The Econometrics of Gatekeeping Experiments

Bo Cowgill*
Columbia University

September 10, 2018

Working paper

Abstract

Gatekeeping is abundant in our economy and society. Hiring, university admissions, job promotion, occupational licensing and startup investment are examples of resources protected by selective gatekeeping policies. Gatekeeping is a common social application of machine learning. Some claim this new technology will make gatekeeping more fair, while others allege it will entrench bias. Evaluating gatekeeping policies is complicated; in many cases, gatekeeping policies generate outcomes visible only for subjects who are admitted, with outcomes missing for rejected candidates. In addition, changes in gatekeeping policy may leave most candidates unaffected. This challenges evaluators to isolate outcomes for the the subset of candidates whose admission status is changed by new policy. These issues complicate evaluation of any gatekeeping policy, including discretionary human gatekeepers.

This paper develops simple econometric methods for evaluating and comparing gatekeeping policies through field experiments. The methods are applicable in a wide variety empirical settings. Under the formal conditions outlined in the paper, these methods permit researchers to make causal statements about the effects of a new gatekeeping policy on characteristics of the admitted (and omitted) populations (such as their diversity and performance outcomes). They also permit clean, interpretable, causally identified estimates of the tradeoffs between a gatekeeper's welfare- or productivity- outcomes and its effects on representation and/or distributional outcomes. These tools do not require the gatekeeper to algorithmic or interpretable thus evaluation and comparisons across a wide range of gatekeeping policies and technologies, including gatekeepers based on human discretion or black-box methods. The resulting evaluations are interpretable estimates that satisfy some practitioners' need for transparency and explanation, but without constraining the underlying algorithms or methods.

*The author thinks Dan Gross, John Horton, Daniel Kahneman, John Morgan, Paul Oyer, Olivier Sibony and seminar participants at Columbia, Harvard, MIT, Stanford and the University of Pennsylvania for valuable feedback.

1 Introduction

Gatekeeping is abundant in our economy and society. Hiring, university admissions, job promotion, occupational licensing and startup investment are examples of resources protected by selective gatekeeping policies.

Many of the biggest questions about gatekeeping policies are causal. If a company uses a new hiring approach, will the change *cause* outcomes to be more biased or unbiased? More fair or unfair? If we use new screening procedures rather than the status quo, will this *cause* fewer women and minorities to be approved for loans?

Gatekeeping is an often-proposed application of machine learning and AI. Some claim this technology will make gatekeeping more fair, while others allege it will entrench bias. The answers to these questions are fundamentally causal.

Field experiments are the gold standard for causal inference. Opening a "FDA for Algorithms" is a commonly-suggested policy solution for algorithmic fairness (Tutt, 2016). Randomized controlled trials are a key component of FDA regulation, but few commentators have specified how this aspect of pharma policy would apply to algorithms.

Gatekeeping may have enormous distributional consequences, and are therefore often taboo as the topic of experiments. Oyer and Schaefer (2011)'s review of employment research notes that field experiments in hiring are rare, asking "What manager, after all, would allow an academic economist to experiment with the firm's screening, interviewing or hiring decisions?" Beyond taboos, such experiments may raise ethical problems for researchers, although some of these may be mitigated through careful experimental design.

Because of these issues, many researchers lack tools and intuition for measuring the causal effects of new gatekeeping. In an era when new technology (machine learning and AI) disrupts traditional gatekeeping, researchers' lack of tools and intuition invites unintended consequences.

This paper aims to fill this gap. In Section 2, I develop a potential outcomes framework for evaluating gatekeeping policies. In Section 3, I compare the potential outcomes approach to related methods. In Section 4, I introduce adjustments to measure effects of new gatekeeping policies on outcomes revealed only for candidates who are admitted.

In Section 5, I outline the formal assumptions for causal inference about all outcomes. Section 6 presents the estimand, its interpretation and relationship to discrimination and interpretability. Section 7 examines methods for measuring effects on "downstream outcomes," realized far past the initial admission decision. I conclude with a brief discussion in Section 8.

2 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)? In this section, I present a stylized potential outcomes framework similar to Neyman (1923/1990) and Rubin (1974, 2005), and apply this framework to gatekeeping to obtain causal estimates. For the exposition below, I will use generic gatekeeping

language wherever possible and use hiring examples for clarification.

First, I introduce some notation. Each observation is "candidate" for a gatekeeping decision, indexed by i. Many applications feature outcome metrics the gatekeeper aims to maximize or balance. For this exposition I will call the characteristic metric θ. θ can represent a characteristic such as performance if hired, or it can represent a demographic or other observable characteristic. Evaluations of gatekeeping may involve multiple such outcomes, and/or a single combination variable weighing several outcomes.

The approach below will cover many definitions of θ. As we will see, the challenging aspect of gatekeeping experiments typically arises when θ represents outcomes that are observable only for candidates who are selected (such as interview performance or on-the-job performance), and interactions with these variables ("minority candidates who pass the interviews," where interview performance is observable only for selected candidates).

Suppose we aim to compare the effects of adopting a new testing criteria, called Criteria B, against a status quo testing criteria called Criteria A. 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. For exposition, suppose that 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.

Researchers may want to evaluate A vs B on a number of dimensions (for example, whether A or B admit more women or minorities). However, a more challenging evaluation is to measure which criteria admits higher performing applicants (since performance is unobserved for non-admitted applicants) – and whether increases or decreases in the quality of applicants is worth changes in the composition of applicants.

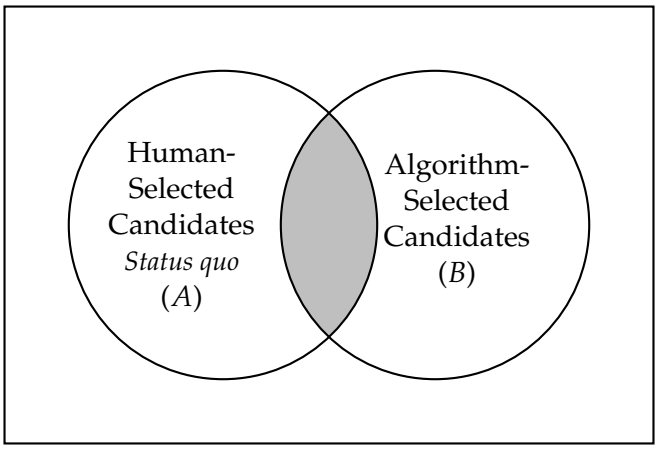
For any given candidate, $A_i = 1$ means that Criteria A suggests testing candidate i and $A_i = 0$ means Criteria A suggests *not* testing i (and similarly for $B = 1$ and $B = 0$). I will refer to $A = 1$ candidates as "A candidates" and $B = 1$ candidates as "B candidates." I'll refer to $A = 1$ & $B = 0$ 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 orientation.

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. Candidates who are rejected (or accepted) by both methods are irrelevant for determining which strategy is different or better.

As I discuss later in this paper, this approach differs from some others – including those arising from Industrial and Organizational Psychology literature and the EEOC's *Uniform Guidelines on Employee Selection Procedures*. As I discuss later in this paper, these result in in provably misleading estimates of the effects of B.

A strikingly few empirical examinations of gatekeeping attempt to characterize the zones of disagreement at all. For example: Even without algorithmic guidance from COMPAS, Florida judges may find many defendants guilty through their own analysis. Judges' unguided opinions may in fact be *more* biased than the COMPAS algorithm criticized by ProPublica. On which defendants did judges and COMPAS disagree? Few researchers have attempted to characterize the disagreement.

Figure 1: Visualization of Candidate Types



Notes: The above is a useful visualization of the decision. 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 $B \setminus A$).

3 Discussion of Related Methods

Before proceeding with implementation and technical aspects of the measurement strategy, I will briefly contrast the measurement approach in this paper – based on potential outcomes and field experiments – to methods used in other literatures. Several (Autor and Scarborough, 2008; Hoffman et al., 2016; Horton, 2013) studies make very similar causal inferences similar to the ones here about the effects of gatekeeping policies.

These papers implicitly use a potential outcomes framework similar to the one I describe here, but there are several differences. First, these papers often examine an entire firm, a division or job opening as a unit of analysis. An experiment of this scale may be difficult for a researcher or even a practitioner hoping to make a practical, business-oriented evaluation.

Setting aside logistics, persuasion and coordination necessary for such an experiment, a typical experiment of this nature may also lack statistical power. Standard errors in this experiment should be clustered at the firm/division/opening, and many researchers may not have access to many independent clusters (even if the number of job applications per cluster is high).

Second: The approach I describe focuses on describing and evaluating the *marginal* candidate who is admitted under one gatekeeping method and rejected by another. As I argue above, these population is critical for measuring the effects of the new gatekeeping policy and understanding its mechanisms.

Other methodological approaches compare *average* outcomes for candidates admitted by A to

average outcomes to candidates admitted by B. This approach pools both marginal and non-marginal candidates together. In theory, these should result in the same conclusion.

However, average outcomes are sometimes noisier than marginal outcomes. As a result, failing to isolate marginal candidates may lead to less precise estimates. This may raise uncertainty about the entire sign of the effect. In addition, a focus on of marginal candidates may be helpful for understanding mechanisms.

The approach outlined here differs more seriously with those used in the field of industrial and organizational psychology ("I/O psychology"). This research – and its methodological recommendations – have been very influential in law and public policy surrounding 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.[1]

These guidelines extensively reference and justifies itself using the standards of academic psychology.[2] No other profession or academic discipline is referenced in the UGESP at all.

The UGESP were adopted in 1978 and contains extensive statistical advice, including recommended p-value thresholds (rare for administrative laws).[3]

Since 1978, statistical practice has changed substantially. Within social science, a "credibility revolution" (Angrist and Pischke, 2010) occurred as a result of improved causal inference methods. The techniques behind this revolution are related to the methods in this paper, but were developed after the UGESP. However, the UGESP have not been substantially revised and are still in use today.[4] They make no reference to causal inference at all.

Policymakers in the public or private sectors often need to assess tradeoffs between the productivity benefits of a selection algorithm and the effects on diversity. In the language of Figure 1, the I/O Psych approach (and the UGESP) evaluates the diversity characteristics of gatekeeper B by measuring differences between B and $\setminus B$. The EEOC's infamous "four-fifths rule" (Feldman et al., 2015) is about B and $\setminus B$, even thought many candidates in both groups (B and $\setminus B$) would be selected (or rejected) in under alternative regimes as well (such as A, or others) – and are therefore not marginal to B.

For productivity impacts, the guidance asks for a different comparison. In particular, it asks for a comparison of 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. In

[1]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.

[2]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 Standards for Educational and Psychological Testing (2014)) as the embodiment of professional standards.

[3]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[.]"

[4]They can be accessed at `https://www.gpo.gov/fdsys/pkg/CFR-2014-title29-vol4/xml/CFR-2014-title29-vol4-part1607.xml` (last accessed August 22, 2018).

the notation of this paper, it compares performance outcomes between $A \cap B$ candidates to $A \setminus B$.

This is problematic for at least two reasons: First the productivity effects are being measured on a different populations ($A \cap B$ vs $A \setminus B$) than the diversity effects (B vs $\setminus B$).

Second: The UGESP evaluation does not require an evaluation of $B \setminus A$ – candidates who are currently rejected by A, but would be admitted under B – even though this is a potentially large source of diversity and productivity changes.

This is most likely because researchers – 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.

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 new 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).[5]

The UGESP thus stands in stark contrast to policy at the FDA, which *requires* counterfactual evaluation through randomized controlled trials.

By contrast, the candidates central in the UGESP and I/O psych approaches – those selected both by both new and old regimes – are irrelevant. Such candidates would be admitted under both policies.[6] The candidates whose shift effects outcomes the most – those admitted by one regim and not the other – are unobservable without the type of experiment this literature says are unnecessary.

These methodological issues are present in the academic I/O Psychology literature that undergird the UGESP. For example: A famous psychology paper by Dawes (1971) shows that for psychology graduate students, a simple linear model more accurately predicts academic success than professors' ratings.[7] This is a classic paper that echoes today's debates about machine and human judgement.

However, the sample in Dawes (1971) consists only of *matriculated* graduate students at one University under the system of professors rating. In his papers, Dawes did not make an effort to change the gatekeeping policy in to his department and evaluate the effects – as would be advocated by the potential outcomes approach.

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. 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

[5] 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.

[6] In some cases, researchers have justified the above approach using an assumption of linearity or monotonicity, but this assumption is rarely empirically tested or made explicit.

[7] In a followup paper, Dawes (1979) showed this result held, even when the linear predictor was misspecified.

writing.

In Appendix A, I show that ignoring these issues could result in either over- or under- estimate the true effect. As such, the method advocated in the UGESP and I/O psych papers does not deliver a useful boundary condition (such as an upper- or lower- bound) of the true effects.

4 When Outcomes Are Revealed Only for Admitted Candidates

If the researcher's data contains A, B and θ labels for all candidates, it would suffice to compare average differences in θs for $B \setminus A$ vs. $A \setminus B$. However, a common problem is for is for some problem θ variables to be unobserved at the time of the admission decision. For example, suppose that θ represents whether a candidate would pass a qualifications test. The test is expensive to administer and cannot be offered to everyone. Gatekeeping policy may reasonably want to maximize the proportion of admitted candidates who are able to pass the qualifications test. However, this ability may not be measurable at the time of the job application.

For every candidate, the potential outcomes are $Y_i = 1$ (passed the test) or $Y_i = 0$ (did not pass the test, possibly because the test was not given). Binary outcomes is used to simplify exposition. For each candidate i, the empirical researcher 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.

Often researchers do not know the full extent of disagreement between A and B because of *interference* between the two methods. The act of selecting a B candidate (say, by scheduling an interview because an algorithm liked the candidate), may pre-empt evaluation by A, making the candidate unavailable for an assessment by A (a human screener) after the candidate is removed from the stack of resumes in order to schedule an interview.

I propose a strategy for addressing these issues below using a generalized 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 $1/2$, for all candidates for whom $B_i = 1$. The

instrument must affect who is interviewed – for example, assume that firm admits or tests all candidates for whom $Z_i = 1$, irrespective of A_i.

In order to measure the marginal yield of Criteria B, we need variation in Z_i within $B_i = 1$. 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 $B = 0$. The instrument Z_i within $B_i = 1$ is "local" in that that it only varies for candidates approved by Criteria 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 instrumental variable ("IV") conditions (Angrist et al., 1996). I outline each condition in theory in the following Section 5, with some interpretation of these assumptions in a hiring or other gatekeeping setting.

5 IV Assumptions in a Candidate-Level Gatekeeping Field Experiment

The five IV assumptions, applied to a gatekeeping experiment, are below.

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

Ignorable assignment of Z. Z_i must be randomly assigned, or $0 < \Pr(Z_i = 1|X_i = x) = \Pr(Z_j = 1|X_j = x) < 1, \forall i, j, x$. This can be empirically tested through a covariate balance test, measuring that $Z_i = 1$ and $Z_i = 0$ do not differ on observables.

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. Test grading that is biased (with respect to A vs B) would violate the exclusion restriction.[8] Double-blind or objective evaluation may help meet the exclusion restriction, so that graders of the

[8] 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.

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. Ideally, the existence of the experiment and instrument are never disclosed to test graders or candidates, and the variable Z_i is hidden from subsequent evaluators.

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

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 $Cov(Z, T|X) \neq 0$). The instrument must have a non-zero effect on who is tested. This can be evaluated by demonstrating a strong correlation between Z_i and T_i, for example performing a $T-$ test or regression of T_i on Z_i.

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.

6 Estimand and Interpretation

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.

Firms that want to fix the quantities can run quantity-limiting experiments. Alternatively: If one Criteria is based on a rankable variable, a researcher can examine subsets of the data that limit analysis to the top N candidates selected by either mechanism.

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

$$\begin{aligned} \hat{\beta} &= E[Y|T = 1, A = 0, B = 1] \\ &= \frac{E[Y_i|Z_i = 1, B_i = 1] - E[Y_i|Z_i = 0, B_i = 1]}{E[T_i|Z_i = 1, B_i = 1] - E[T_i|Z_i = 0, B_i = 1]} \end{aligned} \tag{1}$$

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*." Recall that 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 – in $B \setminus A$.

We can now compare this marginal success rate (B) to other success rates in the applicant pool. In particular, we may compare the success rate of $B \setminus A$ to the overall success rate in $A \setminus B$, or in the

intersection area $A \cap B$. This measures the increase in quality between candidates in B but not A to candidates in A but not B. Switching gatekeeping policies from A to B would yield this amount of quality increase.

We can also use the differences in characteristics between the candidates in $B \setminus A$ and $A \setminus B$ to help explain the nature of the shift, and where differences in productivity come from. In machine learning applications, these explanations may be useful when policymakers or clients would like an interpretable reason for why an algorithm is making a decision (Doshi-Velez and Kim, 2017).

For example: Suppose $B \setminus A$ contained +15 percentage points more math majors than $A \setminus B$, +7 percentage points more knitting enthusiasts, and -8 percentage points fewer ex-employees from rival companies. And it yielded a quality difference of +10 percentage points. If a policymaker asks "why are we adopting this algorithm?" the answer can be "A randomized controlled trial showed an +10 increase in passing job skills tests. This came from selecting more math majors and knitting enthusiasts who were more likely to pass the test, and eliminating ex-employees from rival companies who were unlikely to pass – but were being tested in our previous policy."

It is possible to estimate the coefficient in Equation 1 separately for each subgroup within the dataset, so that a policymaker know: For the marginal candidate approved by B, rejected by A and has characteristic X (black, asian, female, literature major, PHD holder, etc), what is the average success rate of that candidate? This would give a policymaker very fine-grained explanations of why the algorithm was being adopted, and what mechanisms were responsible for the improved outcomes.[9]

This quality comparison can be combined with other variables to estimates clean, well-identified tradeoffs between the quality of candidates and their diversity. In many cases there may be no such tradeoff at all. However: Suppose the quality difference in $B \setminus A$ versus $A \setminus B$ was +10 percentage points. However, in some cases there may be tradeoffs between passing the test and representation. For example, $B \setminus A$ may increase the quality of tested candidates by +10 percentage points, but decreased hiring of non-traditional candidates by 1 percentage point.

A policymaker could place different weights in how much she values increased quality and representation, and determine if the tradeoffs are worthwhile on the A vs B margin. Because these estimates come from experiments in which productivities or test outcomes are realized, they enable policymakers to assess between *realized productivity outcomes* from an experiment, rather than a ML model's "predictive accuracy," and diversity. Measuring tradeoffs with "predictive accuracy" requires the researcher to believe that productivity correlations within the status quo regime (A) extrapolate into new candidates B about which there is no available historical data (because the firm never hired them).

In many cases, firms may care 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).

These downstream outcomes present similar challenges: Some downstream outcomes are only

[9]It may come short of deliver case-by-case explanations to all applications, which is sometimes a requirement for interpretable systems.

realized for admitted candidates who make it several steps downstream (pass test, accept offer and start work). Policy changes are only relevant for candidates about whom there is disagreement between A and B. In addition, policymakers may require explanations of why new policies are adopted that effect downstream outcomes, and what interpretable mechanisms are responsible for these effects. The policymakers may require well-identified estimates of the tradeoffs between downstream productivity outcomes and representation.

Next, I show how to extend this framework further into the production function to measure the effects on these downstream outcomes after early stage testing acceptance.

7 Extension: Downstream Outcomes

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.[10]

We can think of these extensions as a form of imperfect compliance with the instrument. As the researchers 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.

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

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.

7.1 Revisiting IV assumptions for "downstream" hiring outcomes

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. Even if cross-candidate comparisons were absent from test-grading, they might re-appear downstream in offer-acceptances. For example: 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. This may happen even if candidates are not graded on a curve in the initial qualifications test.

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.

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. "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. 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.

Exclusion restriction, or $Y(Z,T) = Y(Z',T), \forall Z, Z', T$. Downstream outcomes may violate the exclusion restrictions even if it it passed the original testing outcome. As previously discussed, the best way to ensure the exclusion restriction is met is to blind evaluation. Although blind evaluation may be feasible for early evaluation, it may be less feasible for downstream outcomes such as promotions – particularly if it becomes publicly known who was admitted under which gatekeeping regime.

8 Discussion and Conclusion

Gatekeeping – in setting such as university admission, hiring, promotion, occupational licensing, lending and startup investments – is a commonly proposed application of machine learning. Advocates of this method claim that new technology will make gatekeeping more fair and objective, however many worry that automation will entrench bias.

This paper has described experimental design and evaluation methods for examining the impact of gatekeeping policies. Evaluating gatekeeping policies is complicated; in many cases, gatekeeping policies generate outcomes visible only for subjects who are admitted, with outcomes missing for rejected candidates. Changes in gatekeeping policy may leave most candidates unaffected. In addition, many policymakers want interpretable explanations of new policies, and interpretable ways to assess the tradeoffs of new gatekeeping policies along a number of dimensions.

Claims about algorithmic fairness are necessarily causal – they imply that an algorithm has caused an unfair impact. They are also inherently *marginal*, because many changes to gatekeeping will leave most applicants unaffected.

This paper develops simple econometric methods for evaluating and comparing gatekeeping policies through a field experiments (the gold standard in causal inference). Unlike the methods in the I/O Psych literature and the UGESP, the methods evaluate effects on diversity and productivity using the same group of candidates. They also focus on marginal candidates – ones who would be rejected or accepted depending on which method is used.

These tools do not require the gatekeeper to algorithmic or interpretable thus evaluation and comparisons across a wide range of gatekeeping policies and technologies, including gatekeepers based on human discretion or black-box methods. Under the formal conditions outlined in the paper, the methods permit clean, interpretable, causally identified estimates of the tradeoffs between a gatekeeper's welfare- or productivity- outcomes and its effects on representation and/or distributional outcomes. The resulting evaluations are interpretable estimates that satisfy many practitioners' need for transparency and explanation, but without constraining the underlying algorithms or methods.

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.

_ , 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.

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.

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

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.

Doshi-Velez, Finale and Been Kim, "Towards a rigorous science of interpretable machine learning," *Working paper*, 2017.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, "Certifying and removing disparate impact," in "Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" ACM 2015, pp. 259–268.

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.

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.

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.

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.

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.

Tutt, Andrew, "An FDA for algorithms," 2016.

A Over and Underestimation in the UGESP Procedure

Overestimation: The $A \cap B$ candidates have passed both Criteria. If both Criteria A and B have some merit, then the $A \cap B$ candidates contains "superstar" who were able to pass both stan dards. Comparing these superstars to candidates who only passed one test (A) may thus overestimate the benefit of switching to B; much of the effect of switching policies would come from selecting B candidates who only passed one test. A fairer approach (advocated in this paper) is to make comparisons between candidates who passed exactly one test. However, this method would require testing (or hiri ng) candidates who wouldn't otherwise be tested.

Underestimation: It's also possible that the I/O psych method could understate, rather than overstate, the benefit of a new method. Suppose that Criterial B identifies lots of high-performing cand idates who were previously not identified at all. In this case, most of the benefit of adopting B would come from these new candidates. Candidates who passed both requirements (A and B) may not per form as well as purely B-only ones. This would mean that the I/O psych analysis understates the benefit of B.