NBER WORKING PAPER SERIES

HELP REALLY WANTED? THE IMPACT OF AGE STEREOTYPES IN JOB ADS ON APPLICATIONS FROM OLDER WORKERS

Ian Burn
Daniel Firoozi
Daniel Ladd
David Neumark

Working Paper 30287
http://www.nber.org/papers/w30287

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2022

This research was funded by the Sloan Foundation. We received helpful comments from Joanna Lahey and from seminar participants at Cardiff University, the University of Kentucky, Texas A&M, and the 2022 NBER Conference on the Labor Market for Older Workers. Any views or opinions expressed are solely those of the authors. The Pre-Analysis Plan for this project was registered on Open Science Framework on December 31, 2020. The research was approved by the UCI Office of Research Institutional Review Board on October 18, 2019: HS# 2015-2017, modification application #26404. The pre-analysis plan is available at https://osf.io/37yaq/. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2022 by Ian Burn, Daniel Firoozi, Daniel Ladd, and David Neumark. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Help Really Wanted? The Impact of Age Stereotypes in Job Ads on Applications from Older Workers
Ian Burn, Daniel Firoozi, Daniel Ladd, and David Neumark
NBER Working Paper No. 30287
July 2022
JEL No. J14,J6,J7,J78

ABSTRACT

Correspondence studies have found evidence of age discrimination in callback rates for older workers, but less is known about whether job advertisements can themselves shape the age composition of the applicant pool. We construct job ads for administrative assistant, retail, and security guard jobs, using language from real job ads collected in a prior large-scale correspondence study (Neumark et al., 2019a). We modify the job-ad language to randomly vary whether or not the job ad includes ageist language regarding age-related stereotypes. Our main analysis relies on machine learning methods to design job ads based on the semantic similarity between phrases in job ads and age-related stereotypes. In contrast to a correspondence study in which job searchers are artificial and researchers study the responses of real employers, in our research the job ads are artificial and we study the responses of real job searchers.

We find that job-ad language related to ageist stereotypes, even when the language is not blatantly or specifically age-related, deters older workers from applying for jobs. The change in the age distribution of applicants is large, with significant declines in the average and median age, the 75th percentile of the age distribution, and the share of applicants over 40. Based on these estimates and those from the correspondence study, and the fact that we use real-world ageist job-ad language, we conclude that job-ad language that deters older workers from applying for jobs can have roughly as large an impact on hiring of older workers as direct age discrimination in hiring.

Ian Burn
University of Liverpool
Department of Economics
Chatham St, Liverpool L69 7ZH
United Kingdom
ian.burn@liverpool.ac.uk

Daniel Firoozi
Department of Economics
University of California, Irvine
3151 Social Sciences Plaza
Irvine, CA 92697
dfiroozi@uci.edu

Daniel Ladd
Department of Economics
University of California, Irvine
3151 Social Sciences Plaza
Irvine, Unit 92697
dladd@uci.edu

David Neumark
Department of Economics
University of California, Irvine
3151 Social Science Plaza
Irvine, CA 92697
and NBER
dneumark@uci.edu

Introduction

Lengthening work lives for those able to work is a crucial part of the policy response to population aging. Because many seniors transition to "partial retirement" or "bridge jobs" at the end of their careers (Cahill et al., 2006; Johnson et al., 2009) or return to work after a period of retirement (Maestas, 2010), reducing age discrimination in hiring is critical to lengthening working lives. There is an extensive body of research that documents the extent to which employers discriminate against older workers in hiring, using correspondence studies (e.g., Bendick et al., 1997, 1999; Lahey, 2008; Farber et al., 2019; Neumark et al., 2019a, 2019b). This research focuses on measuring employer behavior – specifically, whether there is less hiring of qualified older workers – and generally finds evidence consistent with hiring discrimination against older workers. There is little work, however, that studies how workers respond to manifestations of age discrimination in the labor market, including steps employers may take to discourage older workers from applying for jobs.

In this study, we create a bank of job ads for administrative assistant, retail sales, and security guard jobs.[1] We construct these job ads using language from real job ads collected in Neumark et al. (2019a). We post these job ads while randomly varying whether or not the text includes language that is rated as semantically similar to ageist stereotypes using machine learning, and that older workers perceive as age biased. We focus on stereotypes related to communication skills, physical ability, and technology skills. The innovation in this study is that the job ads are artificial, and we are studying the responses of real job searchers – in contrast to a correspondence study in which the job searchers are artificial and researchers study the responses of real employers.[2]

The potential for age stereotypes or other language in job ads to deter applications from older job seekers is real. An extreme version of such language is stating maximum experience levels in job ads. This occurred recently in *Kleber v. Carefusion Corp.*, where the job ad requested "3 to 7 years (no more than 7 years) of relevant legal experience," language that will clearly act to exclude many older applicants.[3] More generally, the U.S. Code of Federal

[1] These jobs have relatively high hiring of older workers.

[2] We obtained IRB approval to post these ads, subject to an IRB-approved protocol to do two things: (i) to quickly inform applicants that the job is not available, so as not to interrupt their job search; and (ii) subsequently, to inform those from whom we have collected data of their inclusion in an experiment, and allow them to opt out of their data being used. (This is the standard protocol in experiments involving real job searchers; see Krause et al., 2012).

[3] See *Kleber v. Carefusion Corp.* (http://www.aarp.org/content/dam/aarp/aarp_foundation/litigation/pdf-

Regulations covering the ADEA currently states, "Help wanted notices or advertisements may not contain terms and phrases that limit or deter the employment of older individuals. Notices or advertisements that contain terms such as age 25 to 35, young, college student, recent college graduate, boy, girl, or others of a similar nature violate the Act unless one of the statutory exceptions applies" (§1625.4).[4] Beyond that, organizations like AARP suggest that "[d]espite protections by the Age Discrimination in Employment Act of 1967 (ADEA), employers have gotten cleverer in masking what is age discrimination" by using ageist phrases in job ads (Brenoff, 2019).

We find strong evidence that ageist language related to communication skills, physical ability, and technology skills, even when it is not blatantly or specifically age-related, deters older workers from applying for jobs. Job ads that feature ageist language attract younger applicants on average than job ads that do not feature ageist language. For example, ads containing a machine-learning generated phrase related to each of these three stereotypes attract job applicants that are 2.5 years younger on average, and more ageist phrases we examine have even stronger impacts.

In our view, we add significant new evidence to the research literature on discrimination. As the first paper to examine how discriminatory language in job ads impacts job search behavior in a field experiment, we provide evidence on the effect of discrimination on the behavior of workers, whereas the research literature on discrimination focuses on the behavior of *employers*. Building on the results from the correspondence study in Neumark et al. (2019a), which showed that employers discriminate against older workers, and Burn et al. (2022), who showed that the discriminating employers in this correspondence study used ageist language in job ads, the present paper answers the question of whether job seekers respond to these phenomena – which may also explain why employers use ageist stereotypes in job ads. Our experiment shows that workers respond to subtle shifts in the language of job ads that might signal that an employer holds ageist stereotypes about older workers or is otherwise less interested in hiring older workers.

Our experimental evidence that ageist stereotypes in job ads discourage older applicants

---

beg-02-01-2016/kleber-amended-complaint.pdf, viewed November 8, 2017). Surprisingly, the court ruled in favor of the defense in this case, reaching a new interpretation that the ADEA does not authorize job applicants to bring a disparate impact claim. (See Button, 2019.)

[4] As described in Pillar et al. (2022), in the Netherlands, the Dutch Equal Treatment Act regulates both explicit mention of age (e.g., "younger than 30 years") and formulations that imply age (e.g., specifically recruiting students). In contrast to the work in Burn et al. (2022), discussed below, Pillar et al. study detection in job ads of discriminatory statements that are explicitly defined and prohibited by the law.

from applying for jobs thus illustrates a more subtle form of age discrimination in the labor market.[5] Age discrimination that deters older workers from applying for jobs has the same effect as direct age discrimination applied to job applicants; both reduce the employment of older workers.[6] Strikingly, our evidence from this new experiment, combined with prior evidence on age discrimination in hiring from the correspondence study (Neumark et al., 2019a), suggests that the effect of direct hiring discrimination may be only slightly larger than the effect of discouraging older workers from applying for jobs – although the evidence from both experiments is specific to the experimental conditions, and may not generalize to the actual incidence of age discrimination in hiring and the effects of age-related stereotypes in job ads in the broader labor market.

Our evidence has significant policy implications regarding age discrimination and its enforcement. Utilizing ageist language in job ads may be rational for employers, despite it being illegal to discriminate against older workers. Shaping the applicant pool by discouraging older applicants has a potential benefit for discriminatory employers, because of the incentives created by age discrimination laws. In particular, a lower representation of older workers in their applicant pool can justify a lower representation of older workers among employees, making it easier to rebut an allegation of age discrimination in hiring. More generally, employers who do not want to hire older workers might, in order to avoid unnecessary search costs, discourage older workers from applying by signaling their ageism. To think about this another way, in the legal system, hiring discrimination cases based on age (or, similarly, membership in other groups) typically hinge on shortfalls of older workers among hires relative to the applicant pool. But if job-ad language deters older workers from applying, these shortfalls may be obscured, and the courts may need to weight other evidence more heavily – including both job-ad language as the source of lower applications from older workers, and comparisons with other benchmarks to assess whether hiring of older workers is notably lower at the firm in question.

To address age discrimination from stereotyped job-ad language that discourages older

[5] Our evidence should be viewed as another dimension of age discrimination in hiring – one that has not been studied or detected in the research literature that tests for hiring discrimination based on age, mainly using correspondence studies. These include Baert et al. (2016); Bendick et al. (1997, 1999); Carlsson and Eriksson (2019); Farber et al. (2017, 2019); Lahey (2008); Neumark et al. (2016, 2019a, 2019b); and Riach and Rich (2006, 2010).

[6] Yet another way to "discourage" older workers from applying for jobs is to target job ads to younger job seekers, as discussed in Ajunwa (2019), who also discusses a class action complaint against Facebook and other companies for such targeting. Moreover, the complaint alleged that Facebook used similarly discriminatory age filters in targeting its own employment ads.

workers from applying for jobs, there are two tools the U.S. Equal Employment Opportunity Commission (EEOC) – or other anti-discrimination authorities – could utilize.[7] First, it could issue stronger guidance to employers on language to avoid that might be interpreted as discouraging older workers from applying. There exists a duty of care for employers to knowingly avoid using language which may deter older workers from applying. Therefore, our evidence provides a basis for further guidance regarding the usage of ageist stereotypes in job ads that may influence job application decisions. Second, the EEOC might consider flagging for potential investigation firms that use age-stereotyped language in their job ads, recognizing that, for these firms, discrimination may be occurring even in the absence of shortfalls between the share of older applicants hired and the share of older workers who apply for jobs. Thus, rigorous evidence on the role of ageist language in job ads could potentially influence policy to reduce age discrimination in hiring and contribute to lengthening work lives. Obviously, similar considerations could apply to other groups protected by discrimination laws.

Previous Related Literature

Very few studies in labor economics explore job ads, and fewer still focus on discrimination. Among studies of issues other than discrimination, Modestino et al. (2016) use text data from job ads to document that "downskilling" occurred during the recovery from the Great Recession, with firms reducing skill requirements in their job ads. Deming and Kahn (2018) use text data in job ads to measure how ten different skills relate to wages. Marinescu and Wolthoff (2020) match text data from job ads to job application data to study the matching process between jobs and applicants. Kuhn and Shen (2013) and Kuhn et al. (2018) explore how gender preferences feature explicitly or implicitly in job ads in China, Chaturvedi et al. (2021) examine gender preferences in job ads in India, and Hellester et al. (2020) study age and gender preferences in job ads in China and Mexico.

A small number of studies are closer in spirit to ours in that they run experiments manipulating job ads and study responses of job seekers. Among these, He et al. (2021) study how job flexibility conditions influence job application behavior. Flory et al. (2015) study job seeker responses to ads for jobs that differ regarding competition and uncertainty about pay. And closer to our work, Flory et al. (2019) examine how signaling interest in employee diversity

[7] European Union law also bars age discrimination. To the best of our knowledge it is less explicit about the forms of discrimination barred, and it also differs in not protecting older workers per se, but rather barring discrimination based on age generally. See Lahey (2010) and European Commission (2000).

affects application behavior of minority and female candidates (as well as firm selection) – although this should probably be viewed as trying to encourage job applications from a particular group, the opposite of the behavior we study.[8]

Two studies connect the text of job ads to measured discriminatory behavior of employers.[9] Tilcsik (2011) identifies words in job ads related to masculine stereotypes (decisive, aggressive, assertive, and ambitious) and links those to hiring outcomes in a correspondence study of discrimination against gay men.[10] And, in the most systematic approach, Burn et al. (2022) identify common age stereotypes from the research literature in industrial psychology, use machine learning to calculate the relationship between the text of the job ads and specific age stereotypes, and then test whether job-ad language related to the stereotypes predicts hiring discrimination against older workers in a correspondence study. As already noted, the present paper builds on this prior study.

There has been no research on how ageist language in job ads affects the decisions of older workers to apply. What is known about how job applicants read job ads for bias focuses exclusively on gender bias in job ads. Gaucher, Friesen, and Kay (2011) found that job ads for male-dominated occupations used masculine wording (i.e., words associated with male stereotypes, such as leader, competitive, dominant) more frequently than advertisements for female-dominated occupations, and women find job advertisements less appealing when they contain more masculine than feminine wording (Bem and Bem, 1973; Gaucher et al., 2011).[11] However, these findings are based on laboratory experiments that ask how subjects perceive job-ad language, whereas our research uses a field experiment that studies the behavior of actual job seekers.

Conceptual Framework, Interpretation, and Model

---

[8] There is other research suggesting that, in other contexts, job seekers respond to job-ad language, including: Belot et al. (2018) and Banfi et al. (2019) on posted wages; Hellester et al. (2020) on gender requests; and Ibanez and Riener (2018), Leibbrandt and List (2018), and Flory et al. (2021) on affirmative action or diversity statements in recruitment materials or job ads.

[9] Though they did not focus on job ads, Hanson et al. (2011, 2016) study language used by mortgage originators and connect this language to their behavior. Hanson et al. (2011) study subtle discrimination through "keywords" used by landlords responding to prospective tenants. Hanson et al. (2016) had research assistants subjectively (and blindly) code the helpfulness and other characteristics of mortgage loan originator responses to prospective borrowers.

[10] In an early small study, Wax (1948) found that summer resorts in Ontario, Canada, were more likely to discriminate against Jewish customers (based on names) requesting accommodations if they used phrases like "restrictive clientele" in their advertising.

[11] Chaturvedi et al. (2021) examine how words that predict the employer having a gender preference are correlated with job applications, but they do not examine the link between the words and stereotypes about men and women.

Why might employers use stereotyped language in job ads? One hypothesis – perhaps the central one – is that employers who discriminate based on age use stereotyped language to try to shape the applicant pool. Using language that conveys positive stereotypes related to young workers might discourage older workers from applying (as might language conveying negative stereotypes related to older workers – although that seems less likely and is, in fact, less common in our data). Employers may introduce this language via job requirements that are correlated with age, appear natural to use in job ads, and are not so blatant as to make the age discrimination clear.

This discouragement from applying would lead to the underrepresentation of older applicants in the applicant pool, and is potentially valuable to a discriminating employer because the probability of a hiring age discrimination claim and of an adverse outcome for the employer is smaller when, ceteris paribus, the ratio of older applicants to younger applicants is lower.[12] Employers could use job-ad language this way whether their discrimination is taste-based or statistical, and, in the case of statistical discrimination, whether or not the language is related to the assumptions they make about older workers (e.g., they might assume older workers will leave the firm sooner). In any of these cases, employers might use ageist language in job ads to deter older workers from applying.

A second hypothesis, which is more complex, is also related to statistical discrimination. Different jobs may have different requirements, which could be stated in job ads without any explicit intent of discouraging older applicants. But employers may hold stereotypes about older job applicants' abilities to meet these job requirements – for example, assuming that older workers are less likely to be able to do the heavy lifting that a job requires, which may well be true on average but of course not of each applicant.[13] Employers may act on these assumptions, and older job seekers, expecting this, may be deterred from applying.

[12] In legal cases, the most compelling data on hiring discrimination comes from comparing hiring rates of the group in question (e.g., older workers) relative to the applicant pool. Hiring charges under the U.S. Age Discrimination in Employment Act (ADEA) made up nearly 5% of total ADEA charges in 2020 – more than double the percentage under Title VII (protecting women, minorities, etc.) or the Americans with Disabilities Act. (This is based on authors' computations using EEOC statistics available at https://www.eeoc.gov/statistics/statutes-issue-charges-filed-eeoc-fy-2010-fy-2020, viewed January 18, 2022.) The representation of hires among applicants is important in anti-discrimination enforcement, as the EEOC uses a "4/5ths" rule (the ratio of the selection rate for the group in question to the group with the highest selection rate) as "a practical means of keeping the attention of the enforcement agencies on serious discrepancies in rates of hiring, promotion and other selection decisions" (U.S. Equal Employment Opportunity Commission, 1979).
[13] It is also possible, in principle, that employers use this language randomly and unintentionally, but it still deters older workers from applying. However, the evidence in Burn et al. (2022), showing that discriminatory employers used ageist stereotypes in job ads, ruled this out.

While social scientists are interested in the nature of discriminatory behavior, both statistical and taste discrimination are illegal under U.S. law. Not surprisingly, language in job ads that refers to age either explicitly or "mechanically" (e.g., referring to recent graduates) is illegal in the United States. The legality of less blatant job-ad language with job requirements that reflect age stereotypes and is associated with lower hiring of older workers is more complex. On the one hand, EEOC regulations state: "An employer may not base hiring decisions on stereotypes and assumptions about a person's race, color, religion, sex (including pregnancy), national origin, age (40 or older), disability or genetic information." (See U.S. Equal Employment Opportunity Commission, n.d.(a).) On the other hand, job requirements that are based on factors related to age are not necessarily illegal. The legality of job requirements related to age generally requires an employer to show that the use of these requirements is based on a reasonable factor other than age (RFOA), even if that factor is correlated with age. An RFOA is defined as "a non-age factor that is objectively reasonable when viewed from the position of a prudent employer mindful of its responsibilities under the ADEA under like circumstances." (See Federal Register, n.d.) In other words, the law recognizes that characteristics of workers that are related to age can sometimes be legitimate for employers to consider.[14]

Our evidence cannot speak directly to the question of taste vs. statistical discrimination or whether the stated job requirements would be viewed as legal. Indeed, we do not study employer behavior in our experiment, although we do use job-ad language from real employers. What our evidence *does* address is whether age stereotypes expressed in job ads affect the likelihood that older job seekers apply for jobs, likely by signaling to job applicants that older workers are less likely to be hired. A response could mean either that the language is perceived as directly reflecting age bias – aversion to hiring older workers – or that the language is perceived as "biased" because it puts older workers at a disadvantage because they may be less likely to satisfy

[14] Indeed, the law even goes further, as in some rare cases employers can even use age as an explicit criterion if a requirement for the job is strongly related to age but hard to assess independently. This exception requires that age can be shown to be a "bona fide occupational qualification" (BFOQ) that is "reasonably necessary to the normal operation of the business." (U.S. Equal Employment Opportunity Commission, n.d.(b)). A key example is *Hodgson v. Greyhound Lines, Inc.*, where the company was sued for having a maximum hiring age. Greyhound prevailed by establishing that driving ability is essential to passenger safety, that older hires would be less safe drivers (because achieving maximum safety took 16-20 years of experience), that some abilities associated with safe driving deteriorate with age, and that these changes are not detectable by physical examination (which could otherwise be a substitute for an age criterion). (See U.S. Court of Appeals, 7th Circuit, 1974.) As discussed by Combs (1982), the issue of the rights of older workers vs. public safety have figured prominently in court decisions regarding age as a BFOQ under the ADEA.

the stated job requirement, or perceived as such by employers. Thus, our evidence can reveal the potential for employers to use job-ad language to discriminate against older workers in hiring, and the potential adverse impact on older job applicants.

A simple model can describe the behavior of job seekers that we use to interpret our evidence. When deciding whether to apply to a job, potential workers read the job ad and decide whether the potential benefit outweighs the costs of applying. Suppose the utility of job j to person i is $U_{ij} = \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0,1)$.[15] S indexes how age-stereotyped the job ad is. The cost of applying for a job is c. The potential benefit of a job is determined by posted wage and the probability of getting a job offer (callback). For younger workers this is $p_y = b$ ($0 < b \leq 1$) if young, and $p_o = b(S)$ ($0 \leq b(S) \leq b$, $b'(S) < 0$) if old. That is, the probability of a job offer for an older worker is a decreasing function of how age-stereotyped (S) is the job ad. An example of a function meeting these conditions is:

$$(1) \quad b(S) = b \cdot e^{-\eta \cdot S}, \ (\eta > 0).$$

A young person applies if $b \cdot \varepsilon > c$ (dropping the i and j subscripts) or $\varepsilon < -c/b$. Given the distributional assumption, the probability of applying is $A_y = \Phi(-c/b)$, where Φ denotes the standard normal distribution function. An old person applies if $b(S) \cdot \varepsilon > c$ (dropping the i and j subscripts) or $\varepsilon < -c/b(S)$, so the probability of applying is $A_o = \Phi(-c/b(S))$.[16]

In this paper, we are interested in estimating $\partial A_o/\partial S$. For old applicants,

$$(2) \quad \partial A_o/\partial S = \frac{-(-c)b'(S)}{b(S)^2} \cdot \phi\left(\frac{-c}{b(S)}\right),$$

$b'(S) < 0$ implies that $\partial A_o/\partial S < 0$, thus predicting a negative response of older applicants to job-ad language with more ageist stereotypes.

Methods

To test whether older workers respond to ageist language in job ads (i.e., is $\partial A_o/\partial S < 0$?), we conduct an experiment where we manipulate S and observe how the applicant pool changes. In our experiment, we post job ads in three occupations in 15 U.S. cities, randomly varying the inclusion of age-related stereotypes in the text of the job ad. The job ads are artificial, and we study the responses of real job searchers. This allows us to test for differences in the applicant pool when otherwise similar ads use age-related stereotypes vs. age-neutral language.

Selecting the Cities and Occupations

[15] The variance can be fixed without loss of generality.

[16] We assume that $\partial A_y/\partial S = 0$, or young people do not respond to the stereotyped language. We could have the probability of an offer for a young applicant *increasing* in S – i.e., the opposite direction from old people – and the qualitative conclusion is the same.

We build on the experiment conducted in Neumark et al. (2019a). The cities in that study were selected due to their large size, their geographic distribution across the U.S., and because they have different population age distributions. For this experiment, we added three more cities with a large online presence for the job board we use. For each city, we post our job ad on the online job board, setting the hiring firms' locations to the central business district. The cities in the experiment are shown in Figure 1 (which also provides additional information on the data collection in the experiment).

We use three of the four occupations from the original study: retail sales (mixed-gender), administrative assistant (female-dominated), and security guards (male-dominated). These occupations are lower-skilled, with jobs often advertised using online job boards. These jobs are also very relevant for older workers seeking new employment. As shown in Neumark et al. (2019a), all three occupations were in the top decile of jobs in terms of the proportions of older people hired.[17,18]

Selecting the Stereotypes

To select the stereotypes we use in our experiment, we start with a list of ageist stereotypes from the industrial psychology literature (see Burn et al., 2022). These are listed in Table 1. Among these, we selected stereotypes that met the following criteria. First, the stereotype is commonly expressed in job-ad language about the ideal or preferred candidate skills or attributes; we did not want to focus on stereotypes that are not often included in job ads (e.g., hearing and memory), even if employers hold these stereotypes based on the industrial psychology literature.

Second, we focus on stereotypes for which we had evidence of a correlation between discrimination and the stereotype (from Burn et al., 2022) and evidence that the stereotype captures a skill that employers view older workers as less likely to possess (from van Borm et al., 2020).[19]

Third, older workers should be aware that employers held that stereotype. As evidence, we drew on various reports put out by AARP; see Brenoff (2019) and Terrell (2019).

[17] As reported in Neumark et al. (2019a), looking at the distribution of the share of 62-70 year-olds hired recently (tenure less than five years) across all occupations, the percentiles for males in the occupations we use were 96.6 for retail salespersons and 93.3 for security guards and gaming surveillance officers. The percentiles for females were 100 for secretaries and administrative assistants, 96.4 for receptionists and information clerks, and 95.2 for retail salespersons.

[18] We omit the janitor jobs also included in Neumark et al. (2019a), because for them the evidence of age discrimination was less clear-cut, and there are many fewer janitor job advertisements posted online.

[19] We did not require this evidence for all three occupations or for both genders, but just for some subset.

Our final list of stereotypes is three skills or abilities for which older workers are stereotyped as deficient: communication skills, physical ability, and technological skills.

Designing the Job Ads

We used one job ad template per city-occupation combination, basing the structure and language of the template on actual ads collected in Neumark et al. (2019a) and recent real ads posted on job boards in the sample cities to capture contemporaneous patterns in their language. We selected a handful of ads to use as our base to create a template and copied their format (location of blocks of text, types of bullet points, and style of text) to ensure our templates were similar in appearance to others on the website. We stripped the ad of all identifying information, so there is no identifiable link between the ad posted and the template we created. The text of each ad was rewritten to give enough details about the company and the position to appear realistic, but not enough details to suggest a specific company. We modified the requirements of the jobs to reduce the number of applicants they potentially exclude. All ads were written to have flexible hours, competitive pay, and the availability of part-time and full-time positions (at the employee's choice). For half of the templates, we included the requirement that applicants must have a high school diploma (randomized by template). Figure 2 provides an example of a job ad for each of our occupations.

The treatment and control ads differ in the job requirements (denoted in bold with asterisks in each template in Figure 2), with three sentences assigned to be either a treatment phrase (stereotyped) or a control phrase (not stereotyped). The requirements we manipulate have to do with a candidate's communication skills, physical ability, and technology skills. Our control phrases express job requirements that are also appropriate for the job but use age-neutral language not related to these age-stereotyped skills or abilities, and as much as possible refer to related skills, while our treatment phrases use language highly related to these ageist stereotypes.

Creating Stereotyped Job Requirements

We use two methods to generate sentences highly related to ageist stereotypes, focusing on constructing sentences that were highly related to only one of the three stereotypes we use. The first uses measures of semantic similarity generated by machine learning methods. Drawing on Burn et al. (2022), we calculate the semantic (cosine) similarity of thousands of phrases to communication skills, physical ability, and technology skills. From this list, we construct our treatment sentences. We iteratively edited the sentences to ensure that only the cosine similarity score of the manipulated stereotype substantively differed between the treatment and control phrases. For example, if the treatment language related to communication skills was also highly related to the stereotype about personality, we identified which words in the sentence were highly related to personality and selected synonyms that were less related to personality. Our control sentences were created to express requirements for similar jobs without referring to ageist stereotypes about skills or abilities. We iteratively removed phrases that were highly related to our stereotypes to minimize the semantic similarity. The sentences for the treatment and control groups are listed in columns (3) and (4) of Table 2.

Figure 3 illustrates the distributions of the treatment and control phrases in the distribution of all text from the job ads collected in Burn et al. (2022). The key insight from this figure is that the control phrases are close to the median and thus should not be regarded as ageist by the average job seeker reading the text, while the treatment phrases are coming from higher in the distribution, close to the 75th percentile.[20]

Our second treatment conveys bias by using ageist language identified by AARP as the text related to communication skills, physical ability, and technology skills. We select three AARP examples that correspond to our respective stereotypes: "cultural fit," "energetic person," and "digital native" (Brenoff, 2019; Terrell, 2019). We adapted the language to fit our job ads and created three sentences, one for each stereotype (Table 2, column 5). Using the text about cultural fit, we created the phrase "You must be up-to-date with current industry jargon and communicate with a dynamic workforce" to reflect stereotypes about communication skills, emphasizing the communication aspect of fitting in. Using the text about energetic persons, we created the sentence "You must be a fit and energetic person" to reflect stereotypes about physical ability. Using the text about digital natives, we created the sentence "You must be a digital native and have a background in social media" to reflect stereotypes about technology skills by emphasizing social media.

We vary the combination of treatment and control phrases used in a job ad to create six job ads from each template: one control ad and five treatment ads. In our control ad, we use all three control phrases to express the skill requirements in language unrelated to ageist stereotypes. Four of the treatment ads utilize machine learning derived stereotyped phrases. We have three ads where we use the stereotyped phrase for either communication skills, physical ability, or technological skills (i.e., only one at a time) and the control phrases for the other two stereotypes, and there is one ad where we use all three treatment phrases. In the AARP treatment, we use all

[20] This is somewhat lower than the types of phrases analyzed in Burn et al. (2022), who focused on phrases above the 90th percentile. But the usage of phrases closer to the 75th percentile provides greater insight into the types of phrases that are more common to observe in job ads, and are also less obviously ageist.

three treatment phrases.

It might seem unsurprising if the AARP phrases deter older workers from applying for jobs. However, our machine-learning generated phrases are far more subtle, and as Figure 3 shows are by no means outliers, relative to the text of job ads, in their semantic similarity with age-related stereotypes. In addition, as shown in Burn et al. (2022), the kinds of age-stereotyped phrases from the job ads that we use help predict age discrimination by employers, as measured in the correspondence study. In other words, our experiment provides evidence on the effects of real-world job ads with language that reflects ageist stereotypes relatively subtly, and that is sometimes used by employers who – based on experimental evidence – discriminate against older workers in hiring.

Validating the Treatment vs. Control Differences

One question about our treatments is how well the stereotyped vs. neutral phrases generated by the machine learning convey the intended stereotypes. In the language of epidemiology, we would like our treatment ads to have high “sensitivity” (conveying ageist stereotypes) and “specificity” (conveying information about the specific ageist stereotype intended).

Figures 4a-4c illustrate how the semantic similarity scores differ across the templates for the treatment and control job ads, and they show that our treatment job ads do activate the intended stereotypes. In these figures, words have been aggregated up to three-word phrases to ensure that we measure semantic meaning more accurately. Information on the distribution of all phrases found in the ads in Burn et al. (2022) is shown in grey, information for the treatment ads is shown with dashed black lines, and information for the neutral ads with solid black lines. The figures show the median to 99th percentile range and the average (with plotting symbols).

These figures indicate that biased (treatment) templates have considerably higher 99th percentiles than the control templates, as well as higher means (and medians, although less so). The implication of the differences in the means and especially the upper tails of the distributions is that the treatment ads we write using the stereotyped language do, in fact, create ads with notably stronger age stereotypes. In addition, we see – importantly – that our treatment ads with single stereotyped phrases only generate a shift in similarity for the stereotypes we are seeking to activate, hence isolating those stereotypes in the job ads. Finally, note that the actual “collected” ads are more similar to the treatment ads – reflecting the fact that actual job ads often use ageist stereotypes (as documented in Neumark et al. (2019a)).

The second way we assessed the validity of the treatment vs. control ads as activating the intended stereotypes was to conduct a validation exercise using Amazon MTURK. We found that the control phrases were not perceived as ageist by applicants, and treatment phrases were perceived as more ageist than the control phrases. The AARP phrases were perceived as the most ageist, with the machine learning phrases intermediate between the AARP and control phrases, as we would expect. The survey and results are detailed in Burn et al. (2021).

Figure 5 presents this evidence, providing a graphical depiction of the answers from the MTURK survey participants. Across the three blocks of the survey that solicited respondents' self-assessments of age bias, their predictions of previous respondents' answers, and their predictions of the answers of respondents over the age of 50, our results were consistent. The participants, on average, strongly disagreed with the notion that anyone would perceive the control phrases as biased against workers over the age of 50.[21] Respondents rated the physical and technology-biased phrases derived from our machine learning methods as more biased than the control phrases, but viewed the communication skills stereotyped phrases as roughly identical to the controls. Views of the AARP-derived treatment phrases were starker, as all three were rated as far more age biased than their respective control counterparts. The absence of evidence for bias for the language related to communication skills may reflect the fact that older workers are not always stereotyped as having worse communication skills but are sometimes, as Table 1 showed, perceived as having better communication skills. In that sense, one might view the evidence of ageist ratings for the physical ability and technology-related stereotypes but not the communications stereotype as further confirmation that respondent perceptions accord with the industrial psychology literature. (Note that the cosine similarity scores from the machine learning do not detect positive vs. negative uses of the language.)

These results, like those in Figures 4a-4c, imply that our phrases capture real ageist sentiments and will be perceived as such by job applicants, so our results should be informative about the effect of ageist language on job ads on job applicants' behavior. That is, failure to find an impact would be informative about job applicant responses to age-stereotyped job-ad language, rather than reflecting a failure to convey these stereotypes in the job ads.

Posting the Job Ads

We had a total of 18 ads to post in each city, six per occupation. We staggered the posting of ads to leave two weeks in between the taking down of one ad and the posting of the next within each city. To avoid p-hacking, we initially planned to run the experiment for 54 weeks, with the

[21] In the figure, values to the right are associated with less perceived bias.

schedule of posting pre-registered.[22] To maximize the number of potential applicants, one ad was to be posted each weekday (Monday through Friday).[23] The rotation of the ads posted was staggered such that there were eight weeks between the same occupation's ad appearing in the same city with different treatment statuses.

This was a complicated process. The job board we used for the experiment makes money from fees for posting job ads, and hence is sensitive to fake ads, ads used for phishing, etc. In the course of the experiment, we encountered problems if we tried to use the same credit card to pay for ads in different cities, or used the same IP address for posting ads in different cities. In addition, there seem to be human "checkers" for each city on the job board, who monitor for highly similar ads or ads that appear to be from fictitious companies.

To get around the payments problem, we used a very large number of gift cards, so we would never use one card more than four times.[24] Even this required some workarounds, as the websites for some gift cards made it difficult or impossible to register a large number of cards from the same IP addresses over a short period of time – sometimes prompting impossible to resolve "are you a robot" questions or tasks. This is apparently because gift cards are used by those who steal credit card information,[25] or others (like money launderers) who want to avoid detection.[26] We thus had to experiment to identify gift cards that did not have this constraint. To get around the problem with IP addresses, we purchased numerous cell phones and SIM cards for each city, using pay-as-you-go plans that randomize the IP address each time the service is restarted. Figure 6 conveys an idea of what was involved.[27]

There was no way around the human checkers. A number of our ads were flagged by the job board as spam and taken down before they had been active for a week, or our payment method was rejected, leading to a delayed or skipped job posting. If the ad was taken down before we received responses, we began to repost it at the end of the study, starting in week 55. For city-occupation cells where multiple ads were taken down, we reposted them in the order that they

[22] As discussed below, we anticipated some difficulties in placing ads in some cities, and the PAP explicitly called for a period subsequent to the initial 54 weeks when we would re-try placing these ads, in the same order.
[23] Initially, all ads were randomized to either Monday or Tuesday, but we had to switch to five days a week to avoid triggering a moderator response. The job ad board suggests not to post ads more often than every 48 hours.
[24] We used gift cards up to $100, which, depending on the city, would cover two to four ads.
[25] See, e.g., https://www.forbes.com/sites/laurengensler/2017/01/11/gift-cards-money-laundering/?sh=5498ac6f1449.
[26] See, e.g., https://www.fraud-magazine.com/article.aspx?id=4294967696.
[27] Our university grants administrator was frequently puzzled by the receipts submitted for reimbursement.

were originally meant to be posted, still leaving one week in between each ad.

For two cities (Boston and Pittsburgh), we were unable to post many ads, due to flagging. Because of this, and because the budget allowed it, we replaced these two cities and added two additional cities (Seattle, Washington, D.C., Minneapolis/Saint Paul, and San Diego), early on in the experiment after we encountered problems. These cities were selected due to having higher numbers of job postings on the job board, which increased the likelihood that ads were being seen. Furthermore, we chose to add more than one city to replace the two problem cities in case problems emerged in other cities, based on our early experience with posting ads in Boston. Because these cities were not specified in the Pre-Analysis Plan (PAP), we also report key results for the originally proposed cities only; the results were qualitatively very similar.

Consistent with our PAP, we collect responses to our ads that we received within one week of the posting. We found, early in the experiment (when we were testing our procedures), that very few responses arrived after one week. Additionally, with our design and schedule, no two ads based on the same template were ever concurrently available on the job posting board.

Collecting Responses

Usually, applicants sent us their resumes when they replied to our job ad. To reduce the cost of applying for our fake job, we informed applicants that they were not selected for an interview and that we had decided to go with another candidate, within 24 hours of their application, via an email.[28] While it is rare for employers to inform their applicants of a negative outcome, we believe that this was important to reduce possible costs to participants.[29] We understood, in designing this study, that there are potential ethical issues involved, and we did not have the capacity to provide real jobs to applicants – as is possible, for example, in experiments on job platforms that offer inexpensive, short-term employment for specific tasks (e.g., Pallais, 2014). We thus chose these procedures to minimize potential harm to job seekers. (And, as noted earlier, respondents were debriefed and offered the opportunity to have their

[28] Emails read: "Thank you for your interest in this position. Unfortunately we will not be pursuing your application at this time." If someone expressed interest in applying *and* had a question, we provided the same response. If someone responded and *only* had a question about the ad, we did not reply nor did we have resume data to include (nor an email address from the resume). We did not respond to recruiters. If a job placement agency (e.g., refugee resettlement) sent a resume on someone's behalf, we sent the response addressed in the third person ("we will not be pursuing X's application").

[29] To try to avoid spam responses and to make sure that the applicants have read the job ad, we included a manipulation check. Each ad contained a cue that applicants were asked to respond to that does not appear out of place in a job ad, such as "Please indicate which days of the week you are available to work." However, a majority (66.9%) did not respond to the cue.

data withdrawn. Only a handful of respondents chose to do so.)

Calculating Applicant Age

Our primary outcome of interest is the age of applicants. We calculate the age of our applicants based on the available information listed on the resume. The first method to calculate age is based on the year of high school graduation (or equivalent).[30] Assuming that an individual was approximately 18 years old when they graduated high school, age is calculated as

$$Age = Date\ of\ Job\ Post - Year\ of\ HS\ Graduation + 18.$$

If the applicant does not provide a year of high school graduation, we calculate age based on the earliest date of work experience listed on the resume.[31] Age is calculated as

$$Age = Date\ of\ Job\ Post - Year\ of\ First\ Job + 16.$$[32]

Applicants were assigned the oldest age calculated across these methods.[33]

A concern is that ageist language may cause older applicants to hide their age, so the above methods to calculate age may undercount the number of older workers applying to the job ads with ageist language, creating a bias towards finding that ageist job-ad language deters older job applicants. To address this concern, we use a binary indicator to record whether or not we can determine an applicant's age from the information on the resume. If ageist language causes older applicants to manipulate their resumes to obscure their age, we should be able to capture this effect by comparing the shares of applicants whose age we cannot ascertain for job ads with ageist language and job ads without ageist language. There was also a small number of applicants who responded to more than one job ad, and for them we can see whether the reporting of age changes in response to the job-ad language. As described later, we find no evidence of applicants manipulating the inclusion of information on age or their indicated ages in response to age stereotypes in job ads.

Empirical Analysis

To test whether ageist language changes the composition of the applicant pool, our primary outcome of interest is the age of applicants. We calculate three measures of the age of the applicant pool using the data on age we extracted from the applications. First, we calculate the

[30] In a small number of cases this could be the year of getting a GED, or starting post-secondary education (even if a GED was received later).
[31] Additionally, we collected the earliest non-work listed year on their resume and calculated age as being 18 in that year. If applicants explicitly listed their age or year of birth we would record their age as such.
[32] This is probably best interpreted as "minimal possible age" assuming one did not start working before age 16.
[33] E.g., for an applicant with a 2021 job posting date who had a high school graduation year of 2014 (implying age 25), an earliest working year of 2007 (implying age 30), and the earliest non-work year listed on their resume of 2010 (implying age 28), we would assign them as being 30 years old.

average age of applicants (excluding, in each case, those who have not provided enough information to approximate their age). Second, we calculate the distribution of the age of applicants and identify the median and 75th percentile. Third, we calculate the share of the applicants aged 40 and over. Finally, as a check on manipulation of applications in response to job-ad language, we calculate the share of applicants to a job who do not provide information to approximate their age.

To estimate the effect of ageist language on whether older applicants apply, we estimate a series of models. We first estimate a regression equation that distinguishes between any type of age-stereotyped job-ad language treatment and the control job ads, defining the dummy variable S^A to equal one in the case of any of the five treatment arms):

$$(3)\ \ A = \alpha + \beta \cdot S^A + X\delta + \varepsilon.$$

A is, alternatively, the average age of applicants to the job posting, the median, the 75th percentile, and the share of applicants over 40 years old. Observations correspond to the city, occupation, and job ad cell.

We estimate the effect of the stereotyped language in an ad (S^A) on the age of applicants conditional on controls (X) for the city and occupation the ads were posted in.[34] Because the job ads vary by city and occupation and the job search behavior of applicants in a city and occupation may be correlated, we cluster the standard errors at the occupation and city levels.

Our null hypothesis is that the presence of ageist language on a job ad has no effect on the share of older workers that apply to the job ad (i.e., $\partial A_o/\partial S = 0$, where S represents, generically, the different versions of stereotyped treatments that we use). The alternative hypothesis is that the presence of ageist language in a job ad will reduce the share of older workers that apply to the job ad (i.e., $\partial A_o/\partial S < 0$). We do not think a two-sided hypothesis test is the most meaningful in our context, but we report test results from both one-sided tests and two-sided tests (because two-sided tests are more conventional).[35] For the one-sided tests, our null hypothesis is that stereotyped language in a job ad does not deter older applicants from applying for a job. If this hypothesis is true, then β will be greater than or equal to zero. If we find that β is negative, this is evidence in favor of the alternative hypothesis that stereotyped language in a job ad deters older applicants from applying for a job (or reporting age).

As can be seen in Figure 5, there is a significant difference in the perceived age bias of

[34] We show that the results are robust to including the full list of controls from the experiment, which in addition to city and occupation includes month posted and day of the week posted.
[35] Our PAP generally focused on the one-sided hypothesis.

the AARP treatments and the machine learning treatments. Therefore, we next estimate whether there is a difference in responses when using stereotyped language determined by the machine learning and when using the stereotyped language provided by AARP (*AARP*). To do this, we modify equation (3) to include an interaction between the dummy variable for stereotyped language in an ad and a dummy variable for using the AARP language.

$$(4)\ \ A = \alpha + \beta_1 S^A + \beta_2 (S^A \times AARP) + \delta X + \varepsilon.$$

Then, we test estimate the response to job ads with a single age-stereotyped phrase, by restricting the sample to the controls and, alternatively, the observations on job-ad language with a single age-stereotyped phrase related to communication skills (*C*), physical ability (*P*), or technology (*T*). In this case, we estimate a version of equation (3) of the form:

$$(5)\ \ A = \alpha + \beta \cdot S^j + X\delta + \varepsilon, j = C, P, or\ T.$$

We next estimate the most comprehensive model that distinguishes the five treatment arms:

$$(6)\ \ A = \alpha + \beta_1 S^A + \beta_2 (S^P \times P) + \beta_3 (S^T \times T) + \beta_4 (S^A \times All\ 3) + \beta_5 (S^A \times AARP) + \delta X + \varepsilon.$$ [36]

In this model, the identification of the effect of the specific stereotypes comes from the machine learning generated phrases and not the AARP phrases because the AARP treatment only ever includes all three stereotypes, while we separately enter each stereotype for the machine learning generated phrases.

Finally, rather than estimate models with dummy variables for different treatments, we "index" the treatments by either the cosine similarity scores of the ads (see Figure 3) or the measure of perceived age bias from the MTURK survey (see Figure 5). In the latter case, we use the Likert scale elicited in the survey, but reversing the order, relative to Figure 5, so higher is more biased – consistent with the cosine similarity score. When there is more than one job ad in the treatment, we compute the average value. We thus estimate:

$$(7)\ \ A = \alpha + \beta \cdot Score^j + X\delta + \varepsilon, j = CSS\ or\ Likert.$$

In both cases, we standardize the index so that β measures the effect of a one standard deviation increase in the index of age bias in the job ad.

Results

Descriptive evidence

[36] The PAP also calls for estimating heterogeneous effects along other dimensions. These analyses are described below, after our main results.

We begin by presenting the empirical cumulative distribution functions of applicant ages for the treatment and control ads, aggregating across all of the treatment ads (Figure 7). The raw data clearly show that the treatment ads that include ageist stereotypes attract fewer older applicants than the control ads. The CDF for the control ads is lower throughout nearly all of the distribution, consistent with the treatment arms – combined – leading to fewer older applicants. In Figure 8, we disaggregate the different treatment arms. There is a good deal of heterogeneity, but one can discern that the CDF for the control ads is generally lower throughout the distribution, again consistent with all of the treatment arms involving ageist stereotypes leading to fewer older applicants. One can also see that the treatment ads with three ageist phrases together, and the AARP treatment, are most pronounced in attracting fewer older applicants. Finally, among the ads with one ageist phrase, ads with an ageist phrase related to communication skills or physical ability tend to attract fewer older applicants than ads with an ageist phrase related to technology.

In Figure 9, we examine the empirical density functions of applicant ages by treatment arm. Displaying the data this way gives a clearer indication of the ages at which job searchers are less likely to apply when faced with ageist phrases in job ads. Looking at any treatment, in the upper-left panel, the under-representation is evident between about ages 40 and 60, whereas there are more younger applicants for the treatment job ads. For the individual communications and physical stereotypes, the same pattern is evident. In the lower-left pane, for the treatment including all three machine-learning generated phrases, the lower application rate from older workers is more marked and extends down to about age 35, indicating a stronger response when the job ad contains three stereotyped phrases. In the lower-right panel, for the AARP phrases, the separation between about ages 40 and 60 is stronger still, as is the greater representation of younger job applicants.

Regression results, combined treatments

We first present the regression results for the simplest specification (equation (3)), comparing the ages of respondents to job ads with any type of age-stereotyped language to the controls. These estimates are reported in the top panel of Table 3. Job applicants to ads with ageist language are on average 2.7 years younger than the applicants to the control ads. The median age for these ads was also 2.7 years younger, and the 75th percentile was 3.1 years younger. The share of applicants over 40 was lower by 9.4 percentage points. All of these estimates are statistically significant at the 5% level or less, in both one-sided and two-sided tests.

The bottom panel introduces an interaction with the AARP treatment (equation (4)). In this case, the estimated coefficients of *Any Treatment* correspond to the machine learning phrases, the estimated effects of the *AARP* dummy variable are the differences between that treatment and the others combined, and the total effect of the AARP treatment is the sum of the coefficients. The first and perhaps most important result is that the effects of *Any Treatment* – which now exclude the AARP treatment – remain large. Job applicants to the ads with machine-learning generated ageist phrases are on average 2.2 years younger than the applicants to the control ad. The median age for these ads was 2.4 years younger, and the 75th percentile was 2.5 years younger. The share of applicants over 40 was lower by 7.9 percentage points. Except for the effect on the 75th percentile, all of these estimates are statistically significant at the 5% level or less, in both one-sided and two-sided tests. The second result is that the incremental effect of the AARP treatment is also large and statistically significant. For example, average age of applicants was lower by an additional 2.3 years, and the share of applicants over 40 was lower by an additional 7.7 percentage points. All these differentials are statistically significant at the 5% level or less, in both one-sided and two-sided tests.

Manipulation of age reporting?

Before moving on to the other specifications, we consider evidence on two other issues. First, because workers could, in principle, strategically mask their age in responding to ageist job ads, we test the effect of ageist language on the share of applicants providing no age-identifying information. This analysis can be viewed as studying selection into reporting age.[37] As reported in the last column of Table 3, we find no evidence of this kind of selection, and hence do not need to be concerned with this potential source of bias. The same is true in the tables that follow; since the result is the same, we include the corresponding column in each table, but do not discuss this result anymore.

A related possibility is that job applicants report information on age, but manipulate this information to appear younger in response to age stereotypes in job-ad language. We do not regard this as very likely, given that applicants will almost certainly have to reveal accurate age information prior to taking the job. We can actually garner some evidence on this question from repeat applicants. In particular, there were about 400 observations on individual applicants who applied to more than one job ad.[38] For these observations, we can test the effect of the experimental manipulation of the job-ad language to see if the reported age information responded. Perhaps the most obvious hypothesis is that they would report information indicating

[37] In a different context, Kang et al. (2016) report interview evidence that minority job seekers (of university age) sometimes conceal or downplay racial cues in job applications, and in a lab experiment do so less in response to job ads that suggest the employer values diversity.
[38] Most of these applied to two ads, but some applied to three, four, or five.

a younger age or be less likely to provide age information for ads with ageist stereotypes. We found no evidence of this. In estimates corresponding to the specification in Table 5 (equation (6), where we distinguish the different treatments, discussed below), only two of the 15 estimates were statistically significant, and only at the 10% level. Moreover, these estimates were inconsistent with the age manipulation we might expect – with a positive effect of the physical stereotype on age, and a negative effect of the *All 3* treatment on reporting no age information. Thus, we conclude that age manipulation in response to the experimental treatment is not an issue.[39]

Multiple testing

Second, we consider multiple hypothesis testing. Given that we pre-registered our analysis plan, we are less concerned with the issue of searching for statistically significant results. Still, we are estimating multiple effects. We use the Simes procedure for the "false discovery rate" (Benjamini and Hochberg, 1995). Controlling the false discovery rate (for example, at the 5% significance level) means that we are 95% confident that at least some of the rejected null hypotheses are false. Given that our hypothesis is somewhat general – that ageist language in job ads may deter older job applicants – and that we do not have a strong hypothesis about either a specific treatment among those we consider, or a specific measure of the age of applicants, the false discovery rate is appropriate (as opposed to more conservative multiple testing methods). The procedure results in a q-value, which has the same interpretation as a p-value, but with the multiple testing adjustment.

We apply the multiple testing correction to each panel of Table 3 and the main tables that follow (Tables 4-6), for the different measures of the age of applicants, for each treatment including in a set of specifications. For example, in Table 3, it is applied to the first four columns of the top panel (for the estimates of equation (3)), and then to the first four columns of the bottom panel, which has two treatment variables. In Table 3, we highlight in boldface those estimates that remain significant at the 5% level in two-sided tests, and in italics those that remain significant at the 10% level. In this table, the statistical conclusions are very robust to multiple testing.[40]

Evidence for single stereotype treatments

Table 4 reports estimates of versions of equation (3) estimate for job ads with a single

[39] These results are available upon request.
[40] Table A1 in online Appendix A reports the p-values and q-values for Table 3, and our other main analyses (Tables 4, 5, and 6).

machine-language generated stereotyped phrase vs. the control arm (equation (5)). In each case, the sample is smaller because only one treatment arm is included. The three separate panels of Table 4 consider each stereotype in turn. For the individual stereotypes, and the smaller samples considered in Table 4, the results are statistically weaker. However, the key result, in our view, is that for every way that we measure the age of applicants – average age, median age, 75th percentile, and the share over 40 – for all three stereotypes, the sign of the estimate indicates that the ageist stereotype reduces applications from older job seekers.

For ads featuring an ageist phrase related to communication skills, the average age of applicants was 2.6 years younger than the control ads. The median and 75th percentile were lower by a similar amount, and the share over 40 lower by 7 percentage points. For ads featuring an ageist phrase related to physical ability, the average age of applicants was 1.8 years younger than the control ads. The effects on the median and 75th percentile were somewhat larger, and the share over 40 was lower by 8.2 percentage points. For ads featuring an ageist phrase related to technology, the average age of applicants was 1.9 years younger than the control ads, and the median was lower by 2.1 years, while the effects on the 75th percentile and the share over 40 were lower than for the other two stereotypes. Some of these estimates are significant at only the 10% level, and none after the correction for multiple testing. But the evidence that all 12 estimates in the first four columns are negative is nonetheless quite compelling.

Regression estimates for all treatment arms

In Table 5, we estimate a single model that uses all of the data, and includes separate variables for each of the different treatment arms (equation (6)). Note, first, that every single estimated effect on the age of applicants in this table (shown in the first four columns) is negative, implying that no matter how we measure age, and for every treatment and stereotype, the age-stereotyped job ads attract fewer older applicants. The estimates for the job ads with a single stereotype are similar to those in Table 4, although a bit stronger statistically for the communications and physical ability stereotypes. We would not expect any difference except because the city-occupation dummy coefficients are estimated from different (more) observations, and the residual variance changes.

The treatment arms including all three machine learning phrases or all three AARP phrases generate large and strongly significant reductions in the ages of job applicants. For ads that feature all three machine learning treatment phrases, job applicants were 2.5 years younger than the applicants to the control ad. The median age for these ads was also 2.5 years younger, and the 75th percentile was 4.2 years younger. The share of applicants over 40 was lower by 11.7

percentage points. All of these estimates are statistically significant at the 5% level or less, in both one-sided and two-sided tests, and at the 5% or 10% level after considering multiple testing.

For ads that feature all three AARP treatment phrases, the estimates are even larger and more strongly statistically significant. Note that in this specification, the *AARP* effect is the full effect of this treatment, not the incremental effect (as in equation (4) and Table 3). For example, the average age of applicants was 4.8 years younger than applicants to the control ads, and the share of applicants over age 40 was lower by 15.6 percentage points. All of these estimates are strongly statistically significant.

The evidence in Table 5 for the *All 3* treatment gives some indication of a "dosage" response, with a job ad that reflects more than one stereotype more strongly signaling to job applicants an employer is less likely to hire older workers and hence reducing applications from older workers by more. However, this is not consistent; it is apparent for the 75th percentile and the proportion over 40, but not for the average or median age. Moreover, the point estimate of the effect of the *All 3* treatment is well below the sum of the effects of the single stereotype treatments, indicating that even a single stereotyped phrase related to one skill or characteristic does not go unnoticed by job applicants and can have a sizable negative effect on job applications from older workers.

Regressions using cosine similarity score or Likert scale

Finally, we report the estimates substituting indices of age bias in the job-ad language for the treatment dummy variables (equation (7)). The top panel of Table 6 reports the estimates using the cosine similarity score (or scores, when averaged across multiple stereotypes). The evidence points strongly to age-biased job-ad language reducing the age of job applicants. When the index was one standard deviation higher, average age was lower by 1.7 years, median age by 1.5 years, and the 75th percentile of age by 3.1 years. The proportion over age 40 was lower by 8.3 percentage points. All of the estimates are statistically significant at the 5% level or less in two-sided tests, and after correcting for multiple testing. The evidence in the bottom panel, using perceived age bias of the language from the MTURK survey, is very similar. The point estimates are very close to those in the top panel, and the evidence is even stronger statistically.

Summary

Overall, we find significant evidence that ageist stereotypes reduce the likelihood that older workers apply, and the effects are substantial. For example, when all three machine-learning generated stereotypes are used in the model with all treatment arms, average age across cities was lowered by about 2.5 years (on a mean of 32.7 for the control group), and the share of applicants over age 40 was lowered by 12 percentage points (on a mean of 20.0% for the control group).

Robustness analyses

We next consider two robustness analyses.[41] In general, we do these analyses for the specifications with all treatment arms (Table 5), and for the indexes of age bias (Table 6). However, in cases where we introduce interactions, to avoid an excessive number of parameters we substitute the simpler specifications from Table 3 for those in Table 5. Given that the structure of the tables is similar to the preceding ones, we discuss these results briefly. The bottom line, though, is that none of these alternative analyses change our conclusion that ageist language in job ads deters older job applicants.

First, we test the effect of varying our definition of older workers by raising the threshold to 50 and then 65 years old. As shown in Tables 7A and 7B, the estimates for the age 50 threshold are quite similar to those for age 40. The estimates for the age 65 threshold are very small and statistically insignificant – although there were very few applicants above age 65, as Figure 7 shows.

Of course, these estimates measure the effect in the change in the proportion of applicants above a particular age as a result of the treatment. Since the baseline proportion of applicants declines the higher the age threshold, a smaller estimated change in the proportion above a higher age threshold can represent a larger relative effect. To provide more detail on the effect of changing the age threshold, but in terms of relative effects, we estimate the model for *Any treatment* (like Table 3) for the effect on the proportion above each age threshold from age 30 through age 70, rescaling the estimated effects by the baseline proportion. (We similarly rescale the confidence intervals, so the statistical inferences are based on the original regressions.) The results are reported in Figure 10. The estimates show that in fact the estimate at age 50 is a bit larger in relative terms (compared to 40). Moreover, the estimates show an increasing effect through about age 53, after which things get a bit less consistent. And at the oldest ages, although imprecise, the estimates are largest.

Second, we also add fixed effects for city and occupation, in Table 8. The estimates are little changed from Tables 5 and 6.

Comparing the potential effects of age-stereotyped job-ad language vs. direct age discrimination in hiring

As noted in the Introduction, age discrimination that deters older workers from applying

[41] These were specified in the PAP. Online Appendix B presents additional robustness checks. Most of these were specified in the PAP. We note the exceptions. None of these alter our conclusions.

for jobs has the same effect as direct age discrimination applied to job applicants; both reduce the employment of older workers. We can compare the estimated effects on the share of older job seekers hired from the discouragement of older applicants from age-stereotyped job-ad language, estimated in this paper, and the direct impact of age discrimination in hiring in the closely related correspondence study (Neumark et al., 2019a).

It is important to keep in mind, though, that the evidence from both experiments is specific to the experimental conditions, and may not generalize to actual incidence of age discrimination in hiring and age-related stereotypes in job ads in the broader labor market. Still, this evidence suggests that the two influences on hiring of older workers in the labor market could be of similar empirical importance.

In our experiment in this paper, the share of applicants over 40 in the control group is 20.00%; the use of ageist language reduces the share of job applicants over age 40 by 4.41 percentage points.[42] In the correspondence study, the overall callback rate for the over 40 group (averaging across those near age 50 and near age 65) was 13.78%, compared with 18.69% for those under 40, a shortfall of 4.91 percentage points. Clearly these effects are of a similar magnitude. However, it is more instructive to calculate the implied effects on the share of older "hires" among all "hires."[43]

- If there were no age discrimination in hiring and no discouragement of applications from older job seekers, the percentage of older workers among all hires would be the same as this percentage in the control group, or 20.00%.
- Age discrimination reduces the percentage of older workers among hires to 15.56%.[44]
- The discouragement of older applicants reduces the percentage of older workers among hires to 16.31%.[45]

[42] We use the individual-level data to be comparable across the two experiments.
[43] We equate "hiring" with "callback" for these calculations.
[44] This is computed by applying the actual hiring rates for older and younger applicants to the shares of applicants in each age group in the control group, to eliminate the discouragement effect:

$$= (0.2000 \times 0.1378)/[(0.2000 \times 0.1378) + (1 - 0.2000) \times 0.1869)].$$

[45] Since the hiring rate for older and younger applicants would be the same, this is computed by simply adjusting downward the proportion of older applicants and recomputing the share of older applicants:

$$= (0.2000 - 0.0441) /[(0.2000 - 0.0441) + (1 - 0.2000)].$$

Thus, the two effects are very similar, with the discouragement effect only slightly smaller.[46] If these numbers roughly generalize to the actual labor market, the implication is that enforcement that focuses only on hiring shortfalls could conceivably miss nearly half of age discrimination – subject also to the caveat discussed earlier that job-ad language that reflects age-related stereotypes may not solely reflect age discrimination.

We also have to be a little cautious in this interpretation, because it is conceivable that age-stereotyped language in job ads may signal job characteristics that older applicants dislike or do not think themselves capable of fulfilling. While it is not possible to disentangle job applicants' thought processes, we are quite confident these interpretations do not fully explain our findings for three primary reasons. First, we observe in Burn et al. (2021) that both older and younger respondents perceive that the machine learning and AARP job requirements are biased against older workers. Moreover, the AARP requirements were explicitly billed as "phrases employers use to mask ageist discrimination" (Brenoff 2019). Second, the evidence from Burn et al. (2022) indicates that employers who used physical and technologically-biased language sometimes discriminated against older men. If older applicants learn from experience that callback rates are lower for ads that include biased language, then we may expect they will be less likely to apply to such ads.

Third, the older workers deterred by our treatment phrases are largely between the ages of 40 and 60. Previous research indicates that age discrimination begins in one's early 40s (Carlsson and Eriksson 2019), which suggests that discrimination begins to appear before an age group becomes obviously less qualified to fulfill the job requirements. The AARP phrases, for example, do not convey any specific or objective skill requirements. In addition, the computer programs listed as required in our machine learning treatments (see Table 2) have been available for over 30 years, so many of our deterred applicants have been familiar with these programs for much of their working lives.

Nonetheless, the aforementioned possibilities imply that the use of ageist language, per se, in job ads does not necessarily imply discrimination, which parallels what we said earlier about the relationship between such language and RFOAs. Job ads that feature bona fide job

[46] Note that the summed effects exceed the overall effect by a little bit. That is because there is a negative interaction from the lower callback rate for older applicants being applied to a reduced number of older applicants. The percentage of older workers among all hires resulting from both effects is 12.56%, computed as (0.1631 × 0.1378)/[(0.1631 × 0.1378) + (1 – 0.1631) × 0.1869)]. This is a reduction of 7.44 percentage points, vs. the sum of the two effects adding to an 8.31 percentage point reduction, or [(0.2000 – 0.1556) + (0.2000 – 0.1631)].

requirements related to ageist stereotypes may be less attractive to older workers. Thus, were job-ad language to be added to the tools of anti-discrimination enforcement, it should be used only as a potential flag for discriminatory behavior – prompting further investigation, including whether employers who use such language are still less likely to hire older job applicants.

One might also object that discouraging older workers from applying for jobs does not ultimately impact their employment, because they can just apply to other jobs. But that assumes there is a very large supply of potential jobs to which older workers might apply, and they can simply pick and choose among them. However, the supply is not that large, especially in the smaller markets. Across city and occupation cells, the median number of job ads in the same category (which remain up for 30 days) was 123, the 25th percentile was 39, and the 10th percentile was 12. (Outside of administrative assistant jobs, the numbers are considerably lower.) Moreover, recall from Figure 3 that our treatment phrases were generally at around the 75th percentile of the distributions of CS scores for the three stereotypes, based on the job ads used in Burn et al. (2022). The implication is that there are many job ads with language as stereotyped or more stereotyped than our treatment job ads. In other words, job ads with the kind of language we use in our treatment ads are not easily avoidable. The same kind of behavioral response we observe in the experiment – discouraging job applicants from older workers – likely also occurs for many real-world job ads.

Discussion and Conclusion

In this paper, we conducted the first field experiment that examines how older job seekers respond to the presence of ageist language in job ads. We manipulated the language of online job ads to feature control phrases that had low relatedness to ageist stereotypes or treatment phrases that were highly related to ageist stereotypes or flagged as such by AARP. The treatment and control phrases were validated using two methods. The first shows that the machine learning phrases are only related to the specific stereotypes and are not related to any of the other stereotypes about older workers. The second method showed the phrases to individuals on MTURK and asked them to rate how ageist they perceived them to be. We found that the treatment job-ad language was viewed as significantly more ageist than the control job-ad language.

We study job ads posted in three occupations in 14 cities, with six job postings in each city-occupation cell. The results indicate that older workers, when faced with ageist language in job ads, are less likely to apply for jobs, with measures like average age and the share of applicants over the age of 40 (as well as other measures) falling. The results indicate that there may be additive or "dose-response" effects of ageist language, with the effect growing with additional ageist phrases in the job ads. Job ads with multiple ageist phrases led to strong declines in applications from older job seekers. For example, when job ads included three machine-learning generated phrases with ageist stereotypes related to communication skills, physical ability, and technology skills, the share of applicants over 40 declined by 12 percentage points, and the average age of applicants fell by 2.5 years. The decline was particularly sharp in the upper parts of the age distribution, with the 75th percentile falling by 4.2 years.

Our evidence has significant policy implications regarding age discrimination. We show that ageist stereotypes in job ads discourage older applicants from applying for jobs. The effects of this discouragement of applications from older job seekers can have as deleterious an impact on the hiring of older workers as can direct age discrimination in hiring; indeed, our evidence suggests the discouragement effect may be nearly as large as the direct discrimination effect. As a result, these results suggest the need for further guidance from the EEOC to employers to avoid age-stereotyped job-ad language that deters older workers from applying for jobs. Using language that explicitly deters older workers from applying is already illegal under the ADEA, but the subtler usage of ageist language that we study suggests that job-ad language that would not be flagged as explicitly illegal can still have pernicious effects on older workers in the labor market, and possibly facilitate age discrimination. Moreover, the EEOC might consider flagging for potential investigation firms that use age-stereotyped language in their job ads, recognizing that, for these firms, discrimination may be occurring even in the absence of shortfalls between the share of older applicants hired and the share of older workers who apply for jobs. The findings also imply that, in assessing evidence of age discrimination in hiring, courts may need to put more weight on evidence aside from differences between the shares of older workers among hires and among job applicants, as the share of older workers among job applicants may itself reflect the discrimination that occurs through job-ad language. Finally, of course, these same considerations may apply to discrimination against other protected groups, but such an assessment awaits research on these groups using methods similar to ours.

References

Adams, Scott. 2004. "Age Discrimination Legislation and the Employment of Older Workers." *Labour Economics* 11: 219-41.

Ajunwa, Ifeoma. 2019. "Age Discrimination by Platforms." *Berkeley Journal of Employment and Labor Law* 40: 1-27.

Baert, Stijn, Jennifer Norga, Yannick Thuy, and Marieke Van Hecke. 2016. "Getting Grey Hairs in the Labour Market. An Alternative Experiment on Age Discrimination." *Journal of Economic Psychology* 57: 86-101.

Banfi, S. and Villena-Roldán, B. 2019. "Do High-Wage Jobs Attract More Applicants? Directed Search Evidence from the Online Labor Market." *Journal of Labor Economics* 37: 715-746.

Belot, Michele, Philipp Kircher, and Paul Muller. 2018. "How Wage Announcements Affect Job Search – A Field Experiment." IZA Discussion Paper No. 11814.

Bem, S.L., and Bem, D.J., 1973. "Does Sex-biased Job Advertising "Aid and Abet" Sex Discrimination?" *Journal of Applied Social Psychology* 3: 6-18.

Bendick, Marc, Jr., Lauren E. Brown, and Kennington Wall. 1999. "No Foot in the Door: An Experimental Study of Employment Discrimination Against Older Workers." *Journal of Aging & Social Policy* 10: 5-23.

Bendick, Marc, Jr., Charles W. Jackson, and J. Horacio Romero. 1997. "Employment Discrimination Against Older Workers: An Experimental Study of Hiring Practices." *Journal of Aging & Social Policy* 8: 25-46.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B* 57: 289-300.

Burn, Ian, Patrick Button, Luis Munguia Corella, David Neumark. 2022. "Does Ageist Language in Job Ads Predict Age Discrimination in Hiring?" *Journal of Labor Economics* 3: 613-67.

Burn, Ian, Daniel Firoozi, Daniel Ladd, and David Neumark. 2021. "Machine Learning and Perceived Age Stereotypes in Job Ads: Evidence from an Experiment." NBER Working Paper No. 28328.

Brenoff, Ann. 2019. "5 Ageist Phrases to be Aware of." AARP: Washington, DC.

Cahill, Kevin E, Michael D. Giandrea, and Joseph F. Quinn. 2006. "Retirement Patterns from Career Employment." *The Gerontologist* 46: 514-23.

Carlsson, Magnus, and Stefan Eriksson. 2019. "Age Discrimination in Hiring Decisions: Evidence from a Field Experiment in the Labor Market." *Labour Economics* 59: 173-83.

Combs, Deborah H. 1982. "Striking a Balance between the Interests of Public Safety and the Rights of Older Workers: The Age BFOQ Defense." *Washington and Lee Law Review* 39: 1371-95.

Chaturvedi, Sugat and Mahajan, Kanika and Siddique, Zahra. 2021. "Words Matter: Gender, Jobs and Applicant Behavior." IZA Discussion Paper No. 14497.

Deming, David, and Lisa B. Kahn. 2018. "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals." *Journal of Labor Economics* 36 (S1): S337-69.

European Commission. 2000. *Council Directive 2000/78/EC of 27 November 2000 Establishing a General Framework for Equal Treatment in Employment and Occupation*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0078.

Farber, Henry S., Silverman, Daniel., and Till von Wachter. 2017. "Factors Determining Callbacks to Job

Applications by the Unemployed: An Audit Study.*” RSF: The Russell Sage Foundation Journal of the Social Sciences* 3: 168-201.

Farber, Henry S., Herbst, Chris M., Silverman, Daniel, and Till von Wachter, T. 2019. “Whom Do Employers Want? The Role of Recent Employment and Unemployment Status and Age.” *Journal of Labor Economics* 37(2): 323-49.

Federal Register. (n.d.). *Disparate Impact and Reasonable Factors Other Than Age Under the Age Discrimination in Employment Act.* https://www.federalregister.gov/documents/2012/03/30/2012-5896/disparate-impact-and-reasonable-factors-other-than-age-under-the-age-discrimination-in-employment (viewed September 15, 2019).

Flory, Jeffrey A., Andreas Leibbrandt, and John A. List. 2015. “Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions.” *Review of Economic Studies* 82: 122-55.

Flory, Jeffrey A., Andreas Leibbrandt, Christina Rott, and Olga Stoddard. 2021. “Increasing Workplace Diversity: Evidence from a Recruiting Experiment at a Fortune 500 Company.” *Journal of Human Resources* 57: 73-92.

Gaucher, Danielle, Justin Friesen, and Aaron C. Kay. 2011. “Evidence that Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality.” *Journal of Personality and Social Psychology* 101: 109-28.

Gordon, Randall A., and Richard D. Arvey. 2004. “Age Bias in Laboratory and Field Settings: A Meta-Analytic Investigation.” *Journal of Applied Social Psychology* 34: 468-92.

Hanson, Andrew, Zachary Hawley, and Aryn Taylor. 2011. “Subtle Discrimination in the Rental Housing Market: Evidence from email Correspondence with Landlords.” *Journal of Housing Economics* 20: 276-84.

Hanson, Andrew, Zachary Hawley, Hal Martin, and Bo Liu. 2016. “Discrimination in Mortgage Lending: Evidence from a Correspondence Experiment.” *Journal of Urban Economics* 92: 48-65.

He, Haoran, David Neumark, and Qian Weng. 2021. “Do Workers Value Flexible Jobs? A Field Experiment. *Journal of Labor Economics* 39: 709-38.

Hellerstein, Judith K., David Neumark, and Kenneth Troske. 1999. “Wages, Productivity, and Worker Characteristics: Evidence from Plant-Level Production Functions and Wage Equations.” *Journal of Labor Economics* 17: 409-46.

Hellester, Miguel D., Peter Kuhn, and Kailing Shen. 2020. “The Age Twist in Employers’ Gender Requests.” *Journal of Human Resources* 55: 482-69.

Ibanez, Marcela, and Gerhard Riener. 2018. “Sorting through Affirmative Action: Three Field Experiments.” *Journal of Labor Economics* 36:437-78.

Johnson, Richard W., Janette Kawachi, and Eric K. Lewis. 2009. “Older Workers on the Move: Recareering in Later Life.” AARP Public Policy Institute. Washington, DC: AARP.

Kang, Sonia K., Katherine A. DeCelles, and András Tilcsik. “Whitened Résumés: Race and Self-Presentation in the Labor Market.” *Administrative Science Quarterly* 61: 469-502.

Krause, Annabelle, Ulf Rinne, and Klaus Zimmerman. 2012. “Anonymous Job Applications of Fresh Ph.D. Economists.” *Economics Letters* 117: 441-4.

Kuhn, Peter, and Kailing Shen. 2013. “Gender Discrimination in Job Ads: Evidence from China.” *Quarterly Journal of Economics* 128: 287-336.

Kuhn, Peter, Kailing Shen, and Shuo Zhang. 2018. “Gender-Targeted Job Ads in the Recruitment Process: Evidence from China.” NBER Working Paper No. 25365.

Lahey, Joanna. 2010. “International Comparisons of Age Discrimination Laws.” *Research on Aging* 32: 679-97.

Lahey, Joanna. 2008. “Age, Women, and Hiring: An Experimental Study.” *Journal of Human Resources* 43: 30-56.

Leibbrandt, Andreas, and John List. 2018. “Do Equal Employment Opportunity Statements Backfire? Evidence from a Natural Field Experiment on Job-Entry Decisions.” NBER Working Paper No. 25035.

Maestas, Nicole. 2010. “Back to Work: Expectations and Realizations of Work after Retirement.” *Journal of Human Resources* 45: 718-48.

Marinescu, Ioana Elena, and Ronald Wolthoff. 2020. “Opening the Black Box of the Matching Function: The Power of Words.” *Journal of Labor Economics* 38: 535-568.

Modestino, Alicia Sasser, Daniel Shoag, and Joshua Ballance. 2016. “Downskilling: Changes in Employer Skill Requirements over the Business Cycle.” *Labour Economics* 41: 333-47.

Neumark, David, Ian Burn, and Patrick Button. 2016. “Experimental Age Discrimination Evidence and the Heckman Critique.” *American Economic Review Papers and Proceedings* 106: 303-8.

Neumark, David, Ian Burn, and Patrick Button. 2019a. “Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment.” *Journal of Political Economy* 127: 922-70.

Neumark, David, Ian Burn, Patrick Button, and Nanneh Chehras. 2019b. “Do State Laws Protecting Older Workers from Discrimination Reduce Age Discrimination in Hiring? Evidence from a Field Experiment.” *Journal of Law and Economics* 62: 373-402.

Pallais, Amanda. 2014. “Inefficient Hiring in Entry-Level Labor Markets.” *American Economic Review* 104: 2565-99.

Pillar, Anna, Kyrill Poelmans, and Martha Larson. 2022. “Regex in a Time of Deep Learning: The Role of an Old Technology in Age Discrimination Detection in Job Advertisements.” *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pp. 13-18.

Riach, Peter A., and Judith Rich. 2006. “An Experimental Investigation of Age Discrimination in the French Labour Market.” IZA Discussion Paper No. 2522.

Riach, Peter A., and Judith Rich. 2010. “An Experimental Investigation of Age Discrimination in the English Labor Market.” *Annals of Economics and Statistics* 99/100: 169-85.

Ryan, Ellen Bouchard, Sheree Kwong See, W. Bryan Meneer, and Diane Trovato. 1992. “Age-Based Perceptions of Language Performance Among Younger and Older Adults.” *Communication Research* 19: 423-43.

Terrell, Kenneth. 2019. “Age Bias That’s Barred by Law Appears in Thousands of Job Listings.” AARP: Washington, DC.

Tilcsik, András. 2011. “Pride and Prejudice: Employment Discrimination against Openly Gay Men in the United States.” *American Journal of Sociology* 117: 586-626.

van Borm, Hannah, Ian Burn, and Stijn Baert. 2021. “What Does a Job Candidate’s Age Signal to Employers?” *Labour Economics* 17: 102003.

Wax, S. L. 1948. “A Survey of Restrictive Advertising and Discrimination by Summer Resorts in the Province of Ontario.” *Canadian Jewish Congress.*

Figure 1: Map of Cities in Experiment

Note: This map shows the cities in the experiment. The relative size of the symbol corresponds to the total number of applicants in each city, ranging from 23 in Salt Lake City to 643 in New York City. The total number of applicants is 2,646.

Figure 2: Examples of Job Ad Templates

Administrative Assistants Template 1 (Admin Assistant)
Psychiatric office is in need of a full or part time Administrative Assistant to assist in front/back office general clerical duties. This individual will work on a several tasks and stay on course at all times. The Administrative Assistant we hire will be trained in various duties that cover the entire office.

This individual MUST possess the following:

-Exceptional customer service background to greet and register patients, answer phones, schedule appointments.
-Can multitask.
-High School diploma or GED.
-Professional attitude.
-*__Communication Skill Requirement__*.
-*__Technology Requirement__*
-*__Physical Requirement__*
-Available for flexible hours.
(Schedule hours and days will alternate every other week)

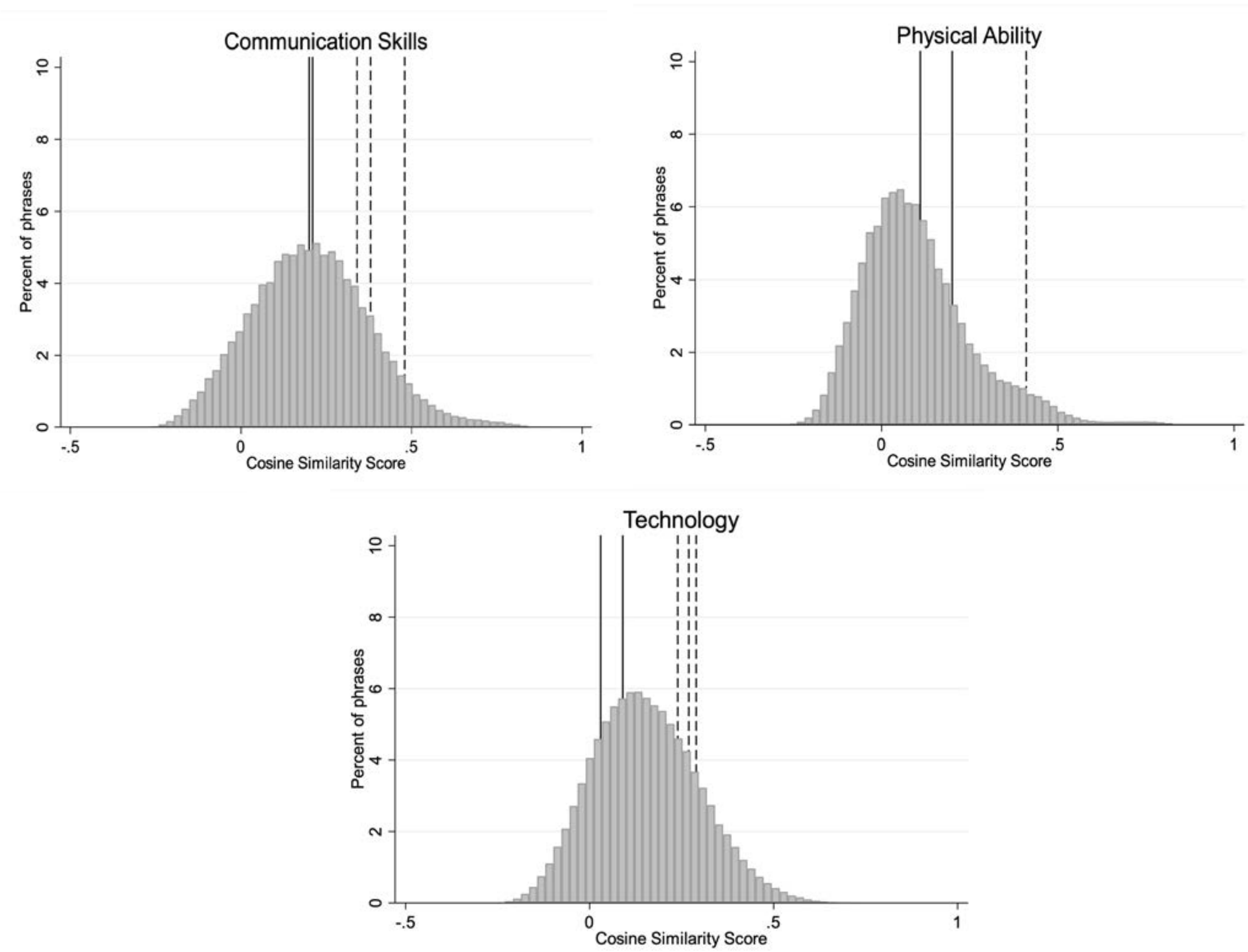
Please email us a CV or resume and put “full-time” or “part-time” in the subject line.

Retail Sales Associate Template 1 (Retail Sales Job)
Our women’s clothing store in *__City__* is looking for a sales associate to help us out weekday afternoons. We are pretty busy store and you must *__Physical Requirement__*__.__ We are looking for someone with open to working in retail, who *__Communication Skill Requirement__*__.__ We need you to *__Technology Requirement__*. So if this sounds like you, send us your resume and your earliest possible starting date and we will be in touch.

Security Guard Template 1 (HIRING UNARMED SECURITY GUARDS)
We currently have a position for a full-time or part-time security officer available. Training and uniforms will provided. We offer flexible working hours and have shifts any day of the week. Our pay scale is competitive. Email your resume and potential work hours to apply.

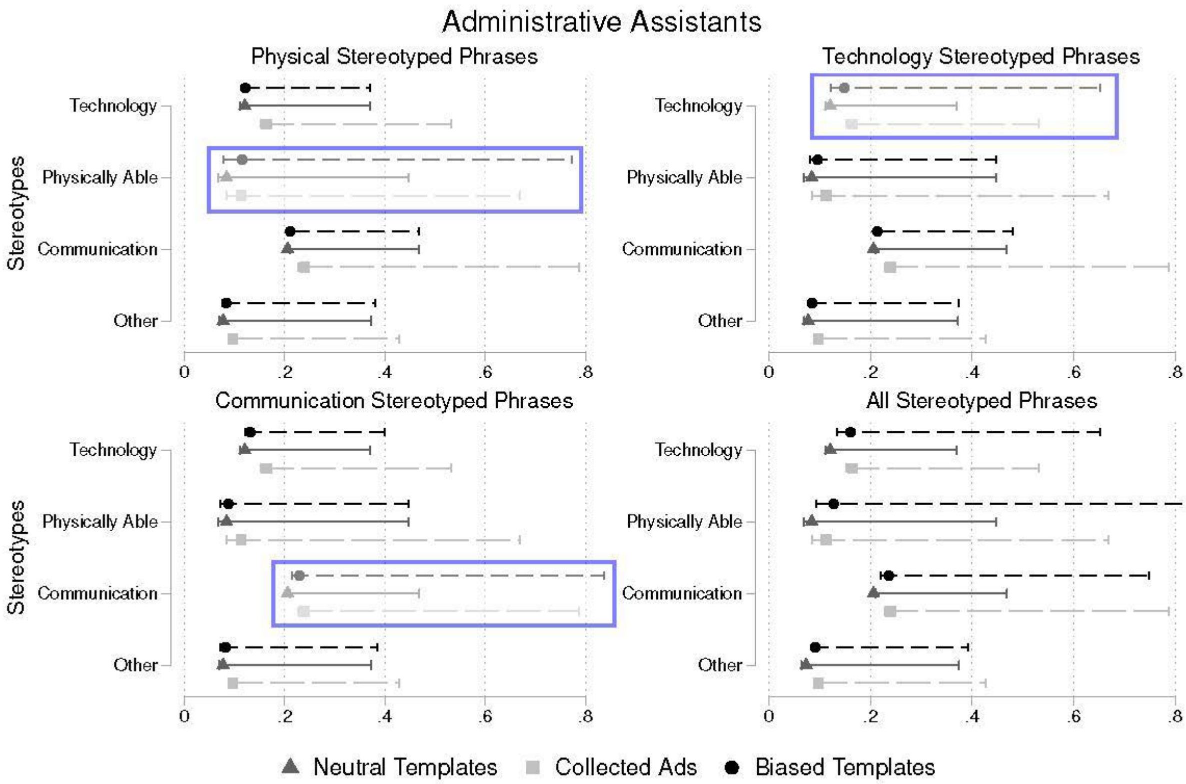
Requirements
-Professional appearance & attitude
-Detail oriented
-*__Communication Skill Requirement__*
-*__Physical Requirement__*
-*__Technology Requirement__*
-At least 18 years of age
-Access to transportation

Figure 3: Locations of Treatment and Control Phrases in the Cosine Similarity Score Distribution of Job Ad Phrases



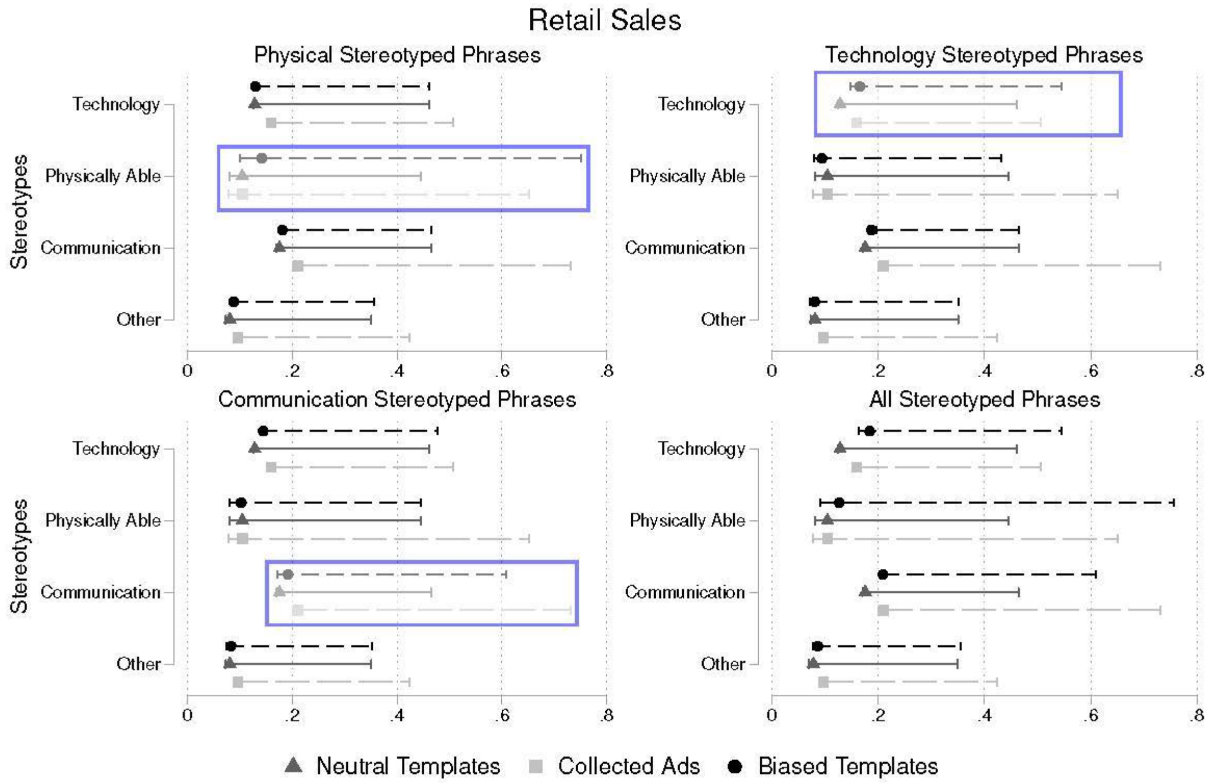
Note: Figure reports the distributions of cosine similarity scores for all trigrams from the job ads with the indicated stereotypes. The higher the cosine similarity score, the more related the trigram is to the stereotype, with a minimum of −1 and a maximum of 1. Solid lines indicate the location of a control sentence in the cosine similarity score distribution. Dashed lines indicate the location of a treatment phrase (for the Machine Learning Treatments shown in Table 2).

Figure 4a: Cosine Similarity Score of Administrative Assistant Templates



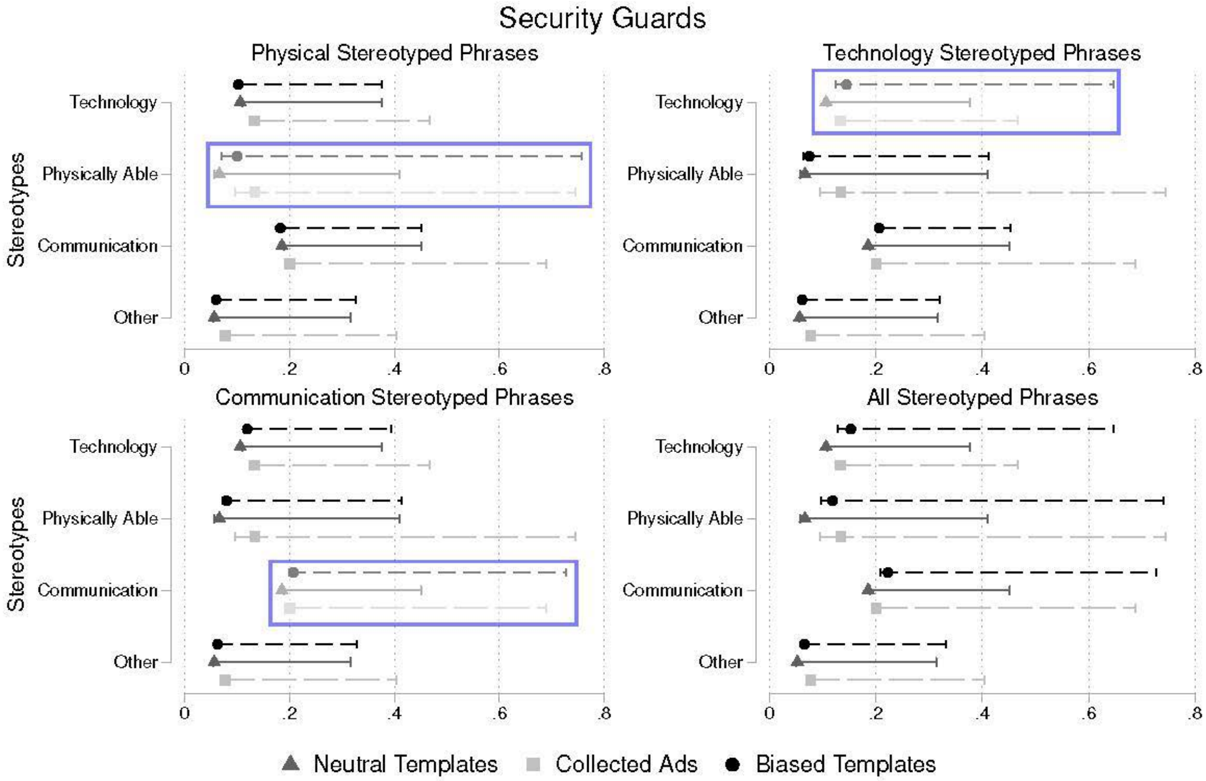
Note: Graphs display median to 99th percentile range of trigram semantic similarity scores for each stereotype for Administrative Assistant ads. The average trigram semantic similarity score for each stereotype is represented by the respective shape for each template. The category "Other" shows the averages for the remaining stereotypes listed in Table 1. Control ("neutral") templates contain trigrams from the created ad templates with only non-stereotyped phrases included. Collected ads comprise trigrams from all Administrative Assistant job ads. Treatment templates contain trigrams from the created ad templates with the respective stereotyped phrase or phrases included.

Figure 4b: Cosine Similarity Score of Retail Sales Templates



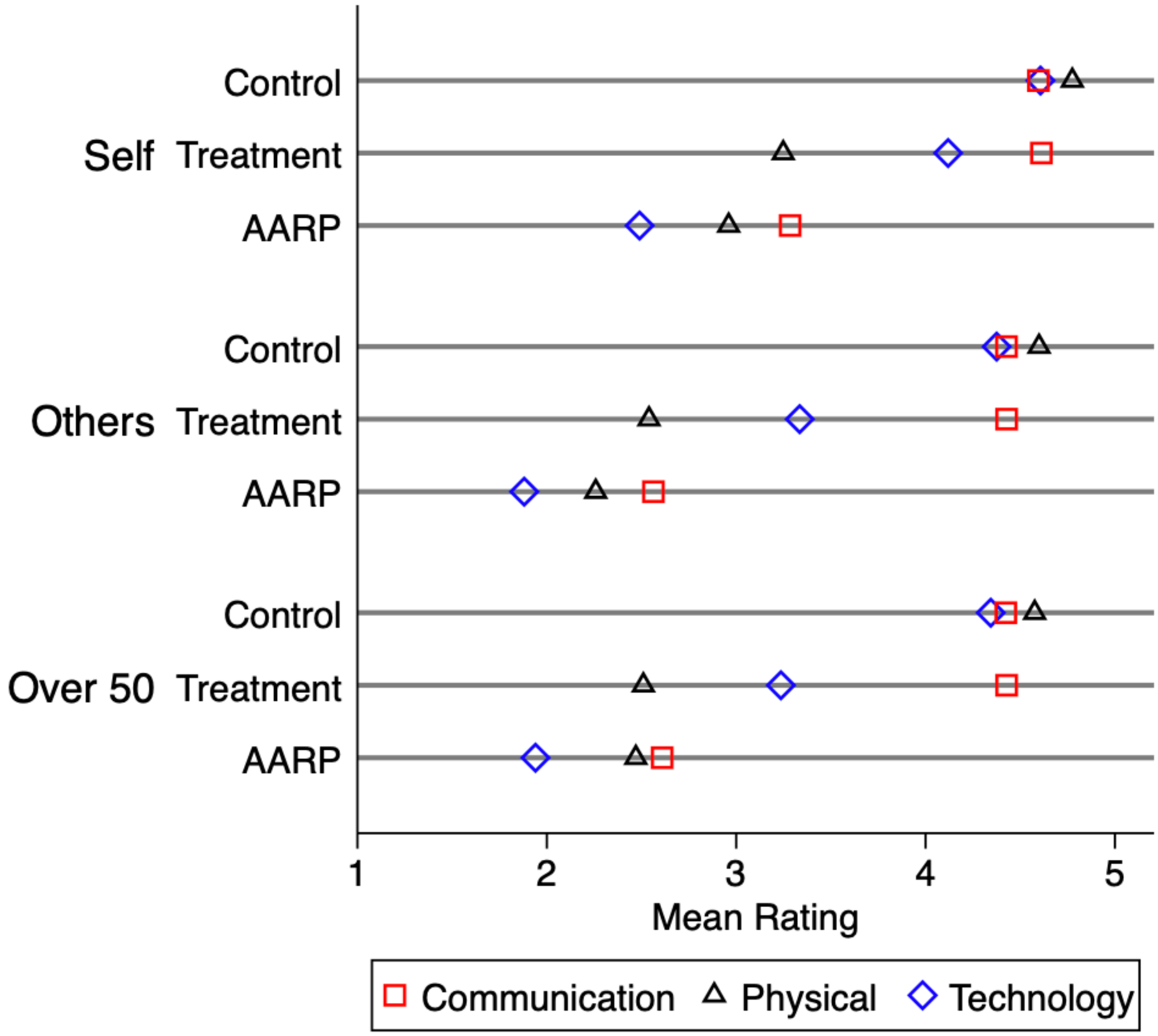
Note: Graphs display median to 99th percentile range of trigram semantic similarity scores for each stereotype for Retail Sales ads. The average trigram semantic similarity score for each stereotype is represented by the respective shape for each template. The category "Other" shows the averages for the remaining stereotypes listed in Table 1. Control ("neutral") templates contain trigrams from the created ad templates with only non-stereotyped phrases included. Collected ads comprise trigrams from all Retail Sales job ads. Treatment templates contain trigrams from the created ad templates with the respective stereotyped phrase or phrases included.

Figure 4c: Cosine Similarity Score of Security Guard Templates



Note: Graphs display median to 99th percentile range of trigram semantic similarity scores for each stereotype for Security Guard ads. The average trigram semantic similarity score for each stereotype is represented by the respective shape for each template. The category "Other" shows the averages for the remaining stereotypes listed in Table 1. Control ("neutral") templates contain trigrams from the created ad templates with only non-stereotyped phrases included. Collected ads comprise trigrams from all Security Guard job ads. Treatment templates contain trigrams from the created ad templates with the respective stereotyped phrase or phrases included.

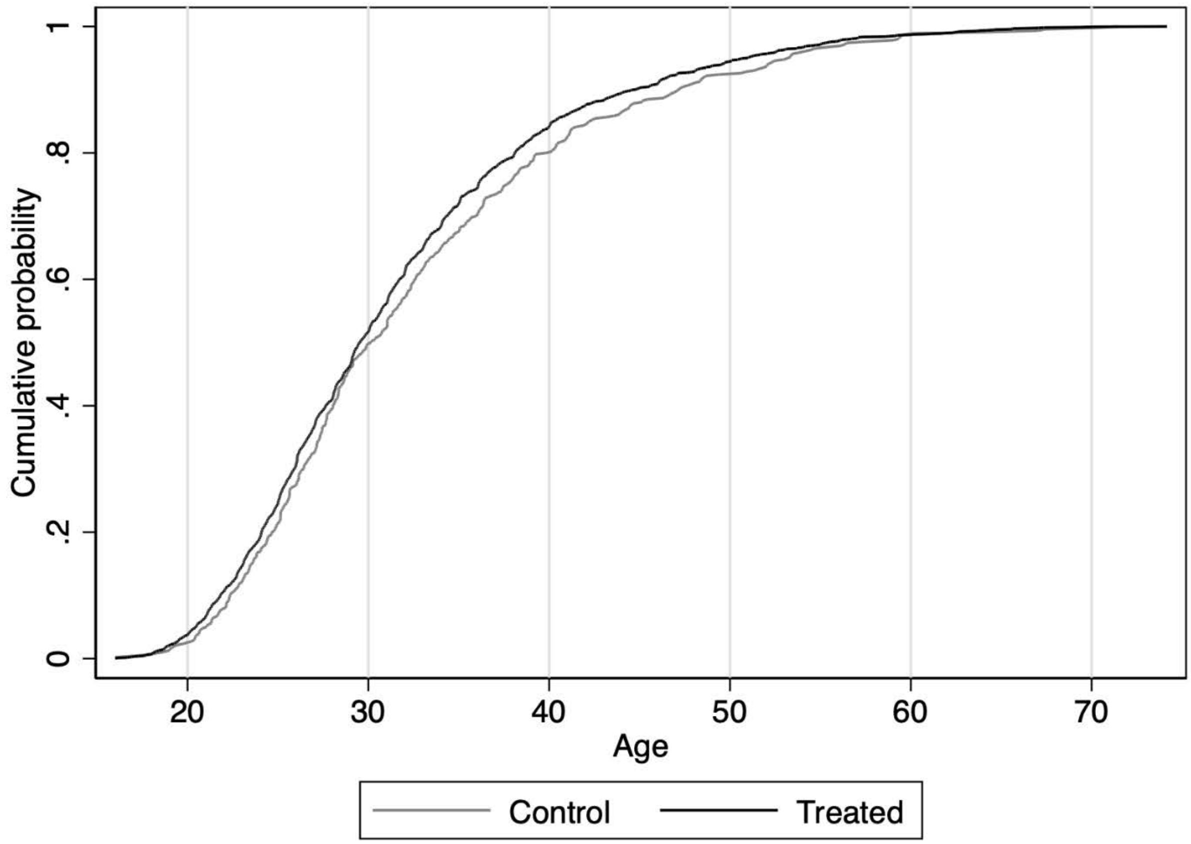
Figure 5: Survey Results



Note: These numerical ratings reflect the degree to which survey respondents rated phrases as age-biased or not age-biased, with lower numbers indicating a greater bias against older workers. Likert Scale ratings were translated to a numerical value such that “Strongly Agree” mapped to 1, “Somewhat Agree” mapped to 2, “Neither agree nor disagree” mapped to 3, “Somewhat Disagree” mapped to 4, and “Strongly Disagree” mapped to 5. The three categories “Self,” “Others,” and “Over 50” refer to which group’s opinions the MTURK respondents were asked to provide or predict in a given survey block. The average bias rating was collapsed on the treatment status of phrases (control, treatment, and AARP) as well as the category of the stereotype (communication, physical, or technology). Hence, each point in the figure above reflects the average bias rating MTURK respondents gave to a given treatment status for a specific stereotype from the perspective of a given group of people. For example, the triangle in the first row of the figure indicates that when respondents were asked for their self-assessment of whether or not the physical stereotype control phrases were age-biased, they, on average, stated that they strongly disagreed.

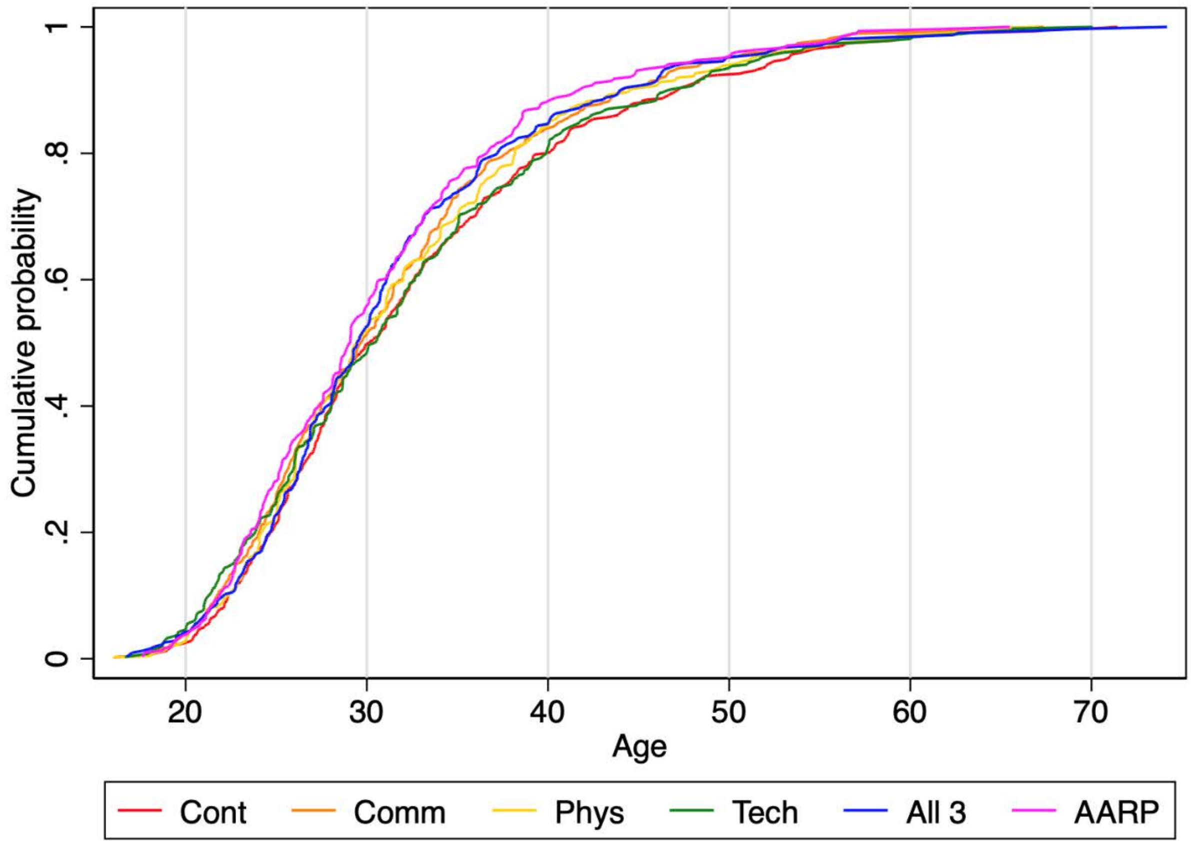
Figure 6: Posting Job Ads Was Not Easy!

Figure 7: Empirical Cumulative Density Functions, Any Treatment



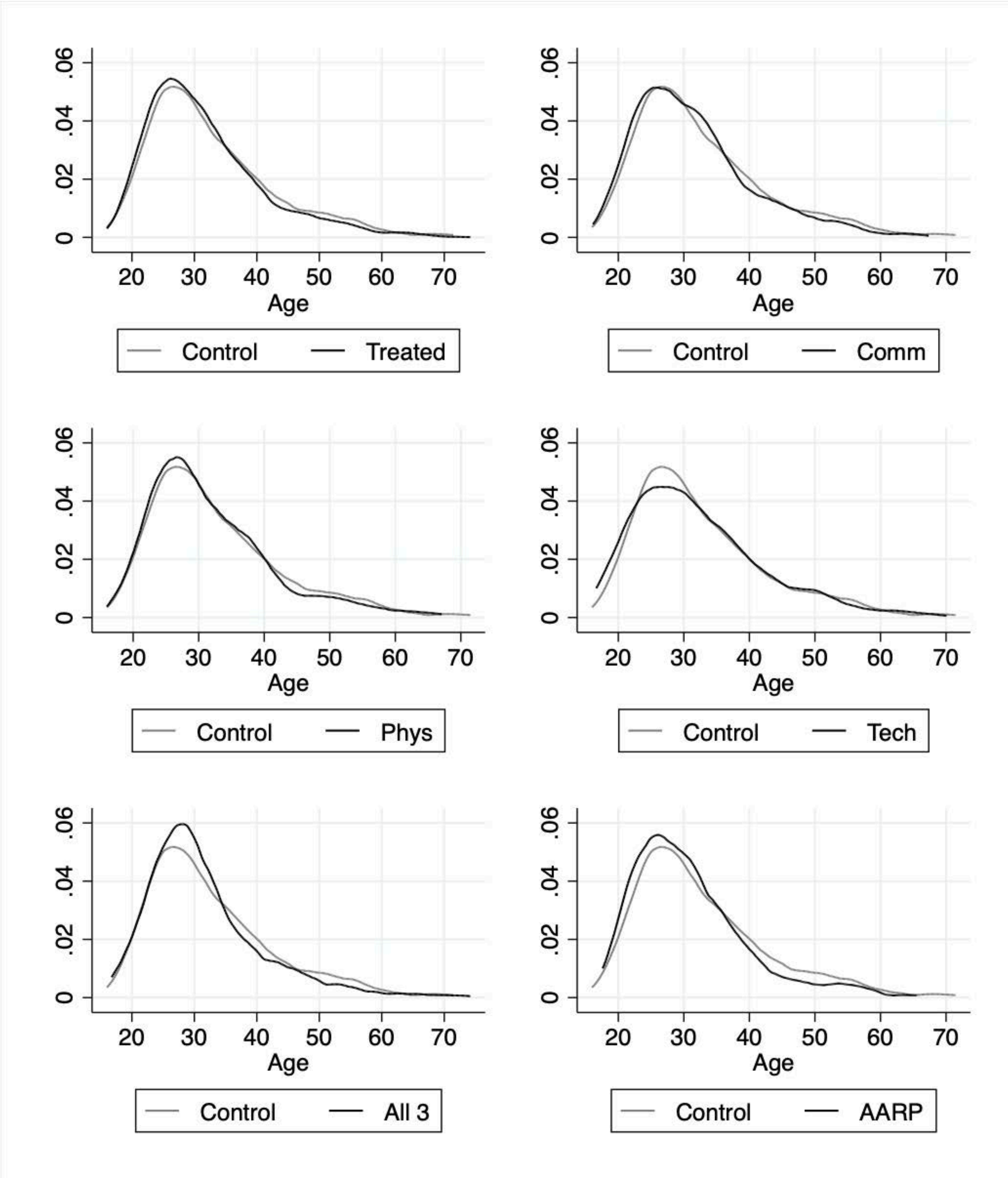
Note: “Treated” refers to any treatment (individual stereotypes, *All 3*, or *AARP*).

Figure 8: Empirical Cumulative Density Functions by Ad Type



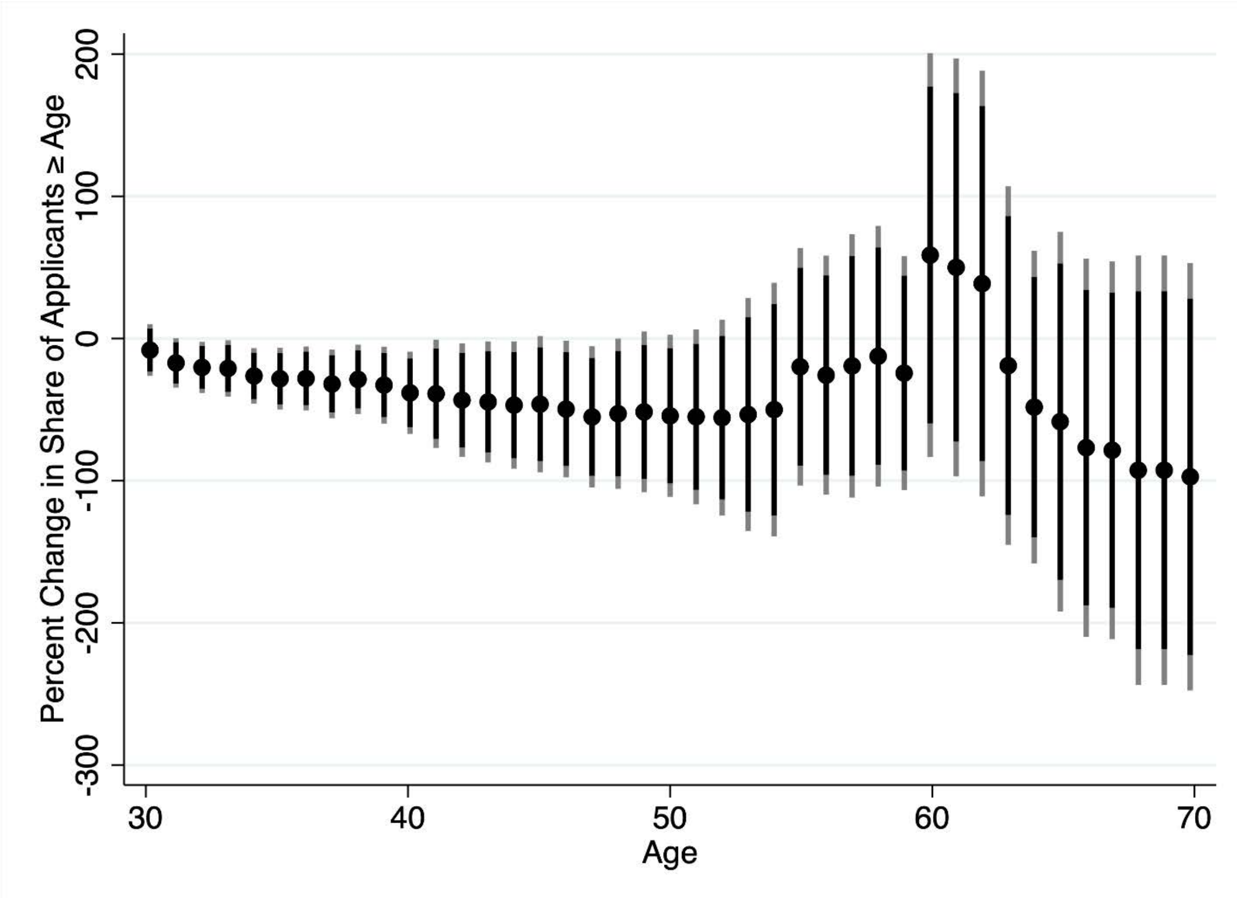
Note: “Cont” refers to controls; “Comm” to communications skills stereotypes; “Phys” to physical ability stereotypes; “Tech” to technology stereotypes”; “All3” to the ads with all three stereotypes reflected in the text; and “AARP” to the ads with AARP ageist language/stereotypes.

Figure 9: Empirical Probability Density Functions for Age



Note: In the upper-left panel, “Treated” refers to any treatment (individual stereotypes, all 3, or AARP). The other labels are explained in the notes to Figure 8.

Figure 10: Estimated Effects of Any Treatment on Proportion Above Each Age Threshold, Scaled by Proportion of Total Applicants Above Age Threshold



Note: 90% and 95% confidence intervals are shown, based on the regression estimates.

Table 1: Age Stereotypes from Industrial Psychology Literature

Health	Personality	Skills
Less Attractive	Less Adaptable	Lower Ability to Learn
Hard of Hearing	Careful	Better Communication Skills
Worse Memory	Less Creative	Worse Communication Skills
Less Physically Able	Dependable	More Experienced
	Negative Personality	More Productive
	Warm Personality	Less Productive
		Worse with Technology

Note: See Burn et al. (2022).

Table 2: Control and Treatment Phrases by Occupation

Occupation	Stereotype	Control	Machine Learning Treatment	AARP Treatment
(1)	(2)	(3)	(4)	(5)
Administrative Assistants	Communication skills	You must be good at working without supervision	You must have good communication and teamwork on tasks	You must be up-to-date with current industry jargon and communicate with a dynamic workforce
Administrative Assistants	Physical ability	You must enter bills and keep track of invoices	You must be able to lift 40 pounds	You must be a fit and energetic person
Administrative Assistants	Technical skills	You must produce and distribute documents such as correspondence memos, faxes and forms	You must use accounting software systems like Netsuite, Freshbook, and QuickBooks	You must be a digital native and have a background in social media
Retail sales	Communication skills	You must be good at working without supervision	You must have good communication with customers and staff	You must be up-to-date with current industry jargon and communicate with a dynamic workforce
Retail sales	Physical ability	You must enter bills and keep track of invoices	You must be able to lift 40 pounds	You must be a fit and energetic person
Retail sales	Technical skills	You must help to clean and organize the store	You must use software such as Microsoft Office/Excel or Google Sheets	You must be a digital native and have a background in social media
Security guard	Communication skills	You must follow instruction from supervisors	You must maintain communication about tasks with supervisors	You must be up-to-date with current industry jargon and communicate with a dynamic workforce
Security guard	Physical ability	You need to carry a flashlight	You must be able to lift 50 pounds	You must be a fit and energetic person
Security guard	Technical skills	You must write patrol records in journal notebook	You must type patrol entries into a journal application on a computer system	You must be a digital native and have a background in social media

Note: See text for a description of how each sentence was created.

Table 3: Estimated Effects on Age Composition of Applicants, Any Stereotype Treatment, and Distinguishing AARP Treatment, All Cities

	Average Age	Median Age	75th Percentile	Over 40	No Age Information
Any Treatment	**-2.687*****†††	**-2.680****†††	**-3.133****††	**-0.094****†††	0.030
	(0.990)	(1.028)	(1.523)	(0.035)	(0.038)
N	228	228	228	228	237
Any Treatment	**-2.241****††	**-2.391****††	-2.474†	**-0.079****††	0.030
	(0.990)	(1.031)	(1.560)	(0.037)	(0.037)
AARP	**-2.318*****†††	**-1.497****††	**-3.420*****†††	**-0.077*****†††	-0.003
	(0.745)	(0.709)	(1.226)	(0.026)	(0.040)
N	228	228	228	228	237

Note: The regressions include all treatment arms and the control arm. In the second panel, the *AARP* variable is equivalent to the interaction between *Any Treatment* and *AARP*. Bold horizontal lines distinguish separate regressions. All specifications include fixed effects for both city and occupation. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test. In first four columns, boldface estimates indicates statistical significance at the 5% level in a two-sided tests, correcting for multiple testing across the estimates in those columns, using the Simes False Discovery Rate (q-values ≤ 0.05). Italicized estimates indicate q-values > 0.05 and ≤ 0.1. (See Table A1 in online Appendix A.)

Table 4: Estimated Effects on Age Composition of Applicants, Separate Stereotype Treatments, All Cities

	Average Age	Median Age	75th Percentile	Over 40	No Age Information
Communication	-2.632**††	-2.948*††	-2.583	-0.070†	0.014
	(1.232)	(1.460)	(2.043)	(0.050)	(0.040)
N	79	79	79	79	81
Physical	-1.828†	-2.062†	-2.454	-0.082†	0.022
	(1.403)	(1.394)	(2.134)	(0.050)	(0.053)
N	79	79	79	79	81
Technology	-1.930†	-2.105†	-0.728	-0.044	0.051
	(1.221)	(1.285)	(1.945)	(0.042)	(0.047)
N	79	79	79	79	81

Note: Each regression includes a single machine learning stereotype arm and the control arm. Bold horizontal lines distinguish separate regressions. All specifications include fixed effects for both city and occupation. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test. In first four columns, boldface estimates indicates statistical significance at the 5% level in a two-sided tests, correcting for multiple testing across the estimates in those columns, using the Simes False Discovery Rate (q-values ≤ 0.05). Italicized estimates indicate q-values > 0.05 and ≤ 0.1. (See Table A1 in online Appendix A.)

Table 5: Estimated Effects on Age Composition of Applicants, All Treatment Arms, All Cities

	Average Age	Median Age	75th Percentile	Over 40	No Age Information
Communication	*-2.683***††	*-2.986***††	-2.704†	-0.075†	0.014
	(1.147)	(1.350)	(1.898)	(0.046)	(0.039)
Physical	-1.879†	-2.071†	-2.385	-0.083*††	0.034
	(1.288)	(1.276)	(2.000)	(0.045)	(0.051)
Technology	-1.889†	-2.002†	-0.707	-0.041	0.056
	(1.165)	(1.222)	(1.822)	(0.041)	(0.045)
All 3	*-2.516***††	*-2.504***††	*-4.156***††	**-0.117*****†††	0.016
	(1.122)	(1.104)	(1.754)	(0.041)	(0.048)
AARP	**-4.559*****†††	**-3.888*****†††	**-5.896*****†††	**-0.156*****†††	0.027
	(1.222)	(1.212)	(1.799)	(0.038)	(0.057)
N	228	228	228	228	237

Note: All specifications include fixed effects for both city and occupation. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test. In first four columns, boldface estimates indicates statistical significance at the 5% level in a two-sided tests, correcting for multiple testing across the estimates in those columns, using the Simes False Discovery Rate (q-values ≤ 0.05). Italicized estimates indicate q-values > 0.05 and ≤ 0.1. (See Table A1 in online Appendix A.)

Table 6: Estimated Effects on Age Composition of Applicants, Cosine Similarity Score (or Average), or MTURK Likert Scale (or Average) of Treatment (Standardized), All Cities

	Average Age	Median Age	75th Percentile	Over 40	No Age Information
CSS	**-1.722****†††	**-1.534****††	**-3.139*****†††	**-0.083*****†††	0.008
	(0.652)	(0.637)	(0.972)	(0.023)	(0.031)
N	228	228	228	228	237
Likert score (perceived age bias)	**-2.048*****†††	**-1.524****†††	**-3.354*****†††	**-0.085*****†††	0.004
	(0.652)	(0.637)	(0.972)	(0.023)	(0.031)
N	228	228	228	228	237

Note: In the top panel, the treatment is the cosine similarity score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. In the bottom panel, the treatment is average MTURK Likert score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. The sign is switched from Figure 5 so that a higher value implies job-ad language perceived as more biased against older workers. Bold horizontal lines distinguish separate regressions. All specifications include fixed effects for both city and occupation. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test. In first four columns, boldface estimates indicates statistical significance at the 5% level in a two-sided tests, correcting for multiple testing across the estimates in those columns, using the Simes False Discovery Rate (q-values ≤ 0.05). Italicized estimates indicate q-values > 0.05 and ≤ 0.1. (See Table A1 in online Appendix A.)

Table 7A: Estimated Effects on Age Composition of Applicants, Different Age Cutoffs, All Treatment Arms, All Cities

	Over 40	Over 50	Over 65
Communication	-0.075†	-0.073*††	-0.005
	(0.046)	(0.037)	(0.010)
Physical	-0.083*††	-0.042	-0.009
	(0.045)	(0.040)	(0.007)
Technology	-0.041	-0.038	-0.004
	(0.041)	(0.036)	(0.009)
All 3	-0.117***†††	-0.081**††	-0.006
	(0.041)	(0.034)	(0.007)
AARP	-0.156***†††	-0.090**†††	-0.008
	(0.038)	(0.035)	(0.008)
N	228	228	228

Note: All specifications include fixed-effects for both city and occupation. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test.

Table 7B: Estimated Effects on Age Composition of Applicants, Cosine Similarity Score (or Average), or MTURK Likert Scale (or Average) of Treatment (Standardized), Different Age Cutoffs, All Cities

	Over 40	Over 50	Over 65
CSS	-0.083***†††	-0.051***†††	-0.004
	(0.023)	(0.019)	(0.004)
N	228	228	228
Likert Score (Perceived Age Bias)	-0.085***†††	-0.040***†††	-0.004
	(0.017)	(0.014)	(0.004)
N	228	228	228

Note: In the top panel, the treatment is the cosine similarity score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. In the bottom panel, the treatment is average MTURK Likert score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. The sign is switched from Figure 5 so that a higher value implies job-ad language perceived as more biased against older workers. Bold horizontal lines distinguish separate regressions. All specifications include fixed effects for city, occupation, month of posting, and day-of-week of posting. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test.

Table 8: Estimated Effects on Age Composition of Applicants, with Month and Day-of-Week Fixed Effects, All Treatment Arms, Cosine Similarity Score (or Average), or MTURK Likert Scale (or Average) of Treatment (Standardized), All Cities

	Average Age	Median Age	75^{th} Percentile	Over 40	No Age Information
Communication	$-3.155^{**\dagger\dagger\dagger}$	$-3.094^{**\dagger\dagger}$	$-3.291^{\dagger}$	$-0.087^{*\dagger\dagger}$	-0.003
	(1.270)	(1.482)	(2.163)	(0.046)	(0.044)
Physical	-1.873	-1.769	-2.620	$-0.087^{*\dagger\dagger}$	0.017
	(1.492)	(1.460)	(2.249)	(0.048)	(0.049)
Technology	$-2.222^{*\dagger\dagger}$	$-2.145^{\dagger}$	-1.017	-0.056	0.046
	(1.288)	(1.293)	(2.056)	(0.048)	(0.041)
All 3	$-2.383^{*\dagger\dagger}$	$-2.160^{*\dagger\dagger}$	$-3.786^{*\dagger\dagger}$	$-0.105^{**\dagger\dagger}$	0.010
	(1.209)	(1.233)	(1.960)	(0.045)	(0.050)
AARP	$-4.232^{***\dagger\dagger}$	$-3.281^{**\dagger\dagger\dagger}$	$-5.601^{**\dagger\dagger\dagger}$	$-0.159^{***\dagger\dagger\dagger}$	0.031
	(1.357)	(1.321)	(2.118)	(0.046)	(0.056)
N	228	228	228	228	237
CSS	$-1.369^{*\dagger\dagger}$	$-1.083^{\dagger}$	$-2.606^{**\dagger\dagger}$	$-0.068^{**\dagger\dagger\dagger}$	0.010
	(0.690)	(0.688)	(1.098)	(0.026)	(0.032)
N	228	228	228	228	237
Likert Score (Perceived Age Bias)	$-1.636^{**\dagger\dagger\dagger}$	$-1.035^{*\dagger\dagger}$	$-2.941^{***\dagger\dagger\dagger}$	$-0.080^{***\dagger\dagger\dagger}$	0.010
	(0.642)	(0.589)	(1.001)	(0.021)	(0.030)
N	228	228	228	228	237

Note: In the second panel, the treatment is the cosine similarity score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. In the third panel, the treatment is average MTURK Likert score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. The sign is switched from Figure 5 so that a higher value implies job-ad language perceived as more biased against older workers. Bold horizontal lines distinguish separate regressions. All specifications include fixed effects for city, occupation, month of posting, and day-of-week of posting. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test.

Online Appendix A: Additional Table for Multiple Hypothesis Testing

Table A1: Estimated Effects on Age Composition of Applicants, Any Stereotype Treatment, All Cities, Heterogeneity

Table			Average Age	Median Age	75th Percentile	Over 40
3	*Any Treatment*	p-value	0.009679	0.012675	0.046128	0.010825
		q-value	0.016899	0.016899	0.046128	0.016899
3	*Any Treatment*	p-value	0.028976	0.025425	0.120609	0.036923
		q-value	0.046362	0.046362	0.120609	0.046761
	Any Treatment x Any 4-yr College	p-value	0.003369	0.040916	0.007967	0.005403
		q-value	0.021247	0.046761	0.021247	0.021247
4	*Communication*	p-value	0.038823	0.050224	0.213521	0.166585
		q-value	0.100449	0.100449	0.213521	0.213521
4	*Physical*	p-value	0.199927	0.14688	0.256707	0.107496
		q-value	0.256707	0.256707	0.256707	0.256707
4	*Technology*	p-value	0.121839	0.108966	0.710265	0.301128
		q-value	0.243679	0.243679	0.710265	0.401504
5	*Communication*	p-value	0.024268	0.032553	0.161749	0.106378
		q-value	0.065106	0.065106	0.190293	0.150205
	Physical	p-value	0.152146	0.112216	0.239974	0.073621
		q-value	0.190183	0.150205	0.266637	0.133857
	Technology	p-value	0.112654	0.109065	0.70008	0.325577
		q-value	0.150205	0.150205	0.70008	0.342712
	All 3	p-value	0.030435	0.028748	0.022642	0.006785
		q-value	0.065106	0.065106	0.065106	0.027142
	AARP	p-value	0.000576	0.002601	0.00214	0.000173
		q-value	0.005761	0.013006	0.013006	0.003454
6	*CSS*	p-value	0.011656	0.020552	0.002435	0.001053
		q-value	0.015541	0.020552	0.004869	0.004213
6	*Likert score (perceived age bias)*	p-value	0.000828	0.010574	0.000208	1.66E-05
		q-value	0.001104	0.010574	0.000416	6.64E-05

Note: The q-value is based on the Simes False Discovery Rate. This is computed using all the estimates in a particular table (for the first four columns that measure the effects of the treatments on the ages of applicants), across the different outcomes and stereotype treatments in that table.

Online Appendix B: Additional Robustness Checks from the Pre-Analysis Plan (PAP)

In this appendix, we discuss additional robustness analyses specified in our PAP. First, as noted above, we added additional cities relative to the original PAP, in part because of (anticipated) difficulties with posting ads in some cities. Table B1 thus presents results for the original cities only (which excludes the cities where we could not post ads). The results are very similar. The point estimates are somewhat larger, but so are the standard errors, with the net result that the strength of the statistical conclusions remains very similar.

Second, we weight the estimates by the number of applicants in each city-occupation cell.[47] This makes the estimates representative of applicants rather than cities and occupations. The weighting reduces the scale of the estimated effects. For example, for the treatment using all three machine-learning generated ageist phrases, the average age of applicants was lower by 1.9 years, compared with 2.5 years in Table 5. But the evidence for the combined treatments in the top panel of Table B2 is as strong statistically – and sometimes stronger. And the evidence for the age bias indexes in the middle and bottom panels is similarly strong statistically.[48]

Third, we consider whether the effects of age biased job-ad language varies with the local unemployment rate, and report the results in Table B3.[49] We report these estimates for the *Any Treatment* specification, the specification for *Any Treatment* and the *AARP* treatment, and for the continuous measures of age stereotyping.[50] Most of the estimated interactions of the treatment variables with the unemployment rate are positive, consistent with ageist language in job ads having a weaker effect in discouraging older applicants where the unemployment rate is higher; that is perhaps not surprising, since job seekers are likely less selective in deciding where to apply in a high unemployment environment. However, only one interaction (for median age for *Any Treatment*) is statistically significant. And the estimated interactions are small relative to the estimated effect of ageist language in the job ad. For example, in the upper left column, *Any*

---

[47] Our PAP did not specify that we would weight the estimates, but we do so as a robustness check.
[48] Note that the estimated reductions in the proportion of applicants over age 40, discussed earlier, were based on unweighted city observations. For example, Table 5 indicated a reduction of 11.7 percentage points, on a mean of 24.5%. These estimates are not strictly comparable, because the latter percentage was based on individual applicants, and hence implicitly weighted. The corresponding estimate in the top panel of Table 10 is a 7.5 percentage point reduction – which is still very large relative to the base. (And the CDFs in Figure 8 are also based on individual – and hence implicitly weighted – data.)
[49] This is measured at the MSA-month level, from the Local Area Unemployment Statistics. The interactive specifications in this table and those that follow always include the main effects as well, although these are not reported.
[50] We prefer not to estimate models with interactions for the model with all arms of the experiment, since that leads to a considerably larger large number of coefficient estimates. However, Table B4 shows these full estimates for the unemployment rate; the results are similar.

Treatment lowers average age by 6.2 years, and the differential is smaller by 0.7 years if the unemployment rate is one percentage point higher. Only if unemployment rates differed by something on the order of 8 or so percentage points would the implied effect of the ageist job language be eliminated.

Finally, we explore evidence on differences in results by occupation, in Table B5. The top panel reports results for any treatment. The estimates are, as we would expect, more variable in this case. But every estimate for the 12 age measures (in the first four columns) is negative, and many are statistically significant on some of the criteria we have considered. There is some evidence that the largest age reductions (for both the average and the median) occurred for the security jobs, which may be because of a perception that physical abilities, in particular, are viewed as most important for older applicants. But the effects at the 75th percentile and on the proportion over 40 are similar for retail. The remainder of the table reports estimates for the *Any Treatment* and the *AARP* treatment, and for the continuous measures of age stereotyping. Two things stand out in further showing that the results are robust across occupations. First, every single estimate of the effect of treatment on age of applicants (the first four columns) is negative. Second, across the different models, there are always negative and significant estimates for each of the three occupations.[51]

[51] An additional heterogeneity analysis listed in the PAP was to estimate models with interactions between the treatments and previous occupational experience, education, current unemployment (of the applicant), and gender. In retrospect, however, this analysis is not so easily interpretable because the job ads do not target job searchers based on their characteristics, so that nonzero coefficients on the interactions can simply reflect compositional effects. Nonetheless, since these were included in the PAP, we include them here (as Table B6). For example, when we interact with gender (based on names), there is an "unknown gender" category. However, older people are less likely to have gender-neutral names; for example, 20.2% of male applicants (applicants with male names) were over age 40, vs. only 9.2% of those with gender-neutral names. Thus, the variables with which we form the interactions can reflect the effects of the experiment.

There were a few other secondary analyses outlined in the PAP that proved impractical. We could not study effects on the number of ads in the occupation available when we posted our ads, because this was not captured sufficiently during the data collection. We could not study the speed with the posted ad leaves the first page on the website because this almost never happened in the one-week period we left the ad up. And we could not study interactions with the length of the unemployment spell because this was not expressed with enough consistency on the resumes.

Finally, we had indicated in the PAP that we would look at results inferring age from year of college graduation when we did not have year of high school graduation. However, the approach we use to approximate age (see the Methods section) already accounts for these cases.

Table B1: Estimated Effects on Age Composition of Applicants, All Treatment Arms, Cosine Similarity Score (or Average), or MTURK Likert Scale (or Average) of Treatment (Standardized), Original Cities

	Average Age	Median Age	75th Percentile	Over 40	No Age Information
Communication	-3.245**††	-2.958†	-3.444†	-0.079	0.030
	(1.524)	(1.814)	(2.496)	(0.063)	(0.052)
Physical	-2.945*††	-2.166	-4.700*††	-0.118*††	0.047
	(1.732)	(1.732)	(2.569)	(0.064)	(0.066)
Technology	-2.204†	-2.029	-0.550	-0.040	0.057
	(1.620)	(1.685)	(2.514)	(0.054)	(0.057)
All 3	-3.066*††	-2.717*††	-4.751*††	-0.139**†††	0.018
	(1.540)	(1.513)	(2.460)	(0.055)	(0.068)
AARP	-5.424***†††	-4.670***†††	-6.936***†††	-0.177***†††	0.011
	(1.572)	(1.595)	(2.241)	(0.047)	(0.073)
N	156	156	156	156	165
CSS	-2.037**††	-1.705*††	-3.534**†††	-0.100***†††	0.005
	(0.888)	(0.883)	(1.343)	(0.031)	(0.043)
N	156.000	156.000	156.000	156.000	165.000
Likert Score (Perceived Age Bias)	-2.484***†††	-2.045***†††	-3.917***†††	-0.101***†††	-0.010
	(0.670)	(0.681)	(0.925)	(0.020)	(0.040)
N	156.000	156.000	156.000	156.000	165.000

Note: In the second panel, the treatment is the cosine similarity score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. In the third panel, the treatment is average MTURK Likert score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. The sign is switched from Figure 5 so that a higher value implies job-ad language perceived as more biased against older workers. Bold horizontal lines distinguish separate regressions. All specifications include fixed effects for both city and occupation. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test.

Table B2: Estimated Effects on Age Composition of Applicants, All Treatment Arms, Cosine Similarity Score (or Average), or MTURK Likert Scale (or Average) of Treatment (Standardized), All Cities, Weighted by Number of Applicants for City/Occupation

	Average Age	Median Age	75th Percentile	Over 40	No Age Information
Communication	-1.207*††	-1.110	-0.358	-0.043†	0.012
	(0.714)	(0.860)	(1.424)	(0.031)	(0.022)
Physical	-1.020	-1.133	-0.943	-0.048**††	0.008
	(0.796)	(0.914)	(1.074)	(0.022)	(0.025)
Technology	-0.752	-0.497	0.701	-0.011	0.028
	(0.942)	(0.943)	(1.595)	(0.032)	(0.026)
All 3	-1.947**††	-1.771**††	-3.071**†††	-0.075***†††	0.011
	(0.825)	(0.784)	(1.267)	(0.026)	(0.023)
AARP	-2.730***†††	-2.195***†††	-2.651*††	-0.101***†††	-0.000
	(0.877)	(0.719)	(1.496)	(0.028)	(0.026)
N	228	228	228	228	237
CSS	-1.502***†††	-1.312***†††	-2.619***†††	-0.057***†††	0.003
	(0.449)	(0.462)	(0.682)	(0.014)	(0.014)
N	228	228	228	228	237
Likert Score (Perceived Age Bias)	-1.407***†††	-1.108***†††	-2.025**†††	-0.057***†††	-0.007
	(0.473)	(0.410)	(0.792)	(0.018)	(0.012)
N	228	228	228	228	237

Note: In the second panel, the treatment is the cosine similarity score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. In the third panel, the treatment is average MTURK Likert score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. The sign is switched from Figure 5 so that a higher value implies job-ad language perceived as more biased against older workers. Bold horizontal lines distinguish separate regressions. All specifications include fixed effects for both city and occupation. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test.

Table B3: Estimated Effects on Age Composition of Applicants, Any Stereotype Treatment, Any Treatment and AARP Treatment, Cosine Similarity Score (or Average), or MTURK Likert Scale (or Average) of Treatment (Standardized), and Distinguishing AARP Treatment, All Cities, Interacted with Unemployment Rate (%)

	Average Age	Median Age	75th Percentile	Over 40	No Age Information
Any Treatment	-6.225*††	-7.435**††	-7.766*††	-0.178†	-0.045
	(3.433)	(3.367)	(4.344)	(0.125)	(0.139)
Any Treatment x MSA Unempl. Rate (%)	0.726	0.976*††	0.916	0.015	0.011
	(0.575)	(0.553)	(0.749)	(0.021)	(0.022)
N	208	208	208	208	217
Any Treatment	-5.420†	-6.958**††	-6.513†	-0.162	-0.051
	(3.306)	(3.242)	(4.308)	(0.129)	(0.132)
Any Treatment x MSA Unempl. Rate (%)	0.654	0.941*††	0.792	0.015	0.012
	(0.556)	(0.536)	(0.742)	(0.021)	(0.021)
AARP	-3.392†	-1.895	-5.454†	-0.050	0.035
	(2.065)	(2.066)	(3.777)	(0.077)	(0.155)
Any Treatment x MSA Unempl. Rate (%)	0.213	0.060	0.432	-0.006	-0.008
	(0.382)	(0.373)	(0.686)	(0.014)	(0.026)
N	208	208	208	208	217
CSS	-3.291†	-3.069	-5.154†	-0.089	-0.047
	(2.470)	(2.545)	(3.289)	(0.082)	(0.119)
CSS x MSA Unempl. Rate (%)	0.339	0.319	0.409	-0.001	0.010
	(0.423)	(0.457)	(0.606)	(0.014)	(0.020)
N	208	208	208	208	217
Likert Score (Perceived Age Bias)	-3.143†	-2.302	-4.503†	-0.055	-0.013
	(2.083)	(2.086)	(2.900)	(0.061)	(0.107)
Likert Score x MSA Unempl. Rate (%)	0.225	0.140	0.232	-0.008	0.002
	(0.353)	(0.355)	(0.522)	(0.011)	(0.017)
N	208	208	208	208	217

Note: In the second panel, the *AARP* variable is equivalent to the interaction between *Any Treatment* and *AARP*. In the third panel, the treatment is the cosine similarity score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. In the fourth panel, the treatment is average MTURK Likert score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. The sign is switched from Figure 5 so that a higher value implies job-ad language perceived as more biased against older workers. Main effects of the interactive variable are also included. There are somewhat fewer observations in this table because the unemployment statistics are produced with a lag, and were not available for the latest experimental observations. Bold horizontal lines distinguish separate regressions. All specifications include fixed effects for both city and occupation. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test.

Table B4: Estimated Effects on Age Composition of Applicants, All Treatments, All Cities, Interacted with Unemployment Rate (%)

	Average Age	Median Age	75^{th} Percentile	Over 40	No Age Information
Communication	$-8.282^{**\dagger\dagger}$	$-10.768^{**\dagger\dagger\dagger}$	$-11.360^{**\dagger\dagger}$	$-0.260^{\dagger}$	-0.135
	(3.792)	(4.052)	(5.341)	(0.156)	(0.136)
Communication x MSA Unempl. Rate (%)	$1.161^{*\dagger\dagger}$	$1.636^{**\dagger\dagger}$	$1.763^{*\dagger\dagger}$	$0.037^{\dagger}$	0.027
	(0.633)	(0.679)	(0.910)	(0.027)	(0.022)
Physical	-2.621	-4.230	-2.624	-0.102	-0.058
	(4.484)	(4.368)	(5.891)	(0.173)	(0.165)
Physical x MSA Unempl. Rate (%)	0.178	0.445	0.049	0.003	0.014
	(0.744)	(0.727)	(0.964)	(0.027)	(0.025)
Technology	$-4.950^{\dagger}$	$-6.262^{\dagger}$	-3.903	-0.118	0.080
	(3.779)	(3.723)	(5.702)	(0.143)	(0.153)
Technology x MSA Unempl. Rate (%)	0.617	$0.886^{\dagger}$	0.626	0.013	-0.011
	(0.670)	(0.631)	(1.029)	(0.025)	(0.027)
All 3	$-5.645^{\dagger}$	$-6.261^{*\dagger\dagger}$	$-8.193^{\dagger}$	-0.170	-0.096
	(3.694)	(3.711)	(5.049)	(0.132)	(0.179)
All 3 x MSA Unempl. Rate (%)	0.658	0.779	0.804	0.008	0.020
	(0.630)	(0.647)	(0.920)	(0.022)	(0.030)
AARP	$-8.810^{*\dagger\dagger}$	$-8.852^{**\dagger\dagger}$	$-11.965^{**\dagger\dagger}$	$-0.212^{\dagger}$	-0.016
	(4.453)	(4.349)	(5.803)	(0.136)	(0.219)
AARP x MSA Unempl. Rate (%)	0.863	$0.996^{\dagger}$	1.218	0.008	0.004
	(0.756)	(0.720)	(1.027)	(0.023)	(0.035)
N	208	208	208	208	217

Note: All specifications include fixed effects for both city and occupation. Main effects of the interactive variable are also included. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test.

Table B5: Estimated Effects on Age Composition of Applicants, Any Stereotype Treatment, Any Treatment and AARP Treatment, Cosine Similarity Score (or Average), or MTURK Likert Scale (or Average) of Treatment (Standardized), and Distinguishing AARP Treatment, All Cities, Interacted with Occupation

	Average Age	Median Age	75th Percentile	Over 40	No Age Information
Any Treatment x Administrative	-1.393*††	-0.888†	-0.805	-0.051**†††	0.018
	(0.718)	(0.648)	(1.186)	(0.019)	(0.026)
Any Treatment x Retail	-1.607†	-1.322	-4.534*††	-0.112**††	0.081
	(1.204)	(1.108)	(2.667)	(0.050)	(0.077)
Any Treatment x Security	-5.031**††	-5.795**††	-4.221	-0.121†	-0.009
	(2.489)	(2.590)	(3.493)	(0.089)	(0.075)
N	228	228	228	228	237
Any Treatment x Administrative	-1.113†	-0.771	-0.478	-0.033*††	0.014
	(0.726)	(0.679)	(1.156)	(0.019)	(0.023)
Any Treatment x Retail	-1.097	-0.869	-3.787	-0.103*††	0.069
	(1.389)	(1.248)	(2.953)	(0.055)	(0.077)
Any Treatment x Security	-4.461*††	-5.478**††	-3.243	-0.104	0.009
	(2.429)	(2.574)	(3.506)	(0.092)	(0.072)
AARP x Administrative	-1.443†	-0.584	-1.675	-0.094***†††	0.024
	(0.993)	(0.937)	(1.653)	(0.035)	(0.027)
AARP x Retail	-2.952*††	-2.611*††	-4.315*††	-0.049	0.066
	(1.512)	(1.384)	(2.300)	(0.048)	(0.099)
AARP x Security	-2.630*††	-1.436	-4.501*††	-0.081†	-0.083
	(1.385)	(1.328)	(2.363)	(0.053)	(0.069)
N	228	228	228	228	237
CSS x Administrative	-0.817	-0.493	-1.559**††	-0.035†	-0.002
	(0.629)	(0.487)	(0.751)	(0.024)	(0.022)
CSS x Retail	-1.852†	-1.578	-5.194**††	-0.141***†††	0.028
	(1.333)	(1.280)	(2.315)	(0.034)	(0.065)
CSS x Security	-3.427*††	-3.627*††	-3.832†	-0.108†	0.004
	(1.794)	(1.897)	(2.324)	(0.069)	(0.087)
N	228	228	228	228	237
Likert Score (Perceived Age Bias) x Administrative	-1.186†	-0.462	-1.635†	-0.074***†††	0.016
	(0.748)	(0.688)	(1.116)	(0.021)	(0.027)
Likert Score x Retail	-2.424***†††	-2.057**†††	-4.762***†††	-0.099***†††	0.063
	(0.753)	(0.811)	(1.135)	(0.026)	(0.068)
Likert Score x Security	-2.556**††	-2.096*††	-3.857**††	-0.083**††	-0.052
	(1.213)	(1.200)	(1.698)	(0.037)	(0.050)
N	228	228	228	228	237

Note: In the second panel, the *AARP* variable is equivalent to the interaction between *Any Treatment* and *AARP*. In the third panel, the treatment is the cosine similarity score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. In the fourth panel, the treatment is average MTURK Likert score for the arm (averaged when there were multiple treatments) with corresponding stereotype or stereotypes; the score (or average) is standardized so the table reports the estimated effects of a 1 standard deviation in the score. The sign is switched from Figure 5 so that a higher value implies job-ad language perceived as more biased against older workers. Main effects of the interactive variable are also included. Bold horizontal lines distinguish separate regressions. All specifications include fixed effects for both city and occupation. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test.

Table B6: Estimated Effects on Age Composition of Applicants, Any Stereotype Treatment, All Cities, Heterogeneity

	Average Age	Median Age	75th Percentile	Over 40	No Age Information
Previous Occupational Experience					
Any Treatment	-8.172*††	-9.333**††	-6.128	-0.198	-0.035
	(4.351)	(4.349)	(5.440)	(0.167)	(0.126)
Any Treatment x Any Previous Occ. Experience	8.338†	10.112*††	4.524	0.157	0.075
	(5.618)	(5.616)	(6.918)	(0.224)	(0.163)
Proportion of those *with* Previous Occ. Experience Aged > 40 = 0.184					
Proportion of those *without* Previous Occ. Experience Aged > 40 = 0.107					
N	228	228	228	228	237
Education					
Any Treatment	-2.953**†††	-2.878**††	-2.684†	-0.025	0.017
	(1.165)	(1.193)	(1.996)	(0.042)	(0.066)
Any Treatment x Any 4-yr College	1.386	1.149	-1.331	-0.272*††	0.045
	(2.798)	(2.810)	(5.697)	(0.135)	(0.154)
Proportion of those *with* Any 4-yr College Aged > 40 = 0.167					
Proportion of those *without* Any 4-yr College Aged > 40 = 0.167					
N	228	228	228	228	237
Unemployed					
Any Treatment	-2.862*††	-3.603**††	-4.063†	-0.100*††	0.075
	(1.529)	(1.663)	(2.722)	(0.057)	(0.083)
Any Treatment x Currently Unemployed	0.133	2.513	2.315	0.006	-0.115
	(2.908)	(3.261)	(6.030)	(0.103)	(0.166)
Proportion of those Unemployed Aged > 40 = 0.182					
Proportion of those *not* Unemployed Aged > 40 = 0.156					
N	228	228	228	228	237
Gender					
Any Treatment	-2.644†	-3.254†	-3.187	-0.088	0.028
	(1.980)	(2.073)	(2.918)	(0.072)	(0.106)
Any Treatment x Female	0.841	2.074	0.352	-0.049	0.026
	(2.537)	(2.575)	(4.134)	(0.094)	(0.136)
Any Treatment x Unknown Gender	-1.213	-0.549	4.266	0.284	-0.070
	(8.432)	(8.548)	(11.083)	(0.341)	(0.424)
Proportion of Females Aged > 40 = 0.157					
Proportion of Males Aged > 40 = 0.202					
Proportion of Unknown Gender Aged > 40 = 0.092					
N	228	228	228	228	237

Note: The regression includes all treatment arms and the control arm. In Panels A-C *Any Treatment* is interacted with the indicated proportion; so the sum of the main and interacted effect would equal the implied effect if the latter proportion was equal to 1. A name is assigned a gender if more than 80% of babies born in the US in the SSN database since 1950 were assigned the corresponding sex at birth. All specifications include fixed effects for both city and occupation. Main effects of the interactive variable are also included. Standard errors clustered at the city-occupation level are reported in parentheses. Data are collapsed to the city-occupation-job ad level. ***, **, or * indicates statistically significant at the 1%, 5%, or 10% level in a two-sided test. †††, ††, or † indicates statistically significant at the 1%, 5%, or 10% level in a one-sided test. 31622