Adaptive Correspondence Experiments

By Hadar Avivi, Patrick Kline, Evan Rose and Christopher Walters*

A large social science literature uses correspondence experiments to measure discrimination (Bertrand and Duflo, 2017; Baert, 2018). Kline and Walters (Forthcoming) argue that experiments sending multiple applications to each job can be used to reliably detect discrimination by particular employers. A practical impediment to such exercises is that employer callback rates are often very low, leading a large fraction of applications to be wasted on unresponsive jobs. Sending many applications to a particular job may also compromise the callback evidence by alerting an employer to the experiment.

This paper considers the potential for dynamic correspondence experiments to reduce the costs of detecting discrimination by particular employers. We consider an experimental design in which the researcher adapts the number and characteristics of applications sent to each job in response to prior callback outcomes. Our analysis is inspired by proposals in the medical sciences to personalize treatments based on time-varying health information (Brown et al., 2009; Chakraborty and Murphy, 2014) as well as econometric procedures that update estimators, decision rules, and experimental designs in response to new data (Dimakopoulou et al., 2018; Kasy and Sautmann, Forthcoming; Tabord-Meehan, 2020). In the discrimination context, adaptive experimentation provides a potential tool for regulatory agencies such as the Equal Employment Opportunity Commission (EEOC), which are charged with preventing and remedying discrimination by individual employers in the labor market.

We begin by building a statistical model of job callback decisions, which we fit to data from a recent correspondence study by Nunley et al. (2015) (henceforth, NPRS), who submitted four applications with distinctively black or white names to each of 2,305 entry-level jobs for new college graduates in the US. One can imagine such training data coming from a pilot study commissioned by the EEOC. We then ask how an auditor who has learned the distribution of discrimination from this pilot might send applications to new vacancies to find discriminatory jobs at minimum cost.

Simulating the performance of the optimal auditing strategy, we find that adaptive experiments can cut the number of applications needed to detect a fixed number of discriminators by more than half without increasing the prevalence of type I errors. This feat is accomplished primarily by giving up early on jobs with very low callback rates and those that demonstrate a willingness to call black applicants. These results reinforce the conclusion of Kline and Walters (Forthcoming) that correspondence experiments can potentially be used by regulatory agencies to target investigations more efficiently.

I. A model of callbacks

Following Kline and Walters (Forthcoming), we model callbacks at each job as independent Bernoulli trials. A fictitious applicant of race $r \in \{w, b\}$ (white or black) possessing observable characteristics x has a callback probability $p_{jr}(x)$ of being called back by job j. We assume that

$$p_{jr}(x) = \Lambda\left(\alpha_j - \beta_j 1\{r = b\} + x'\gamma\right),$$

where $\Lambda(z) \equiv [1 + \exp(-z)]^{-1}$ is the logit link function, the parameter α_j governs the white callback rate, β_j governs the call-

* Avivi: UC Berkeley, 530 Evans Hall #3880, Berkeley, CA, 94720, havivi@berkeley.edu. Kline: UC Berkeley, 530 Evans Hall #3880, Berkeley, CA, 94720, pkline@econ.berkeley.edu. Rose: Microsoft Research, 1 Memorial Dr, Cambridge, MA 02142, ekrose@gmail.com. Walters: UC Berkeley, 530 Evans Hall #3880, Berkeley, CA, 94720, crwalters@econ.berkeley.edu.

Table 1— Mixed logit censored-normal results, Nunley et al. (2015) data

	(1)	(2)
α_0	-4.922	-4.918
	(0.234)	(0.234)
σ_α	4.968	4.963
	(0.240)	(0.240)
β_0	-5.035	-5.022
	(0.176)	(0.329)
σ_β	6.347	6.521
	(0.148)	(0.154)
ρ		-0.013
		(0.017)
Likelihood	-2788.3	-2788.3
Number of jobs	2305	2305

Note: This table presents simulated maximum likelihood estimates of the mixed logit censored-normal model in the Nunley et al. (2015) data. Models also include demeaned resume covariates: gender, industry, high socioeconomic status indicator, work history gaps, business degree, internship experience and GPA. Robust standard errors in parentheses.

back penalty for being perceived as black, and $x'\gamma$ is an application quality index. We treat the parameters (α_j, β_j) as random draws from a bivariate distribution. Kline and Walters (Forthcoming) were unable to reject the hypothesis that white names are never discriminated against in the NPRS experiment. We therefore assume that $\beta_j = \max\left\{0, \tilde{\beta}_j\right\}$, which censors discrimination from below at zero. The model is completed by the following distributional assumption:

$$\begin{pmatrix} \alpha_j \\ \tilde{\beta}_j \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \rho \\ \rho & \sigma_\beta^2 \end{bmatrix} \right), \tag{1}$$

which allows for continuous heterogeneity in overall callback rates and discrimination severity, as well as a positive mass of jobs that do not discriminate at all.

Maximum likelihood estimates of this model are presented in Table 1. We cannot reject that $\rho = 0$, and therefore impose this restriction in what follows. Using the estimates in column 1, the share of jobs with white callback probabilities lower than 1% is $\Phi\left(\frac{4.922-\ln(99)}{4.968}\right) \approx 0.53$, suggesting that a majority of the jobs in this study are essentially unresponsive to applications of either race. The share of jobs with $\beta_j = 0$ is $\Phi\left(\frac{5.035}{6.347}\right) \approx 0.79$. Hence, discrimination is confined to a small minority of jobs. However, the average severity of discrimination among this minority is intense, as $\mathbb{E}\left[\beta_j|\beta_j > 0\right] = \beta_0 + \sigma_\beta \frac{\phi(\beta_0/\sigma_\beta)}{\Phi(\beta_0/\sigma_\beta)} \approx 3.6$, which implies the odds of being called back by a discriminating job are roughly 36 times higher for whites than equivalent blacks. This finding of rare but intense discrimination accords closely with the non-parametric analysis of the NPRS data in Kline and Walters (Forthcoming) and suggests it may be possible to ascertain whether or not a *responsive* job is discriminating with very few applications.[1]

II. The auditor's problem

Consider now a hypothetical auditor who knows the parameters of Table 1 and can draw additional jobs from the same distribution. Her goal is to find as many discriminating jobs as possible by sending fictitious applications. The auditor may send up to 8 applications to each job and is free to choose the race and quality q of each application. To simplify the problem, we coarsen applicant quality to two levels (labeled "high" and "low") that correspond to setting the covariate index $x'\gamma$ to one standard deviation above or below its estimated mean. If the auditor believes a job is discriminating she can initiate an investigation. Once an investigation is initiated, the job's true type is revealed and the auditor receives a payoff

$$\underbrace{\frac{1}{2} \sum_{q \in \{h,l\}} \left[p_{jw}(q) - p_{jb}(q)\right]}_{\equiv S_j} - \kappa, \tag{2}$$

where S_j gives the average severity of discrimination at job j and κ is the cost of conducting an investigation. Choosing not to investigate yields a payoff of zero. Hence, an auditor with $\kappa = .01$ is indifferent about investigating a job that contacts black ap-

[1] Kline and Walters (Forthcoming) find that at least 17% of jobs in the NPRS data are discriminating and that the distribution of discriminatory severity is highly variable and skewed.

plicants a percentage point less often than comparable white applicants. The linear formulation in (2) implies that the auditor cares about the expected number of callbacks lost to racial discrimination, which reflects both discrimination severity and the baseline callback rate for white applicants. Severity is unknown and must be assessed by sending job applications.

Rather than send 8 applications to each job, the auditor solves a sequential problem: in each period, she may send an application to a job and choose its race and quality, launch an investigation, or give up. The auditing history H_n encodes the assigned race and observable characteristics of the n prior applications sent to a given job along with whether each application was called back. The auditor's value function can be written

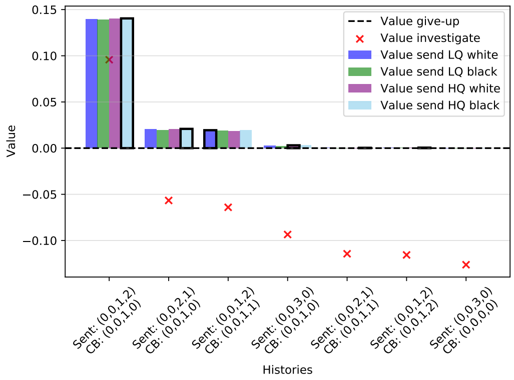
$$\mathrm{V}(H_n)=\begin{cases} \max\left\{\underbrace{\max_{r,q} v_{rq}\left(H_n\right)}_{\text{send new app}}, \underbrace{v_I\left(H_n\right)}_{\text{investigate}}, 0\right\} & \text{if } n<8 \\ \max\left\{\underbrace{v_I\left(H_n\right)}_{\text{investigate}}, 0\right\} & \text{if } n=8, \end{cases}$$

where $v_{rq}(H_n) = -c + \mathbb{E}[V(H_{n+1})|H_n]$ is the expected value of sending an application of race r and quality q net of the cost c of sending a new application. The value of giving up on a job equals zero, while the auditor's expected payoff from investigating is $v_I(H_n) = \mathbb{E}[S_j|H_n] - \kappa$. All expectations are evaluated via Bayes' rule starting from a prior distribution based on the parameters in column 1 of Table 1. As Kline and Walters (Forthcoming) emphasize, the use of Bayes' rule allows the auditor to borrow strength from the experience of the pilot study, which can generate informative conclusions about a particular job even when very few applications have been sent.

III. Optimal auditing policy

The solution to the sequential problem was computed numerically by backwards induction. Figure 1 shows the decision problem that arises after 3 applications when setting $\kappa = .13$ and $c = 10^{-4}$. Seven distinct values of H_3 arise under optimal auditing, which we have ordered by the auditor's posterior expectation of discrimination severity. If sending another application is optimal, the bar representing the value of that action is bolded.

Figure 1. The auditor's expected value and optimal strategy after sending three applications ($\kappa = .13$ $c = 10^{-4}$)



Note: This figure depicts the expected value associated with each possible action given a job history H_3. Job history is characterised by the number applications of each of race and quality level that are sent and the number that are called back (CB). On the horizontal axis, the values in parentheses are ordered as follows: (LQ white, LQ black, HQ white, HQ black). If sending another application is the optimal action, the bar representing the value of that action is bolded.

Least suspicious is the case where 3 high-quality white applications are sent and none are called back. To ascertain whether a job is responsive, the auditor always begins by sending high-quality white applications. The estimates in Table 1 imply that 72% of jobs will fail to call back any of the first 3 applications, at which point the auditor gives up. Doing so this early saves the auditor $\frac{0.72 \times 5}{8} \times 100 = 45\%$ of the maximum possible number of applications.

If the auditor's first high-quality white application is called back, her optimal response is to begin sending high-quality black applications. Receiving no callbacks for any high-quality black applications produces the most suspicious configuration in Figure 1. Because the posterior probability that the job is discriminating is only about 18.8% in this scenario, the auditor decides to send another high quality black application to obtain greater certainty regarding the severity of any discrimination present.

Receiving one callback for a high-quality black application produces the third most suspicious configuration in Figure 1. While the callback to the black application makes it unlikely that this job is discriminating, the evidence could also be rationalized by a very high baseline white callback probability ($\alpha_j \gg 0$) in conjunction with nontrivial discrimination severity ($S_j > \kappa$). To assess this possibility, the auditor opts to send a low quality white application. Similar logic applies to the auditor's choices for additional histories, the full set of which appear in Figure A1 in the appendix.

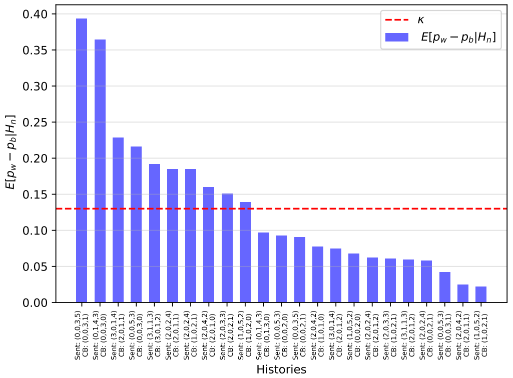
Once 8 applications have been sent, the auditor can no longer send additional applications and the decision problem simplifies to a binary choice: either launch an investigation or give up. Figure 2 shows the auditor's posterior expectation $\mathbb{E}[S_j|H_8]$ of the job's discrimination severity given its callback history H_8. The relevant values of H_8 have been depicted by the 23 distinct callback configurations that might arise under optimal auditing. Expected discrimination severity is maximized when 3 high quality white and 5 high quality black applications have been sent and all 3 white applications but only 1 black application have been called back.[2] A horizontal line gives the investigation cost κ. When the expected benefit exceeds κ, an investigation is launched. Otherwise, the auditor gives up.

IV. The gains from sequential auditing

To evaluate the performance of various auditing strategies, we borrow concepts from the medical literature on diagnostic testing. The *sensitivity* of an auditing strategy is the probability that an investigation is launched when a job is engaged in discrimination (i.e., when $\beta_j > 0$). The *specificity* of an auditing strategy is the probability that an investigation is not launched when a job is not discriminating (i.e., when $\beta_j = 0$).

Figure 3 plots the average number of applications per job associated with a given auditing strategy against its sensitivity when jobs are drawn randomly from the data generating process described in column 1 of Table 1. To enumerate auditing strategies, the parameters (κ, c) have been varied over a range that results in the probability of investigation falling in the interval $[.055, .06]$; each dot, therefore, corresponds to a different strategy that results in roughly the same expected number of investigations.

Figure 2. The auditor's value after sending eight applications ($\kappa = .13$, $c = 10^{-4}$)

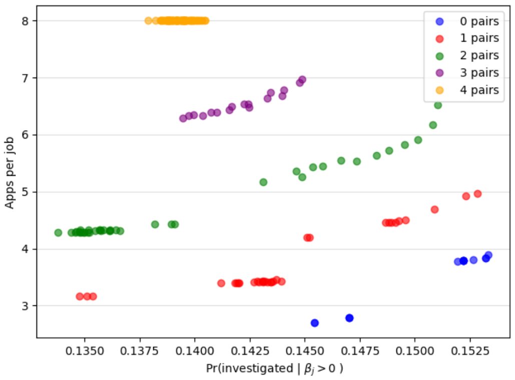


Note: This figure depicts the value of investigation given a job history H_8. Job history is characterized by the number of applications of each race and quality level that are sent and the number that are called back (CB). On the horizontal axis, the values in parentheses are ordered as follows: (LQ white, LQ black, HQ white, HQ black).

To illustrate the gains from sequential optimization, we include strategies where the auditor's discretion is constrained by requiring her to send a fixed number of mixed race application pairs (of random quality) before behaving optimally. Hence, the case where 4 pairs must be sent corresponds to a static auditing strategy, where the auditor must either decide to investigate each job or give up. As the auditor is given more discretion, the number of applications sent to each job falls. For example, when the auditor is allowed to optimize after sending 3 pairs, it is possible to maintain a sensitivity of 14% by sending an average of only 6.3 applications to each job. A fully unconstrained (0 pairs) auditor can achieve a sensitivity of 14.5% while sending fewer than 3 applications per job. In sum, by giving up

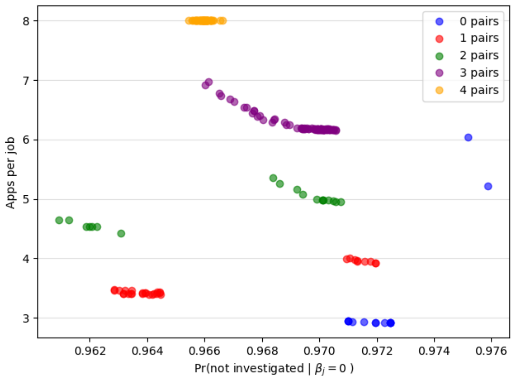
[2] All histories with zero black callbacks lead the auditor to investigate or give up before $n = 8$.

Figure 3. Applications sent vs. sensitivity (investigation probability fixed)



Note: This figure plots the average number of applications sent per job against the sensitivity of the auditing strategy for strategies that result in 5.5-6% of jobs investigated. The number of initial pairs refers to the number of applications of each race that are sent before the auditor begins optimizing. The curves are generated by varying κ between 0.01 and 0.09 and c between 10^{-5} and 0.004.

Figure 4. Applications sent vs. specificity (sensitivity fixed)



Note: This figure plots the average number of applications per job against the specificity of the auditing strategy for strategies where 14-14.5% of discriminating jobs are investigated. The number of initial pairs refers to the number of applications of each race that are sent before the auditor begins optimizing. The curves are generated by varying κ between 0.01 and 0.09 and c between 10^{-5} and 0.004.

on unresponsive jobs and jobs that call back some black applicants in early trials, the dynamic auditor is able to correctly identify more discriminators than the static auditor with fewer than half as many applications.

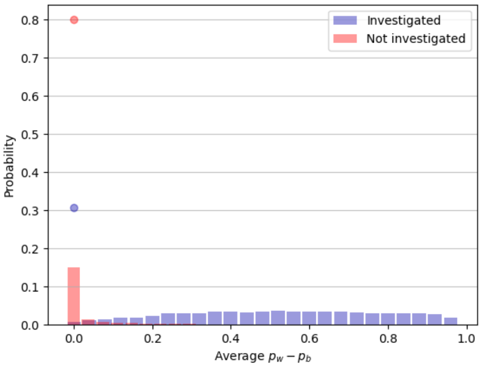
Figure 3 held the probability of investigated fixed, which can be thought of as approximating a setting where the auditor faces an ex-ante budget constraint on the expected number of investigations to be conducted. Figure 4 depicts an analogous exercise fixing the sensitivity of the auditing strategy in the interval [.14,.145] but allowing the marginal probability of an investigation to vary. This scenario can be thought of as one where the auditor plans to continue to experiment until a desired number of discriminators are investigated. Because investigations of non-discriminating jobs constitute type I errors, points in the Southeast quadrant of Figure 4 are preferable. Evidently, an unconstrained auditor (0 pairs) can achieve the same sensitivity as a static auditor (4 pairs) while incurring fewer false positives and utilizing less than half as many applications.

Figure 5 provides a histogram of discrimination severity for investigated and non-investigated jobs for an auditor with cost parameters of $\kappa = .13$ and $c = 10^{-4}$. The unconstrained auditor investigates only 3.5% of the jobs, but the jobs she investigates tend to be heavy discriminators. Roughly 30% of investigated jobs are not engaged in any discrimination, far below the 79% prevalence of non-discrimination in the overall population. Given her low investigation rate, the auditor's type I error rate evaluates to only $\frac{0.3 \times 0.035}{0.79} \times 100 = 1.3\%$. Discrimination tends to be much less severe among jobs that the auditor chooses not to investigate: 80% of the jobs not investigated are not engaged in discrimination at all, 15% engage in negligible discrimination with $S_j \in (0, 0.03]$, and the remainder exhibit only mild racial gaps in callback rates.

V. Conclusion

Correspondence experiments are a widespread tool used throughout the social sciences to detect discrimination in many contexts. While such experiments typically send a predetermined number of applications to each job, substantial cost reductions can, in principle, be achieved by giving up on unresponsive jobs. Our analysis suggests that adaptive correspon-

Figure 5. Discrimination severity by investigation status (0 pairs)



Note: This figure depicts the conditional probability mass function of S_j given investigation status for a dynamic auditor with payoff parameters $\kappa = 0.13$ and $c = 10^{-4}$. Dots show fractions of jobs with $S_j = 0$..

dence experiments can yield substantial reductions in the number of applications per job needed to achieve a desired level of sensitivity and specificity of investigative decisions.

Our analysis was predicated on the existence of a pilot study from which the distribution of unit heterogeneity could be learned in a first step. While dividing the problem into separate "exploration" and "exploitation" steps simplified our analysis considerably, in practice it may be desirable to combine these steps to reduce the costs of making an initial determination of market-wide discrimination. Likewise, updating estimates of the job heterogeneity distribution may be important for enforcement if one is worried that the parameters of the callback process are drifting over the course of a study, perhaps because of the enforcement activities themselves. An interesting topic for future research is the potential for reinforcement learning techniques (e.g., Kasy and Sautmann, Forthcoming) to balance the exploration and exploitation goals of correspondence experiment design without relying on parametric assumptions regarding the job callback process.

REFERENCES

Baert, Stijn. 2018. "Hiring discrimination: an overview of (almost) all correspondence experiments since 2005." In *Audit studies: Behind the scenes with theory, method, and nuance.* 63–77. Springer.

Bertrand, Marianne, and Esther Duflo. 2017. "Field experiments on discrimination." In *Handbook of Field Experiments.* Vol. 1, , ed. Esther Duflo and Abhijit Banerjee. Elsevier.

Brown, C Hendricks, Thomas R Ten Have, Booil Jo, Getachew Dagne, Peter A Wyman, Bengt Muthén, and Robert D Gibbons. 2009. "Adaptive designs for randomized trials in public health." *Annual review of public health*, 30: 1–25.

Chakraborty, Bibhas, and Susan A Murphy. 2014. "Dynamic treatment regimes." *Annual review of statistics and its application*, 1: 447–464.

Dimakopoulou, Maria, Zhengyuan Zhou, Susan Athey, and Guido Imbens. 2018. "Estimation considerations in contextual bandits."

Kasy, Maximilian, and Anja Sautmann. Forthcoming. "Adaptive treatment assignment in experiments for policy choice." *Econometrica.*

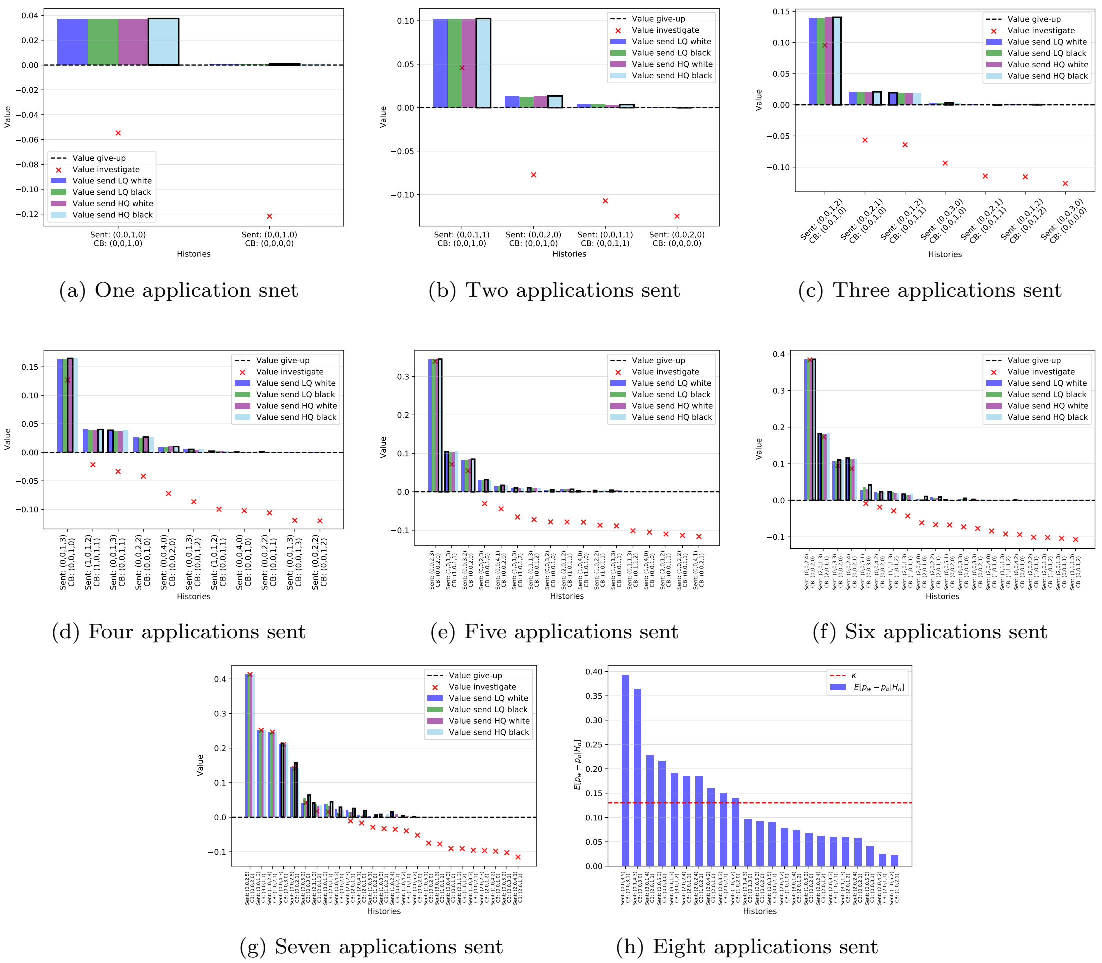
Kline, Patrick M, and Christopher R Walters. Forthcoming. "Reasonable doubt: Experimental detection of job-level employment discrimination." *Econometrica.*

Nunley, John M, Adam Pugh, Nicholas Romero, and R Alan Seals. 2015. "Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment." *The BE Journal of Economic Analysis & Policy*, 15(3): 1093–1125.

Tabord-Meehan, Max. 2020. "Stratification trees for adaptive randomization in randomized controlled trials."

Appendix

Figure A1. The auditor's expected value and optimal strategy, $\kappa = .13$ $c = 10^{-4}$



(a) One application snet

(b) Two applications sent

(c) Three applications sent

(d) Four applications sent

(e) Five applications sent

(f) Six applications sent

(g) Seven applications sent

(h) Eight applications sent

Note: This figures presents the auditor's expected value from each possible action given a job history H_n. Job history is characterised by the number of applications of each race and quality level that are sent and called-back. On the horizontal axis, the values in parentheses are ordered as follows: (LQ white, LQ black, HQ white, HQ black). If sending another application is the optimal action, the bar representing the value of that action is bolded. Expectations are formed assuming the data was generated from the censored logit model in column (1) of Table (1).