powerful enough in the labor market to create general equilibrium effects through the testing of specific candidates.

2. **Ignorable assignment of Z**. Z_i must be randomly assigned, or $0 < \Pr(Z_i = 1|X_i = x) = \Pr(Z_j = 1|X_j = x) < 1, \forall i, j, x$.

3. **Exclusion restriction**, or $Y(Z, T) = Y(Z', T), \forall Z, Z', T$. The instrument only affects the outcome through the decision to administer the test. For a given value of T_i, the value of Z_i must not affect the outcome.

   In a testing setting, one implication of the exclusion restriction is that the test must be graded fairly, so that the resulting pass/fail out are not biased to reflect the grader's preferences for Criteria A vs B. Biased test grading would violate the exclusion restriction.[46] Double-blind or objective evaluation may help meet the exclusion restriction.

   A satisfied exclusion restriction lets us write $Y(Z, T)$ as $Y(T)$. Assumption 1 lets us write $Y_i(T)$ as $Y_i(T_i)$.

4. **Inclusion restriction**. The instrument must have a non-zero effect on who is tested ($E[T_i(1)T_i(0)|X_i] \neq 0$, or $Cov(Z, T|X) \neq 0$).

5. **Monotonicity**, or $T_i(1) \geq T_i(0)$ or $T_i(1) \leq T_i(0), \forall i$. This condition requires there to be no "defiers," for whom testing is less likely if the instrument is zero.

D Revisiting IV assumptions for "downstream" hiring outcomes

Introducing a new downstream outcome (Y') and endogenous variables (T') require revisiting the IV assumptions. Even if the IV requirements were met for Y and T (the original variables), this does not automatically mean the IV requirements are met for our second endogenous variable (T') and the downstream outcome (Y').

All IV assumptions must be revisited. Below, I mention a few particular areas where the IV criteria may fail for downstream outcomes in a testing or hiring setting – even if they are first met in upstream ones.

SUTVA. In my empirical setting, there are no cross-candidate comparisons ("grading applicants on a curve") necessary to pass the test; if they were, it would introduce SUTVA violations.

However, even if cross-candidate comparisons were absent from test-grading, they might reappear downstream in offer-acceptances. If an employer has a finite, inelastic number of "slots" (Lazear et al., 2016), then test-passers' acceptance decisions could interact with each other. A candidate who accepts a spot early may block a later one from being able to accept, creating a SUTVA violation.

[45] If candidates were graded by relative ranking, SUTVA would be violated when one candidate's strong performance adversely affects another's chances of passing.

[46] In many instances, test graders may have a preference for what which criteria are used. In the example above: Suppose the test grader was biased against the CEO's opinion (Criteria A) and wanted the evaluation to look poorly for the CEO. Such a grader he/she may CEO-approved candidates if he/she knew them, violating the exclusion restriction.

Similarly, SUTVA violations may arise for performance metrics (such as promotions) that are given in a tournament-like setting with an inelastic number of winners. In this case, the treatment and control outcomes could interact through the contest system; the one group's successes could crowd out the other's. Cowgill (2016) finds such workforce tournaments are common, however promotions are not included as an outcome of this experiment.

I raise these issue as an example of how downstream SUTVA requirements can fail, even if they pass upstream. These issues do *not* arrive in my empirical setting, where the employer wants to hire as many people as could pass the test and does not have a finite quota of offers or slots.[47] I do not use promotions as an outcome variable.

The main-on-the-job performance outcomes that are included in this paper are lines of code added (and deleted). There is finite amount of code that can be submitted. It may still be possible for treated employees and control employees to interact with each other in other ways. However, no two applicants hired through this experiment were assigned to the same direct manager or had the same set of co-workers at any point during the sample. This limits the opportunity for interactions between treatment and control employees.

Inclusion restriction (instrument strength). An instrument Z that has a strong effect on which candidates are tested is not necessarily a strong effect on which candidates are hired. Z could be a much weaker instrument for a downstream T' than for the earlier T. This is partly because there are fewer candidates who passed T and were eligible to take T' – effectively there is a smaller sample size.

E Econometric method compared to approaches from current federal policy and other disciplines

The field of industrial and organizational psychology ("I/O psychology") has a literature about personnel assessment and hiring criteria. This research – and its methodological recommendations – have been very influential in law and public policy surrounding employment selection mechanisms. For example, the Uniform Guidelines on Employee Selection Procedures ("UGESP"), was adopted in 1978 by the Civil Service Commission, the Department of Labor, the Department of Justice, and the Equal Opportunity Commission in part to enforce the anti-employment discrimination sections of the 1964 Civil Rights Act.[48]

These guidelines extensively reference and justifies itself using the standards of academic psy-

[47] It is possible that SUTVA violations may arise if multiple test-passers were to make a single group decision about where to work together (or apart) as a group. For example, if Candidates i and j wanted to join the same firm and made decisions together, this could violate SUTVA. The candidates in this study applied individually to the employer via an online job application."Joint" offer-acceptance decisions are more common in merger or acquisition settings. It is impossible to know if this is happening in this dataset, but the author inquired with the recruiting staff if they knew of any "joint" offer acceptance decisions in this sample. The recruiters reported no known instances.

[48] The UGESP creates a set of uniform standards for employers throughout the economy around personnel selection procedures from the perspective of federal enforcement. The UGESP are not legislation or law; however, they provide highly influential guidance to the above enforcement agencies and been cited with deference in numerous judicial decisions.

chology.[49] No other profession or academic discipline is referenced at all in the UGESP, including economics.

The UGESP were adopted in 1978 and contains extensive statistical commentary about hiring criteria.[50] Since 1978, statistical practice in a number of social science fields has changed substantially (Angrist and Pischke, 2010). However, the UGESP have not been substantially revised and are still in use today.[51]

The methodology outlined in UGESP hiring criteria ignores the sample selection issues discussed in (Heckman, 1979). Thee guidance asks for comparisons candidates selected by both screening methods methods (the intersection, or candidates admitted by both old and new criteria) against those only passing the old criteria.[52] This is common in psychology papers studying similar topics.[53] The guidelines explicitly state that no experiment or intervention is required for measurement purposes.[54]

This research papers – and human resource practitioners following government guidance for compliance purposes – usually feature a dataset containing performance outcomes only on hired workers, without experimental variation in who is hired.

By contrast the candidates selected both by A and B – those who play a central role in the evaluation of the UGESP and the I/O psych literature – are irrelevant to evaluating a potential shift of A vs B, because such candidates would be admitted under both policies. In some cases, researchers have justified the above approach using an assumption of linearity or monotonicity, but

[49] For example, the UGESP requires that assessment tests that are "consistent with professional standards," and offers "the A.P.A. Standards" (an American Psychological Association book called **(alias?)**) as the embodiment of professional standards.

[50] For example, the requirement that the "relationship between performance on the [job] procedure and performance on the criterion measure [test] is statistically significant at the 0.05 level of significance[.]"

[51] They can be accessed at `https://www.gpo.gov/fdsys/pkg/CFR-2014-title29-vol4/xml/CFR-2014-title29-vol4-part1607.xml` (last accessed December 5, 2016).

[52] Using my notation, the I/O psych approach compares performance outcomes between $A \cap B$ candidates to $A \setminus B$.

[53] For example: A famous psychology paper (Dawes, 1971) shows that for psychology graduate students, a simple linear model more accurately predicts academic success than professors' ratings. In a followup paper, Dawes (1979) showed this result held, even when the linear predictor was misspecified.

Dawes interpreted this finding was to mean that linear predictors should be used in the graduate students' admissions. However McCauley (1991) showed that two decades after Dawes' finding, linear predictors were still not often used often not graduate student selection for PhD programs. This author's casual survey indicates this practice is still rare still rare in academic psychology as of this writing, but is gaining in popularity in the corporate world as discussed in Section 3.1.

A closer reading of Dawes (1971) shows the sample consists only of *matriculated* graduate students at one University. For reasons studied in Heckman (1979), Dawes' correlations within a selected sample may not generalize to the entire applicant pool. The direction of the bias cannot be signed, and the true correlation may not even have the same direction as in the selected sample.

An experiment would be necessary to measure the causal effect of changing selection criteria on ultimate graduate student achievement. Despite the popularity of Dawes' 1971 finding, no one to date has performed this experiment, or attempted to address the potential bias via another source of identification.

[54] The UGESP specifically clarifies it does not require an experiment, or another intervention that would sample from outside the firm's status quo, in order to evaluate a particular hiring method: "These guidelines do not require a user to hire or promote persons for the purpose of making it possible to conduct a criterion-related study." (Section 14B). Besides an experiment, one way to avoid this issues is to test all applicants. Within the economics literature, Pallais and Sands (2016) used the strategy of hiring all applicants to a job opening in her study of referrals in hiring for routine cognitive tasks (basic computations and data entry) on oDesk.

this assumption is rarely empirically tested or made explicit.

F Comparison to Dawes

Table 14: Dawes Approach vs. Experimentally-derived Estimates

	Job Offer	Accept Offer
Dawes-style Estimate	0.18***	-0.14**
	(0.011)	(0.061)
Experimentally-derived Estimates	0.31***	0.07***

Notes:

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

References

Angrist, Joshua D and Jörn-Steffen Pischke, "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics," *The Journal of Economic Perspectives*, 2010, *24* (2), 3–30.

Association, American Educational Research, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, and Psychological Testing (U.S.), *Standards for Educational and Psychological Testing*, American Psychological Association, 2014.

Commission, Equal Employment Opportunity et al., "Uniform Guidelines on Employee Selection Procedures," *Federal register*, 1978, *43* (166), 38295–38309.

Cowgill, Bo, "Competition and Productivity in Employee Promotion Contests," 2016.

Dawes, Robyn M, "A case study of graduate admissions: Application of three principles of human decision making.," *American psychologist*, 1971, *26* (2), 180.

_ , "The robust beauty of improper linear models in decision making.," *American psychologist*, 1979, *34* (7), 571.

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

Lazear, Edward P, Kathryn L Shaw, and Christopher T Stanton, "Who Gets Hired? The Importance of Finding an Open Slot," Technical Report, National Bureau of Economic Research 2016.

McCauley, Clark, "Selection of National Science Foundation Graduate Fellows: A case study of psychologists failing to apply what they know about decision making.," *American Psychologist*, 1991, *46* (12), 1287.

Pallais, Amanda and Emily Glassberg Sands, "Why the Referential Treatment? Evidence from Field Experiments on Referrals," *Journal of Political Economy*, 2016, *124* (6), 1793–1828.