ECONOMIC POLICY

MINISTÈRE DE L'ÉCONOMIE, DES FINANCES ET DE LA RELANCE
Liberté
Égalité
Fraternité

Direction générale du Trésor

75th Economic Policy Panel Meeting

7- 8 April 2022

Machine Learning and Perceived Age Stereotypes in Job Ads: Evidence from an Experiment

Ian Burn, Daniel Firoozi, Daniel Ladd & David Neumark

CEPR CESifo SciencesPo.

Machine Learning and Perceived Age Stereotypes in Job Ads: Evidence from an Experiment[*]

Ian Burn
University of Liverpool

Daniel Firoozi
University of California-Irvine

Daniel Ladd
University of California-Irvine

David Neumark
University of California-Irvine

January 2022

Abstract

We explore whether ageist stereotypes in job ads are detectable using machine learning methods measuring the linguistic similarity of job-ad language to ageist stereotypes identified by industrial psychologists. We then conduct an experiment to evaluate whether this language is perceived as biased against older workers. We find that language classified by the machine learning algorithm as closely related to ageist stereotypes is perceived as ageist by experimental subjects. These methods could potentially help enforce anti-discrimination laws by using job ads to predict or identify employers more likely to be engaging in age discrimination.

[*] This research was supported by the Sloan Foundation. Any views expressed are our own and not those of the Sloan Foundation. This experiment was approved by the UCI Institutional Review Board, HS# 2015-2107. We are very grateful for helpful comments from the editors and anonymous reviewers.

Introduction

Lengthening work lives for those able to work longer is an important part of the policy response to population aging. Reducing age discrimination in hiring is critical to achieving this goal, because many seniors transition to part-time or shorter-term "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). There is an extensive body of research testing for employer discrimination against older workers in labor markets, using correspondence studies to test for discrimination in hiring (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.[1] There is little work, however, that studies how workers respond to age discrimination in the labor market.

In this paper, we explore potential worker responses to one potential manifestation of age discrimination in the labor market – in particular, whether workers perceive job requirements using language related to ageist stereotypes as biased against older workers. The use of ageist stereotypes in job ads, if it discourages older workers from applying for jobs, can have the same adverse outcome on the hiring of older workers as employers discriminating against older job applicants.[2]

We conduct an experiment to explore how job-ad language using age-related stereotypes, or more blatantly ageist language, is interpreted by potential older job applicants. To do this, we begin by using machine learning methods (partly developed in Burn et al., forthcoming) to identify phrases in job ads that are linguistically related to ageist stereotypes drawn from the industrial psychology literature.[3] We use these

[1] For example, Neumark et al. (2019a) study artificial applicants aged 29-31 (younger), 49-51 (middle-aged), and 64-66 (older). For women, there is a distinct pattern of the highest callback rates for the younger applicants, lower for the middle-aged applicants, and lowest for the older applicants. Compared to young applicants, the callback rate for old female applicants for administrative jobs was 47 percent lower (7.58 percent vs. 14.41 percent). In sales, the difference was a bit smaller – a 36 percent lower callback rate. For male job applicants in sales, security, and janitor jobs, there was a lower callback rate for older applicants than younger applicants, although the age pattern is not as consistent or pronounced across the three age groups.

[2] Viewed in the context of labor market search, direct discrimination by employers reduces the likelihood that employers make an offer to an older workers, whereas discouraging them from applying reduces the likelihood that older workers find a match. Both thus reduce the arrival rate of job offers, lengthening unemployment durations.

[3] The job ads were collected as part of a large-scale correspondence study of age discrimination (Neumark et al.,

phrases to construct typical job-ad language that reflects specific age stereotypes. We then conduct an MTURK experiment that asks whether respondents perceive this job-ad language – which the machine learning algorithm classified as related to ageist stereotypes – as ageist. Our experimental evidence shows that job-ad sentences that are classified as closely related to ageist stereotypes by the machine learning algorithm are also rated as ageist by respondents in many cases.

Utilizing ageist language in job ads to shape the applicant pool by discouraging older applicants has a potential benefit for discriminatory employers, because of the incentives created by age discrimination laws. 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.

Our evidence does not directly address the actual behavior or intent of employers that might underlie the use of ageist stereotypes in job ads. Using such language to deter older job applicants could reflect taste discrimination or statistical discrimination. That is, employers may intentionally use this language to deter older workers from applying, either because of taste discrimination (a simple aversion to hiring older workers) or statistical discrimination (an assumption that older workers are not as productive), and this may "work" because respondents perceive age stereotypes in job ads as directly reflecting age bias. Alternatively, employers could be stating actual job requirements; they may have no discriminatory intent in doing this, but may engage in statistical discrimination by assuming that older workers who apply for jobs with these requirements are less likely to meet the requirements. Older job searchers, knowing this, might perceive the job-ad language as ageist because job ads with this language are less likely to result in job offers when older workers apply.[4]

However, we believe that this more subtle story of no discriminatory intent is not the operative one,

2019a). The present research, in turn, is being used to develop a field experiment on how actual job applicants respond to ageist language in job ads.

[4] Moreover, this would be a reasonable expectation, given the correspondence-study evidence from prior work that the kinds of age-stereotyped phrases from the job ads that we use help predict age discrimination by employers (Burn et al., forthcoming).

and rather that employer behavior and respondent perceptions pertain to a desire to avoid hiring older workers. First, in the original correspondence study from which the job ads are drawn (Neumark et al., 2019), evidence of statistical discrimination based on age-related worker skills and characteristics was largely ruled out. Second, the treatment phrases we use based on the machine learning (described in more detail below) do not seem to describe skills that are notably different between workers over and under 50 (the dividing line for "older" in our experiment). As examples, the phrases are things like "good communication and teamwork," "accounting software systems like Netsuite…," and "lift 40 pounds." Thus, our sense is that the job-ad language is perceived more as a signal of ageism than as a signal of job requirements that are strongly related to age.

Regardless, in each of these scenarios, our evidence has the potential to identify job-ad language that can deter job applications from older workers. And as discussed in the next section, any of these scenarios could reflect age discrimination as either social scientists or the law define it.

Our paper demonstrates the promise of machine learning methods to help reduce age discrimination in the labor market. We present two types of evidence as "proof of concept." First, we verify that our machine learning methods detect the presence of stereotyped language in our constructed job ads, even when only one sentence in the job ad is highly related to the ageist stereotype. Second, our main evidence, from our experiment, indicates that this age-stereotyped language is viewed as biased against older workers, which we believe indicates that older job seekers would be less likely to apply to job ads using such language.

Two recent papers have focused on how ageist stereotypes impact the job prospects of older workers. Burn et al. (forthcoming) show that ageist language in job ads helps predict age discrimination by employers, and van Borm et al. (2020) show that employers use ageist stereotypes to help them evaluate resumes. In contrast, we focus on the potential responses of workers to job-ad language that reflects ageist stereotypes. By examining the behavior of workers rather than employers, we are the first in this literature to show that language reflecting age-related stereotypes is viewed as ageist by employees, and that machine learning methods developed to identify age-related stereotypes have potential applications for enforcing non-discrimination protections for older workers.

It might seem unsurprising that job ads using ageist or age-stereotyped language are perceived as ageist by respondents. Indeed, one of our treatments uses language suggested by AARP that is sufficiently blatant (e.g., "energetic person") that the results might not appear surprising at all. A real-world example that is similarly blatant 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.[5]

However, we emphasize two points that make the evidence much more interesting and applicable to real-world job ads that do not use this kind of blatant – and exceptional – language. First, we use phrases from actual job ads (approximately 14,000 job ads collected in the age discrimination correspondence study by Neumark et al., 2019), selecting phrases that appear in these ads *and* are semantically related to age stereotypes. As shown later in the paper, these phrases are far more subtle.[6] Second, 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 (Neumark et al., 2019). In other words, our methods can be used to identify actual age-stereotyped language, and the same methods we use in this paper to study perceptions of potential job applicants also classify job-ad language in a manner that helps predict employer discrimination. Thus, we can garner evidence on whether the same kind of job-ad language that is associated with discriminatory employers also might be likely to discourage older workers from applying for jobs by signaling age discrimination.[7]

Conceptual Framework and Implications of the Evidence

Why might employers use stereotyped language in job ads? One hypothesis is that employers who discriminate based on age use stereotyped language to try to shape the applicant pool, to reduce the

[5] See *Kleber v. Carefusion Corp.* (http://www.aarp.org/content/dam/aarp/aarp_foundation/litigation/pdf-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 (Button, 2019).

[6] Admittedly, the language in *Kleber v. Carefusion Corp.* was from a real-world job ad but was not subtle; but this is an extreme exception.

[7] This kind of discouragement could lead to age patterns in application and hiring data that understate age discrimination, including in correspondence studies of age discrimination if discriminatory employers do not discriminate as much against older applicants because ageist language has already reduced the number of older applicants.

likelihood that age discrimination is detected. 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). 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 the ratio of older applicants to younger applicants is lower.[8] Employers could use job-ad language this way regardless of the nature of age discrimination, 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 either case, employers might use ageist language in job ads to deter older workers from applying, but introduce this language via job requirements that are correlated with age, natural to use in job ads, and not so blatant as to make the age discrimination clear.

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. 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.

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" is illegal in the United States. The U.S. Code of Federal Regulations covering the ADEA currently states, "Help wanted notices or advertisements may not contain terms and phrases that

[8] 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).

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).[9]

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.

Indeed the law even goes further, as in some rare cases employers can even use age as an explicit criterion if it is inherently related to a requirement for the job that is related to age but hard to assess independently. This 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

[9] 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).

criterion). (See U.S. Court of Appeals, 7th Circuit, 1974.)[10]

Thus, our research asking whether we can reliably detect age stereotyping in job ads and whether this language is interpreted as disadvantaging older applicants can provide information and tools to parties that enforce age discrimination laws, by helping to identify job-ad language that may predict intent to discriminate on the basis of age in hiring and adverse impact on older job applicants.[11]

Our evidence cannot speak directly to the question of taste vs. statistical discrimination or whether the 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. Rather, in our experiment we ask respondents if they perceive job-ad language pertaining to job requirements as "biased against workers over age 50" (as explained in more detail below). A positive 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 the stated job requirement. Similarly, our evidence does not speak to whether a stated job requirement would be legal. What our evidence *does* address is whether age stereotypes expressed in job ads likely signal to job applicants that older workers are less likely to be hired. Thus, our evidence can reveal the potential for employers to use job-ad language to discriminate against older workers in hiring, and potential adverse impact on older job applicants. Challenges remain in fully understanding the behavior underlying the actual use of such language in real job ads, and the legality of doing so.

Nonetheless, evidence on age stereotypes in job language could help identify employers that may be discriminating based on age, providing an additional tool in identifying potential discriminators, above and

[10] 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.

[11] Secondarily, our paper makes a contribution regarding methods used in the literature in industrial psychology on employer and worker beliefs about stereotypes. Much of the previous literature in industrial psychology utilizes surveys to understand how employers view older workers or how workers view job ads. To incentivize the elicitation of respondents' true beliefs, we asked respondents to guess how the average respondent to our survey rated each statement, and respondents were paid based on how close to the true value their answers were. When asked to state their own beliefs, respondents in our survey were less likely to rate a statement as ageist. But when asked how they thought the average respondent would view the same statement, they rated statements as more ageist on average. These findings suggest that standard surveying methods that do not incentivize responses may lead to an underreporting of perceptions of ageism; one potential explanation is that it is not socially desirable to perceive the language we use as ageist (Cherry et al., 2015), since doing so indicates that one attributes the characteristics of people used in this language as applying more to older people.

beyond current enforcement mechanisms that rely on worker-initiated complaints.[12] For example, if ageist stereotypes in job ads discourage older applicants from applying for jobs, the EEOC can respond in two ways. First, it could use the text of job ads as a potential indicator of age discrimination in hiring that could be investigated further. Second, the EEOC might issue guidance to employers to avoid language that might discourage older workers from applying.[13] Moreover, the methods explored in this paper have applications that extend to a wide range of situations where text may reflect bias against marginalized groups, such as performance reviews and letters of recommendation.

Studying Job Ads

Very few studies 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 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. Focusing on discriminatory language, Kuhn and Shen (2013) and Kuhn et al. (2018) explore how gender preferences feature explicitly or implicitly in job ads in China, Hellester et al. (2020) explore age and gender preferences in job ads in China and Mexico, and Arceo-Gomez and Campos-Vazquez (2019) study gender and attractiveness preferences in job ads in Mexico.

Two studies, to date, connect the text of job ads to measured discriminatory behavior of employers.[14] 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. And, in the most systematic approach, Burn et al. (forthcoming) identify common age stereotypes from the research literature in industrial psychology, use machine learning to calculate the relationship between

[12] See https://www.eeoc.gov/how-file-charge-employment-discrimination.
[13] If ageist language did less to discourage older workers from applying to discriminatory firms, the ability of the EEOC to identify potential discriminatory behavior from hires relative to applications would be increased.
[14] Though they did not focus on job ads, Hanson et al. (2011) and Hanson et al. (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.

the text of the job ads and specific age stereotypes, and 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 work.[15]

There has been no research on whether the ageist language in job ads is perceived as ageist by potential job applicants. Obviously, the use of such language in job ads is much more troublesome if it is perceived as ageist and thus discourages older workers from applying for jobs. If this happens, it 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, mainly using correspondence studies.[16]

What is known about how job applicants read job ads focuses exclusively on gender bias. Gaucher et al. (2011) found that job ads for male-dominated occupations used words associated with male stereotypes (such as "leader," "competitive," or "dominant") more frequently than advertisements for female-dominated occupations, and women find job advertisements less appealing when they contained more masculine than feminine wording (Bem and Bem, 1973; Gaucher et al., 2011). Chaturvedi et al. (2021) use machine learning to study job ads, identifying words that are predictive of a gender preference; they find that wage offers are lower in jobs with language expressing a preference for women, whether directly, or implicitly through gendered language related to skills, personality traits, and flexible work.

Methods

Selecting the Stereotypes

To select the stereotypes we study, we start with a list of 17 ageist stereotypes from the industrial psychology literature, identified in earlier research (Burn et al., forthcoming). We conducted a detailed review of the industrial psychology, communications, and related literature to identify age stereotypes that this research identifies as applying to workers in their 50s and 60s. We relied on studies that were more likely to cover more recent older cohorts, since age stereotypes may change over time (Gordon and Arvey,

[15] 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.

[16] 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).

2004); we avoided studies published before the 1980s and studies of non-Western countries. We reviewed an extensive set of literature reviews and meta-analyses to identify the relevant studies, but we draw our stereotypes from papers that tested for stereotypes rather than papers that simply reported or aggregated the evidence on stereotypes from other studies. We compiled lists of the stereotypes that these studies identified as applying to older workers. Since studies often have similar stereotypes but phrase them differently, we grouped very similar stereotypes into aggregate categories in a similar manner to the literature review and meta-analysis papers (e.g., Posthuma and Campion, 2007). To focus the analysis on stereotypes on which research agrees, we included a stereotype in our analysis only if at least two studies confirmed the stereotype. This process led to 17 stereotypes of older workers, listed in Table 1.[17]

For our analysis of job ads, we selected a subset of these stereotypes that met the following criteria. First, the stereotype is commonly expressed in job-ad language about the ideal or preferred candidate's 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, according to the industrial psychology literature, employers hold these stereotypes. Second, we focused on stereotypes for which we had evidence of a correlation between discrimination and the stereotype (from Burn et al., forthcoming) and evidence that the stereotype captures a skill that employers view older workers as less likely to possess (from van Borm et al., 2019). Third, older workers should be aware that employers held the stereotype. As evidence, we drew on various reports put out by AARP; see Brenoff (2019) and Terrell (2019). 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.[18]

Industrial psychology research focuses on the skills that employers desire in workers, but in which older workers are perceived deficient. In contrast, job ads rarely use negative formulations of a skill requirement, but instead turn the language to a positive formulation (e.g., ads will ask that a candidate be "adaptable," rather than that they are "not stubborn"). When describing skills and requirements related to our

[17] See Burn et al. (forthcoming) for documentation of the sources used to identify these stereotypes and the larger set of phrases that correspond to them, and additional details about the literature search. Note that a few of these appear as both positive and negative stereotypes about older workers.
[18] Communication skills are one of three stereotypes in Table 1 that appear both positively and negatively. Later, we discuss the implications of this for our evidence.

stereotypes, employers use words like "outgoing" (Stewart and Ryan, 1982), "sociable" (Kite et al., 1991), and "conversational skills" (Ryan et al., 1992; Schmidt and Boland, 1986) to describe communication skills. Physical ability is expressed using words like "energy," "speed," and "physical capability" (Levin, 1988; van Dalen et al., 2009). And technological skills focus on the ability to use "new technology" or on "technological competence" (AARP, 2000; McCann and Keaton, 2013; McGregor and Gray, 2002).

Creating the Treatment (Stereotyped) and Control Job-Ad Language

We create two sets of phrases: treatment phrases and control phrases. The treatment and control job-ad sentences differ in the job requirements expressed and the type of language used in these phrases; we try as much as possible to have the treatment and control phrases describe similar skills, although we had to allow for some differences to be appropriate to the occupation (see Table 2).[19] Our control sentences express job requirements that are also appropriate for the job but use age-neutral language not related to these age-stereotyped skills or abilities, while our treatment sentences use language highly related to these ageist stereotypes.

Our main method for generating phrases and sentences highly related to ageist stereotypes uses measures of semantic similarity generated by machine learning methods. Moreover, to isolate the effects of the different stereotypes, we used the results from these machine learning methods to construct sentences that were highly related to only one of the three stereotypes we use. We calculate the semantic similarity of nearly one million (997,562) phrases from the approximately 14,000 job ads collected in Neumark et al. (2019) to the communication skills, physical ability, and technology skills stereotypes, measuring semantic similarity by the "cosine similarity score," a metric that ranges from – 1 (completely unrelated) to 1 (identical).

We provide a brief overview, explanation, and example of these machine learning methods and the cosine similarity score.[20] We first identify a corpus of the English language that we will use to measure how

[19] For example, for the machine learning treatment for communication skills, the phrase for administrative assistants is "You must have good communication skills and teamwork on tasks," while for retail sales the phrase is "You must have good communication skills with customers and staff." In contrast, the control phrase – "You must be good at working without supervision" – is the same.

[20] See Burn et al. (forthcoming) for a thorough discussion.

similar words and phrases are to each other. We use as the "corpus" the entirety of English-language Wikipedia, which contains all words in the English language. With this corpus, the "input data" are the sentences and paragraphs of Wikipedia, and we compute the semantic similarity between any two words based on the frequency with which they appear together in either sentences or paragraphs, a common procedure in computational linguistics. This procedure results in a vector space (we use 200 vectors), with each phrase (we use three-word phrases, or "trigrams") being represented by weights on each one of these vectors. The vector weights are chosen by a typical machine learning algorithm that iteratively selects these vector weights so as to accurately predict which phrases are near each other (in the same sentence or paragraph) in Wikipedia.

We use these vector weights to compute the similarity between all three-word phrases in our ads and the ageist stereotypes drawn from the industrial psychology. Using the vector weights for these computed from the Wikipedia corpus (which can always be created by breaking down phrases into words), we calculate the cosine similarity (CS) score between these trigrams from the job ads and each stereotype, defined as:

$$[1] \qquad \mathrm{CS}(\mathrm{trigram}, \mathrm{stereotype}) = \frac{\text{dot product}(\mathrm{trigram}, \mathrm{stereotype})}{\|\mathrm{trigram}\| \, \|\mathrm{stereotype}\|}$$

where "trigram" and "stereotype" in the equation refer to the vectors of weights.[21]

The CS score varies between −1 and 1. A score of −1 means the words never appear in the same sentences or paragraphs in Wikipedia. As the CS score increases, the usage of the words becomes more similar; that is, they are used more often in the same sentences or paragraphs, suggesting that they are often used to discuss the same topic. This is what the literature defines as greater semantic similarity. If the words coincide perfectly, the CS score equals 1.

As an example, Figure 1A shows the distribution of cosine similarity scores of all three-word phrases (trigrams) with a particular stereotype; the distribution is centered above zero, which makes sense since we are looking at the text from job ads. To provide some examples, referring to the panel for communications skills, trigrams at the lower end of the distribution are highly unrelated (such as "Christmas season near,"

[21] The || notation indicates the Euclidean norm – e.g., $\|[x, y]^T\| = (x^2 + y^2)^{1/2}$.

with a score of around −0.3), and trigrams at the higher end are more closely related (such as "interactions excellent phone," with a score of around 0.5). The examples provided in the other panels – for physical ability and technology – similarly show low CS scores for unrelated phrases and high CS scores for related phrases.

We use the list of words and phrases from our job ads to 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 sentences, whereas the cosine similarity scores for the other stereotypes listed in Table 1 (including the other two treatment stereotypes) were similar for the treatment and control sentences. 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 then we 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 words and phrases that were highly related to our stereotypes to minimize the semantic similarity. The resulting sentences, for the treatment and control groups, are listed in columns (3) and (4) of Table 2, along with their cosine similarity scores with the stereotype.

Figure 1B repeats the histograms of CS scores from Figure 1A, but now overlaying the positions of these control and treatment phrases. The figure shows that the treatment phrases are a good deal higher in the distributions than the control phrases, and especially for communication skills and physical ability are in the upper tails. Figure 1B and Table 2 illustrate that our experimental results based on the treatment phrases derived from our machine learning will not be "contrived" results that pertain to unusual job-ad language designed to evoke the responses we find. Rather, we are testing whether somewhat subtle shifts in language, which echo actual and reasonable job-ad language variation, affect perceived age bias that could affect job application behavior.

We also use a second stereotype treatment that conveys bias by using ageist language identified by the AARP as related to communication skills, physical ability, and technology skills. We select three AARP examples that best 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 sentence "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 thought it useful to use these rather blatant examples of stereotyped language suggested by AARP as a way of validating our methods and potentially learning more about perceptions of job-ad language related to age stereotypes. One might think that if these phrases are not identified as ageist, it is less likely that our more subtle sentences would be. Conversely, at the other extreme, the experimental responses to the AARP phrases might give a sense of the upper bound of perceived stereotyping we could expect. With reference to the earlier discussion of whether the job-ad language reflects outright age discrimination/bias or stereotyping, one might regard the AARP language as more clearly reflecting discrimination/bias.

On the other hand, note that – as shown in Table 2 – in every case, the cosine similarity score with the stereotype for the machine learning phrase is higher than for the AARP language. This does not necessarily mean that the AARP language is less ageist; rather, the AARP language may be perceived as more ageist, while the language is less directly aligned with a particular ageist stereotype. For example, the AARP language we use for the communications stereotype includes "up-to-date," "jargon," and "dynamic," all of which may reflect ageist stereotypes that are not strongly related to communications. This is not surprising since the AARP stereotypes were not chosen by machine learning with the goal of high semantic similarity with a specific stereotype and not with others. Indeed, as we noted, the AARP language used for communications is in fact described as "cultural fit," which could be much broader.

It is also true that the treatment phrases for some stereotypes are stronger than for others. In particular, the cosine similarity score is always lowest for the phrase corresponding to technological skills

than for the phrases corresponding to the other two stereotypes. As it turns out, however, the same is true for the control phrases. Thus, the impact of the difference between treatment and control phrases may not be expected to differ as much across stereotypes (if, in fact, they are perceived as ageist).

Validating the Treatment vs. Control Differences

While the AARP language is quite blatant and should be perceived as ageist by workers, a first question is how well the stereotyped vs. neutral sentences generated from the machine learning results leads to ads that convey the intended stereotypes. In the language of epidemiology, we would like our treatment sentences to have high "sensitivity" (conveying ageist stereotypes) and "specificity" (conveying information about the specific ageist stereotype intended).

To test whether our phrases are powerful enough to be detected in a job ad, we embedded the treatment and control sentences in job-ad templates we created to correspond to actual job ads. In particular, we created 12 templates per occupation using actual ads collected in Neumark et al. (2019) as a guide to creating our experimental job ads. We supplemented the sample of ads with recent real ads posted on the same job boards used in that study to capture contemporaneous patterns of behavior. Figure 2 provides a few examples, and online Appendix A provides the full set of job-ad templates.

The treatment and control ads differ in the job requirements (denoted in bold with asterisks in each template), with three sentences assigned to be either a treatment phrase (stereotyped) or a control phrase (not stereotyped). As explained above, 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, while our treatment phrases use language highly related to these ageist stereotypes.

Figures 3-5 illustrate how the semantic similarity differs across the templates for the treatment and control job ads and show that our treatment job ads do activate the intended stereotypes. Information on the distribution of all phrases found in the actual approximately 14,000 collected job ads is shown in grey, information for the ads with the treatment job-ad language is shown with dashed black lines, and information for the ads with the control neutral job-ad language with solid black lines. The figures show the median to

99th percentile range and the average (with plotting symbols).[22] We show these for the three stereotypes we study, and then averaged across the other stereotypes.[23]

These figures display a few key results. First, the biased (treatment) job ads have considerably higher 99th percentiles than the control job ads, as well as higher means (and medians, although less so). For example, this is apparent in Figure 3, looking at the bars and symbols for the Physically Able stereotype in the upper left-hand panel, the bars and symbols for the Technology stereotype in the upper right-hand panel, and the bars and symbols for the Communication stereotype in the lower left-hand panel. (In these three panels, we manipulate *only* the indicated stereotype, using the neutral language for the other two.) On the other hand, for the remaining stereotypes – the ones we do *not* manipulate – the control/neutral templates, treatment templates, and collected ads generally have similar medians, means, and 99th percentiles; see, for example, the bars for Other in the upper right-hand panel in Figure 3. The implication of the differences in the means and especially the upper tails of the distributions is that the ads we write using the treatment sentences do, in fact, create ads with notably stronger age stereotypes for the "target" stereotype we are trying to convey. But our manipulated treatments do *not* create similarity with the other stereotypes, as shown by the distribution of all other stereotypes in the job ads (besides the three we are trying to study). That is, our treatment ads only generate a shift in similarity for the stereotypes we are seeking to activate, hence isolating those stereotypes in the job ads.

This key result is also apparent from the lower right-hand panel in each figure, where we use the treatment ads in which we manipulate all three stereotypes at once. If one compares the bars and symbols for any of the three manipulated stereotypes in this panel to the corresponding bars and symbols in the first three panels, the results look almost identical. Again, this reinforces the conclusion that machine learning generated semantic similarity scores are powerful enough to pick up the presence of stereotyped language, even when only one or a few sentences in the ad are actually related to the ageist stereotype.

[22] We think very high percentiles are relevant because they are potentially associated with strongly stereotyped phrases in the job ads, which readers are more likely to notice than potentially very subtle language with lower semantic similarity with ageist stereotypes.
[23] We show results for each of the separate stereotypes in Appendix Figures B1-B3.

In addition, the treatment effect is accentuated by using the control ads rather than simply using all the collected ads (as we would expect). The figures show that the control ads are more neutral than the full set of collected ads, as evidenced by the much lower values of the 99th percentiles for the control ads than for the collected ads, for each of the stereotypes we manipulate, but not for the other stereotypes; for example, compare the bars for Physically Able and for Other in the upper left-hand panel of Figure 3.

Experimental Evidence

Our final step, and the key new contribution of this paper, was to conduct an experiment using Amazon MTURK to measure whether and to what extent job-ad language using phrases with high cosine similarity scores with age-related stereotypes – especially those phrases identified by machine learning methods – are perceived as ageist by potential job applicants, including older applicants.

We recruited participants through the Amazon MTURK online platform. We restricted the sample to U.S. residents. To guarantee that the median age of the sample was roughly 50, we used age-based quotas with a third of the sample in each of the following age bins: 25 to 35, 45 to 55, and 55+. Because the age bins are pre-set by Amazon MTURK, and MTURK's age data may not be up-to-date, we ask participants to self-report their age. The bins we use to collect self-reported age (25 to 35, 35 to 50, and 50+) broaden the MTURK bins to cover a gap in the age distribution and adjust the highest cutoff to 50, in line with our benchmark age for older workers in the survey. We did not balance the recruitment sample on race or gender.[24]

Our experiment consisted of two parts, the baseline survey and the incentivized survey, which were

[24] Respondents to our surveys who met the sample restrictions were excluded if they failed a manipulation check as the first step of the surveys. This manipulation check acted as both a Turing test (ensuring our respondents were human) and a check for American English language fluency (to reduce the chance someone has masked their location). All questions were free response to prevent individuals or computer programs from clicking through and getting the right answers by accident. The first three questions test for English understanding by asking respondents to complete the analogy (e.g., Canine is to Dog as Feline is to ___, for which cat, kitten, Cat, or cats would all be correct). One of these questions (Pen is to Whiteout as Pencil is to ___) was designed to elicit different responses between American and U.K. English speakers. If participants listed "eraser," the correct answer in American English, they were allowed to proceed, while participants who responded with "rubber," the answer in U.K. English, were excluded. The last question was a free response that asked respondents to write two sentences of at least 140 characters telling us who their favorite band or artist is and why. These were checked after the fact to ensure the participant was paying attention and was capable of writing in coherent English sentences. The manipulation checks overall screened out nine respondents. See Appendix Figure C1 for the manipulation checks.

conducted using Qualtrics. In the baseline survey, we recruited 50 respondents who met our criteria and passed the manipulation checks. The responses from the baseline survey were used solely to provide "correct" answers for the next step when we incentivize our second group of respondents to predict the responses of this first group. For the incentivized survey, we recruited 151 respondents who met our criteria and passed the manipulation checks.[25]

In Table 3, we report the self-reported demographic composition of our sampled MTURK respondents (for the incentivized survey). The sample is relatively more-educated, white, and female than the U.S. population as a whole. Consistent with the age-based quotas we set for recruiting participants, the sample is also relatively older, which may explain some of the differences in the other demographic characteristics. In line with our target, we generated a sample with a median age near 50, with roughly 55% of participants over that threshold.

Baseline Survey

The baseline survey had three separate blocks of questions. In the first block of questions, subjects were asked to give their informed consent to participate in the survey (Appendix Figure C2). In the second block, participants were shown a series of job requirements and told they were from job ads posted online. For each requirement, respondents were asked whether they personally agreed or disagreed with the statement that "[Treatment or control requirement] is biased against workers over 50."[26]

Within each of the blocks that asked about whether or not phrases were perceived as biased, there were three pages for each respective stereotype: communication, physical, and technology. All the treatment and control phrases were tested on their respective pages, but the pages were not labeled by stereotypes that respondents viewed. Respondents had to proceed sequentially through the pages and could not go back or skip between them. The ordering of the questions was randomized within each page. We adopted this approach to force respondents to compare treatment and control phrases in a specific stereotype category

[25] We recruited 150 respondents initially – 50 in each age bin – but one person completed the survey and then refused to accept payment. Thus, we ended up with 151 subjects because the quota filling on MTURK only counts people who accepted payment.

[26] An example is shown in Online Appendix Figure D1. In principle, an experiment like ours could also ask respondents questions about their views of true correlations of age (or other groups studied) with the worker characteristics captured in the treatment phrases, as opposed to whether the phrases just signal bias.

without prompting them about the specific age-related stereotype in question. The order of questions on a page was randomized for each respondent. Responses to the questions were given on a Likert scale with the options: Strongly agree, Somewhat agree, Neither agree nor disagree, Somewhat disagree, Strongly disagree.

In the third block, we concluded by collecting the demographic characteristics of our sample (Appendix Figure C3). They were asked to report their age, gender, education, and race/ethnicity.

Incentivized Survey

The second survey was split into five blocks of questions. The first asked the participants to give their informed consent. The second block of questions repeated questions from the baseline survey and asked respondents about their own opinions on job requirements.

The next two parts formed the crux of the incentivized survey. In each part, the respondents were asked to guess how the participants in the baseline survey rated the job requirements, and they were rewarded for how correctly they guessed. Before starting each of these blocks of questions, respondents were sent to a landing page that emphasized the new prompt and cash incentive.

The third set of questions asked respondents to predict what the previous survey respondents answered when they were shown job requirements from the job ads.[27] For each requirement, they were asked whether they thought previous respondents agreed or disagreed with the statement that "This job requirement is biased against workers over 50." They were shown the same Likert scale shown to the baseline respondents. Respondents were informed that either the third set of questions in the survey or the fourth (described below) would be randomly selected for a cash incentive based on their answers to the questions in that part. They were told they would earn bonus pay, which was to be calculated based on how close they were to the correct answer. If they correctly predicted what the average participant said, they would earn \$0.32 per question.[28] Incorrect answers received less money, with the penalty increasing the further they were from the correct answer. Payouts were calculated using the quadratic scoring rule

$$[2] \qquad P_{iq} = 0.32 - 0.02 \times (\overline{A_q} - A_{iq})^2 ,$$

[27] See Appendix Figure D2.
[28] Average pay was \$7.22.

where P_{iq} was the payout to individual i for question q based on the average response to question q by previous respondents ($\overline{A_q}$) and their answer about how previous respondents answered question q (A_{iq}).[29] All the treatment and control phrases were tested on their respective pages, but the pages were not labeled by stereotypes that respondents viewed. Respondents had to proceed sequentially through pages and could not go back or skip between them. The question order was randomized within each page.

The fourth block of questions differed from the third in that we asked respondents to guess how the older participants in the baseline survey (those over 50 years old) responded.[30] Before starting this block of questions, respondents were sent to a landing page that emphasized both the scoring rule and the age of respondents for whom they were guessing. The instructions read, "For each requirement, please state whether you think previous respondents over the age of 50 agree or disagree with the statement that "This job requirement is biased against workers over 50." The structure of this section was identical to the third block of questions.

Analyses

To test for differences in how the treatment and control phrases were perceived, we employ a series of regression models. We begin with a simple regression testing for differences in respondent beliefs between treatment and control phrases conditional on observable characteristics:

$$[3] \qquad A_{iq} = \alpha + \beta T_{iq} + X_i\delta + \varepsilon_{iq}$$

The ranking that individual i gave to question q (A_{iq}) is the dependent variable. We include controls for gender, level of education, race, and age (X_i). If respondents view the treatment phrases as more biased against older workers than the control phrases, we should find that β is less than zero, because the responses (A_{iq}) range from 1 for strongly agree to 5 for strongly disagree. If we observe β is positive, then this is evidence that treatment phrases were rated as less biased by respondents. In Equation [3], and all subsequent regressions, the standard errors are clustered at the respondent level.

[29] Respondents were told that their earnings, "would be calculated according to the formula: M=\$0.32 − \$0.02 *(Average Previous Answer − Your Prediction)^2." To illustrate their payoff, respondents were told they would earn \$0.24 if the correct answer was "somewhat agree" and they guessed "somewhat disagree."
[30] See Appendix Figure D3.

Our next step is to explore two forms of heterogeneity in our estimated treatment effects.[31] The first examines whether respondents view treatment phrases differently depending on the stereotype to which the requirement is related. To test this, we define dummy variables for each of our three pairs of a stereotype treatment and the corresponding control (omitting one from the regression), and interactions between these dummy variables and the dummy variables for each stereotype treatment (communication skills, physical ability, or technology skills):

$$[4] \quad A_{iq} = \alpha + \beta_1\left(T_q \times ComSkill_q\right) + \beta_2\left(T_q \times PhysAb_q\right) + \beta_3\left(T_q \times TechSkill_q\right) + \gamma_1\left(PhysAb_q\right) + \gamma_2\left(TechSkill_q\right) + \delta \mathrm{X_i} + \varepsilon_{iq}$$

Thus, the coefficients β_i, i = 1, 2, 3, capture how biased respondents view the treatment phrase for the stereotype relative to the control phrase.

The second form of heterogeneity we examine is the difference between the machine learning derived treatment phrases and the AARP treatment phrases. To do this, we add an additional interaction for when the treatment phrase is the AARP phrase:

$$[5] \quad A_{iq} = \alpha + \beta_1\left(T_q \times ComSkill_q\right) + \beta_2\left(T_q \times PhysAb_q\right) + \beta_3\left(T_q \times TechSkill_q\right) + \theta_1\left(T_q \times ComSkill_q \times AARP_q\right) + \theta_2\left(T_q \times PhysAb_q \times AARP_q\right) + \theta_3\left(T_q \times TechSkill_q \times AARP_q\right) + \gamma_1\left(PhysAb_q\right) + \gamma_2\left(TechSkill_q\right) + \delta \mathrm{X_i} + \varepsilon_{iq}$$

The coefficients θ_i , i = 1, 2, 3, identify how much more biased respondents view the AARP treatment phrases relative to the machine learning derived treatment phrases for the same stereotype. More importantly, perhaps, only with additional interactions added do we get separate estimates of the treatment effects that exclude the AARP language – i.e., the phrases based on machine-learning from the job ads. Their effects are identified from the estimates of β_i, i = 1, 2, 3, in Equation [5]. This is potentially important because the AARP phrases may be quite distant from what would be viewed as normal or acceptable job-ad language.

[31] We also examined heterogeneous differences in the treatment phrases across demographic groups. We found little evidence to support the hypothesis that different groups view job requirements differently. The pattern of results observed in Table 4 held by sex, age, education, or race. These results are available upon request.

Results

Figure 6 provides 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. Respondents rated the physical and technology-biased phrases derived from our cosine similarity score index as more biased than the control phrases, but viewed the communication skills stereotyped phrases as roughly identical to the controls. Opinions 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 stills 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.)

In Table 4, we estimate regression models for the survey responses that delve into more detail. In all cases, standard errors are clustered at the respondent level. In column (1), we report the estimated coefficient from a simple model of the responses for self-beliefs on a dummy variable for whether the response is to any treatment phrase. The estimated coefficient of −0.886 implies that responses are lower by almost one category of the Likert scale. Recall that the responses range from 1 for strongly agree to 5 for strongly disagree, so a negative estimate implies the phrase was perceived as more ageist. The estimate is strongly statistically significant. To help interpret the magnitude, the third number (in square brackets) reports the

implied effect in terms of standard deviations of the responses.[32]

Column (2) expands the specification to differentiate the treatment by the type of stereotype – communications, physical ability, or technology skills – without differentiating the machine learning phrases from the AARP phrases. We find significant negative effects (implying more ageist phrases) for all three, with the largest estimate (whether looking at the coefficient or the standard deviation effect) for physical ability, followed by technology skills, and the smallest estimate for communications. In this model, we also include controls for the different stereotype phrases (whether treatment or control) so that the Treatment × stereotype interactions measure the differences relative to the paired control phrase.

Column (3) expands the specification to differentiate between machine learning and AARP treatment phrases. In this column, the estimated coefficients of the Treatment × stereotype coefficients measure the effects of the machine learning phrases, and the Treatment × stereotype × AARP estimates measure the differential effects of the AARP treatments relative to the machine learning treatments. We see that in every case the estimated effects of the AARP treatments are larger, in the direction of more perceived bias. All of the estimated differences for the AARP phrases are statistically significant, and the magnitudes are considerably larger for the communications and technology skills stereotypes. For the machine learning stereotypes, the estimated treatment effects for physical ability and technology skills are sizable, significant, and negative, while the effect for communications is near zero and insignificant (paralleling what we found in the raw data). On the one hand, this suggests that the phrases generated by machine learning for communications stereotypes do not evoke ageism as strongly, while the phrases in the AARP treatment do; on the other hand, recall the earlier caution that the stronger evidence for the AARP treatment for communications may not isolate the communications stereotype well.

Columns (4)-(6) report estimates of the same specifications, but for the beliefs about other respondents – i.e., how respondents think others would perceive the language. The qualitative pattern of estimates is the same, but the estimated impacts of the treatments are generally larger. This can be seen most

[32] We estimated the specifications in Table 4 using an ordered probit model as well, to account for the fact that our dependent variable is actually ordinal, rather than cardinal. This led to qualitatively very similar results, so we report the OLS results for simplicity. Results available upon request.

simply by comparing the estimated treatment effects between columns (5) and (2). The estimated coefficients are substantially larger – especially for the physical ability and technology skills stereotypes – and in all three cases, the differences are strongly statistically significant (as indicated by the "daggers"). Comparing columns (6) and (3) indicates that the differences between self-beliefs and perceived beliefs of others for the physical ability and technology stereotypes are driven by the machine learning phrases, as their estimated coefficients are substantially larger in column (6) than in column (3), whereas the estimated interactions with the AARP phrases are not uniformly larger or smaller. As noted earlier, the differences in responses for perceived beliefs of others and self-beliefs may reflect the incentives we offered in eliciting the former, which could counter social desirability biases.

Finally, columns (7)-(9) focus on the perceived beliefs of those over age 50. These estimates are quite similar to those in columns (3)-(6), suggesting that respondents did not particularly believe that older individuals were more likely to perceive language as ageist.[33] Of course, given that the stereotyped language *was* perceived as ageist, the impact on behavior would likely be stronger the older is the person reading job ads with these phrases.

Our last analysis compares the perceptions (self-beliefs, or of others) to the cosine similarity scores for the phrases we use (see Table 2), providing a useful graphical depiction of our survey results that captures many of our key points. Figure 7, Panel A does this for self-beliefs. Note the vertical axis is decreasing in the reported belief because a lower number implies stronger perceived ageism. Consider first the points plotted for control phrases. Referring back to Table 2, there are two control phrases for physical ability and two for communications, so there are two circles for each of these. But there are three circles for technology skills, for which there are three control phrases. The horizontal axis measures the cosine similarity score for these, and they are clustered towards zero. The vertical height measures the perceived bias of these phrases, and – by design – they are low.[34] The triangles are for the machine learning phrases. As Table 2 showed, these

[33] As further evidence that these perceptions that younger and older people to not perceive the ageist phrases differently, the estimated effects of the treatments on self-beliefs of respondents aged 50 or under and over age 50 were very similar. Results available upon request.

[34] We do this analysis for the mean survey responses, rather than the regression coefficients, because we want to depict these for each occupation and the regressions do not estimate separate effects by occupation.

generally have the highest cosine similarity scores with the corresponding stereotypes. The squares are for the AARP stereotypes, which generally have lower cosine similarity scores; there are only three of these plotted because there are only three phrases. Comparing the height of the triangles to the squares, we see that the AARP phrases are generally perceived as more biased, even though the cosine similarity scores with the stereotypes are lower.

Panels B and C in Figure 7 show the same kind of evidence of beliefs about others' perceptions in general, and then for others over age 50. The qualitative patterns are the same, but Panels B and C, in comparison to Panel A, show the stronger perceived bias reported when asked about others' perceptions. This is apparent in Panels B and C, for both the machine learning phrases (triangles) and the AARP phrases (squares); there is no apparent shift for the control phrases. This has an interesting potential implication. If an older job applicant thinks others will perceive job-ad language as more age biased than the individual herself perceives the language as age biased, they might expect less competition from older applicants, which might boost the likelihood the person applies for a job relative to the case where her perceptions were the same as those she ascribes to others. This does not mean the age-stereotyped language will not deter the older applicant from applying for a job, but it does mean that the greater perceived age bias by others may mitigate the effect.[35]

Assessing Actual Job Ads

To help the reader contextualize our results, we now connect the results to real job ads in a more concrete way. In particular, in Figure 1C, we return to the distributions of CS scores from the job ads, but we overlay, for each of the three stereotypes, the average treatment and control CS scores (for the machine learning treatments), and we also show the estimated effect of the treatments on perceived ageism. The panels in this figure give a sense of how one might, in principle, relate our results to actual job-ad language

[35] Bursztyn et al. (2000) describe a related result in a much different context. In particular, they show that young married men in Saudi Arabia are more supportive of women working outside the home than they think other similar men are. In this case, the authors do an experimental intervention to correct men's beliefs about others' beliefs, and find that doing so – making them aware that other men are more supportive – increases the likelihood that their wives apply for jobs outside the home, suggesting that agents can be responsive to perceptions about others' beliefs. Of course in our case, correcting the apparent misperception of how others perceive job ads could have an adverse effect, because in our context the misperception might encourage older job seekers to apply for jobs.

in a set of job ads, by showing how our experimental manipulation and the effects of that manipulation are related to a large set of actual job ads. For example, language with CS scores near or above the levels of our experimental treatments could be flagged as potentially indicating discriminatory behavior; and as discussed earlier, our treatment relates to actual and reasonable job-ad language, not extreme phrases that are blatant (and perhaps very unlikely to be used). To be clear, though, we would not advocate for this being viewed as definitive evidence of discrimination, but rather – at most – as a potential indicator for further investigation into actual hiring behavior.

Conclusions and Discussion

In this paper, we explore whether the type of ageist stereotypes used in job ads is detectable using machine learning methods and whether this language is perceived as biased against older workers. This is important for three reasons. First, workers may respond to this language, with older workers applying to a narrower set of jobs or perhaps choosing not to apply at all, hence diminishing their job market opportunities. Second, the mechanism we study is plausible, as employers who want to discriminate against older workers but also want to avoid getting caught might manipulate job-ad language to discourage older workers from entering the applicant pool. And third, if ageist job-ad language can be detected by machine learning methods, then these methods could, in principle, be used to help enforce anti-discrimination laws by helping to predict or identify employers more likely to be engaging in age discrimination.

We use machine learning methods to identify phrases in job ads that are linguistically related to standard ageist stereotypes from the industrial psychology literature. We use these phrases to construct typical job-ad language that reflects specific age stereotypes. We show that machine learning methods are sensitive enough to detect the presence of stereotyped language, even when only one sentence in the job ad is highly related to the ageist stereotype. We then conduct an MTURK experiment that asks whether respondents perceive this job-ad language – which the machine learning algorithm classifies as related to ageist stereotypes – as ageist. We also used some more blatant ageist phrases identified by AARP.

Our experimental evidence shows that sentences that are classified as closely related to ageist stereotypes by the machine learning algorithm are generally perceived as ageist by respondents in our

MTURK experiment (and more so when asked, with incentives, how they will be perceived by others). These results imply that the different age stereotypes we study capture real ageist sentiments and will be perceived as such by job applicants.

Although the AARP phrases were perceived as more ageist than those generated by our machine learning methods, the latter were more directly and more distinctly related to specific ageist stereotypes. This is potentially significant from a policy perspective, as it implies that machine learning can be used to identify ageist stereotypes in job ads the pertain to *specific* stereotypes. Because the legality of an age stereotype in a job ad might hinge on whether the language pertains to a job requirement based on a reasonable factor other than age (RFOA), it is important to be able to ascertain the type of job requirement to which the language might refer. In contrast, the AARP phrases we use, while perceived as more ageist, are harder to tie to specific stereotypes and hence, perhaps, to specific job requirements. This distinction may be useful with regard to the issues of what our evidence implies about underlying behavior and how our evidence might assist in enforcing age discrimination laws. It seems safe to say that job-ad language with the same ageist flavor of the AARP phrases will fairly reliably help identify employers engaging in illegal age discrimination. In contrast, job-ad language identified by machine-learning methods should be interpreted more cautiously, as it could reflect legitimate job requirements, and may not have as adverse an impact on older workers looking for jobs. Nonetheless, the machine-learning methods could be helpful in flagging potentially discriminatory behavior.

References

AARP. (2000). *American business and older employees*. AARP: Washington, DC.

Arceo-Gomez, E.O., & Campos-Vazquez, R.M. (2019). Double discrimination: Is discrimination in job ads accompanied by discrimination in callbacks? *Journal of Economics, Race, and Policy, 2*, 257-68.

Baert, S., Norga, J., Thuy, Y., & Van Hecke, M. (2016). Getting grey hairs in the labour market. An alternative experiment on age discrimination. *Journal of Economic Psychology, 57*, 86-101.

Bem, S.L., & Bem, D.J. (1973). Does sex-biased job advertising "aid and abet" sex discrimination? *Journal of Applied Social Psychology, 3*, 6-18.

Bendick, M., Jr., Brown, L.E., & Wall, K. (1999). No foot in the door: An experimental study of employment discrimination against older workers. *Journal of Aging & Social Policy, 10*, 5-23.

Bendick, M., Jr., Jackson, C.W., & Romero, J.H. (1997). Employment discrimination against older workers: An experimental study of hiring practices. *Journal of Aging & Social Policy, 8*, 25-46.

Brenoff, A. (2019). *5 ageist phrases to be aware of.* AARP: Washington, DC.

Burn, I., Button, P., Munguia Corella, L.F., & Neumark, D. Older workers need not apply? Ageist language in job ads and age discrimination in hiring. Forthcoming in *Journal of Labor Economics.*

Bursztyn, L., González, A.L., & Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *American Economic Review, 110,* 2997-3029.

Button, P. (2019). Population aging, age discrimination, and age discrimination protections at the 50th anniversary of the Age Discrimination in Employment Act. In *Current and emerging trends in aging and work,* Czaja, S.J., Sharit, J., & James, J.B., eds., 163-88. New York, NY: Springer.

Cahill, K.E., Giandrea, M.D., & Quinn, J.F. (2006). Retirement patterns from career employment. *The Gerontologist, 46,* 514-23.

Carlsson, M., & Eriksson, S. (2019). The effect of age and gender on labor demand – evidence from a field experiment. *Labour Economics, 59*, 173-83.

Chaturvedi, S., Mahajan, K., & Siddique, Z. (2021). Words matter: Gender, jobs and applicant behavior. IZA Discussion Paper No. 14497.

Cherry, K.E., Allen, P.D., Denver, J.Y., & Holland, K.R. (2015). Contributions of social desirability to self-reported ageism. *Journal of Applied Gerontology, 34*, 712-33.

Combs, D.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.

Deming, D., & Kahn, L.B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics, 36*, 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, H.S., Silverman, D., & von Wachter, T.M. (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, H.S., Herbst, C.M., Silverman, D., & von Wachter, T.M. (2019). Whom do employers want? The role of recent employment and unemployment status and age. *Journal of Labor Economics, 37*, 323-49.

Federal Register. (n.d.). *Disparte 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).

Gaucher, D., Friesen, J., & Kay, A.C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology, 101*, 109-28.

Gordon, R.A., Arvey, R.D. (2004). Age bias in laboratory and field settings: A meta-analytic investigation. *Journal of Applied Social Psychology, 34*, 468–92.

Hanson, A., Hawley, Z., & Taylor, A. (2011). Subtle discrimination in the rental housing market: Evidence from email correspondence with landlords. *Journal of Housing Economics, 20*, 276-84.

Hanson, A., Hawley, Z., Martin, H., & Liu, B. (2016). Discrimination in mortgage lending: Evidence from a correspondence experiment. *Journal of Urban Economics, 92*, 48-65.

Hellester, M.D., Kuhn, P., & Shen, K. (2020). The age twist in employers' gender requests. *Journal of Human Resources, 55*, 482-69.

Johnson, R.W., Kawachi, J., & Lewis, E.K. (2009). *Older workers on the move: Recareering in later life*. AARP Public Policy Institute. AARP: Washington, DC.

Kite, M.E., Deaux, K., & Miele, M. (1991). Stereotypes of young and old: Does age outweigh gender? *Psychology and Aging, 6*, 19-27.

Kuhn, P., & Shen, K. (2013). Gender discrimination in job ads: Evidence from China. *Quarterly Journal of Economics, 128*, 287-336.

Kuhn, P., Shen, K., & Zhang, S. (2018). Gender-targeted job ads in the recruitment process: Evidence from China. NBER Working Paper No. 25365. Cambridge, MA.

Lahey, J. (2010). International comparisons of age discrimination laws. *Research on Aging, 32*, 679-97.

Lahey, J. (2008). Age, women, and hiring: An experimental study. *Journal of Human Resources, 43*, 30-56.

Levin, W.C. (1988). Age stereotyping: College student evaluations. *Research on Aging, 10*, 134-48.

Maestas, N. (2010). Back to work: Expectations and realizations of work after retirement. *Journal of Human Resources, 45*, 718-48.

Marinescu, I.E., & Wolthoff, R. (2020). Opening the black box of the matching function: The power of words. *Journal of Labor Economics, 38*, 535-68.

McCann, R.M., & Keaton, S.A. (2013). A cross cultural investigation of age stereotypes and communication perceptions of older and younger workers in the USA and Thailand. *Educational Gerontology, 39*, 326-41.

McGregor, J., & Gray, L. (2002). Stereotypes and older workers: The New Zealand experience. *Social Policy Journal of New Zealand, 18*, 163-77.

Modestino, A.S., Shoag, D., & Balance, J. (2016). Downskilling: Changes in employer skill requirements over the business cycle. *Labour Economics, 41*, 333-47.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature, 56*, 799-866.

Neumark, D., Burn, I. & Button, P. (2016). Experimental age discrimination evidence and the Heckman critique. *American Economic Review Papers and Proceedings, 106*, 303-8.

Neumark, D., Burn, I. & Button, P. (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, D., Burn, I. Button, P., & Chehras, N. (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.

Posthuma, R., Campion, M.A. (2007). Age stereotypes in the workplace: Common stereotypes, moderators,

and future research directions. *Journal of Management, 35*, 58–88.

Riach, P.A., & Rich, J. (2006). An experimental investigation of age discrimination in the French labour market. IZA Discussion Paper No. 2522. Bonn, Germany.

Riach, P.A., & Rich, J. (2010). An experimental investigation of age discrimination in the English labor market. *Annals of Economics and Statistics, 99/100*, 169-85.

Ryan, E.B., See, S.K., Meneer, W.B., & Trovato, D. (1992). Age-based perceptions of language performance among younger and older adults. *Communication Research, 19*, 423-43.

Schmidt, D.F., & Boland, S.M. (1986). Structure of perceptions of older adults: Evidence for multiple stereotypes. *Psychology and Aging, 1*, 255-60.

Stewart, M.A., & Ryan, E.B. (1982). Attitudes toward younger and older adult speakers: Effects of varying speech rates. *Journal of Language and Social Psychology, 1*, 91-109.

Terrell, K. (2019). *Age bias that's barred by law appears in thousands of job listings.* AARP: Washington, DC.

Tilcsik, A. (2011). Pride and prejudice: Employment discrimination against openly gay men in the United States. *American Journal of Sociology, 117*, 586-626.

U.S. Court of Appeals, 7th Circuit. 1974). *Hodgson v. Greyhound Lines, Inc.*, 499 F.2d 859 (7th Cir. 1974). https://casetext.com/case/hodgson-v-greyhound-lines-inc (viewed January 7, 2022).

U.S. Equal Employment Opportunity Commission. (1979). *Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures.* Federal Register, Vol. 40, No. 43, March 2. https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines (viewed January 18, 2022).

U.S. Equal Employment Opportunity Commission. (n.d.(a)). *Prohibited Employment Policies/Practices.* http://www1.eeoc.gov//laws/practices/index.cfm?renderforprint=1 (viewed September 15, 2019).

U.S. Equal Employment Opportunity Commission. (n.d.(b)). *Fact Sheet: Age Discrimination*. https://www.eeoc.gov/laws/guidance/fact-sheet-age-discrimination (viewed January 7, 2022).

van Borm, H., Burn, I., & Baert, S. (2019). What does a job candidate's age signal to employers? IZA Discussion Paper No. 12849. Bonn: Germany.

van Dalen, H., Henkens, K., & Schippers, J.J. (2009). Dealing with older workers in Europe: A comparative survey of employers' attitudes and actions. *Journal of European Social Policy, 19*, 47-60.

Wax, S.L. (1948). Discrimination by Summer Resorts in Ontario. *Information and Comment: Committee on Social and Economic Studies of the Canadian Jewish Congress, 7, June*, 10-13.

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. (forthcoming).

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 (CSS = 0.20)	You must have good communication and teamwork on tasks (CSS = 0.48)	You must be up-to-date with current industry jargon and communicate with a dynamic workforce (CSS = 0.23)
Administrative Assistants	Physical ability	You must enter bills and keep track of invoices (CSS = 0.11)	You must be able to lift 40 pounds (CSS = 0.41)	You must be a fit and energetic person (CSS = 0.30)
Administrative Assistants	Technological skills	You must produce and distribute documents such as correspondence memos, faxes and forms (CSS = 0.08)	You must use accounting software systems like Netsuite, Freshbook, and QuickBooks (CSS = 0.29)	You must be a digital native and have a background in social media (CSS = 0.22)
Retail sales	Communication skills	You must be good at working without supervision (CSS = 0.20)	You must have good communication with customers and staff (CSS = 0.34)	You must be up-to-date with current industry jargon and communicate with a dynamic workforce (CSS = 0.23)
Retail sales	Physical ability	You must enter bills and keep track of invoices (CSS = 0.11)	You must be able to lift 40 pounds (CSS = 0.41)	You must be a fit and energetic person (CSS = 0.30)
Retail sales	Technological skills	You must help to clean and organize the store (CSS = 0.09)	You must use software such as Microsoft Office/Excel or Google Sheets (CSS = 0.27)	You must be a digital native and have a background in social media (CSS = 0.22)
Security guard	Communication skills	You must follow instruction from supervisors (CSS = 0.21)	You must maintain communication about tasks with supervisors (CSS = 0.38)	You must be up-to-date with current industry jargon and communicate with a dynamic workforce (CSS = 0.23)
Security guard	Physical ability	You need to carry a flashlight (CSS = 0.20)	You must be able to lift 50 pounds (CSS = 0.41)	You must be a fit and energetic person (CSS = 0.30)
Security guard	Technological skills	You must write patrol records in journal notebook (CSS = 0.03)	You must type patrol entries into a journal application on a computer system (CSS = 0.24)	You must be a digital native and have a background in social media (CSS = 0.22)

Note: See text for a description of how each sentence was created. The average cosine similarity score with the stereotype for each phrase (averaging over the cosine similarity score of each word contained in the phrase) is reported in parentheses. CSS = "cosine similarity score."

Table 3: Demographics of MTURK Sample

Demographic Characteristic	Number of Respondents	Percent of Sample
A. Level of Education		
Postgraduate Degree	25	16.6%
Bachelor's Degree	60	39.7%
Some College or 2 Year Degree	49	32.4%
High School Graduate or Less	17	11.3%
B. Age Group		
21 to 35 years old	45	29.8%
35 to 50 years old	23	15.2%
Over 50 years old	83	55.0%
C. Sex		
Female	83	55%
Male	68	45%
D. Race and Ethnicity		
White	125	82.8%
Black or African American	8	5.3%
Asian	6	4.0%
Hispanic or Latino	4	2.7%
Other	3	2.0%
Two or More	5	3.3%

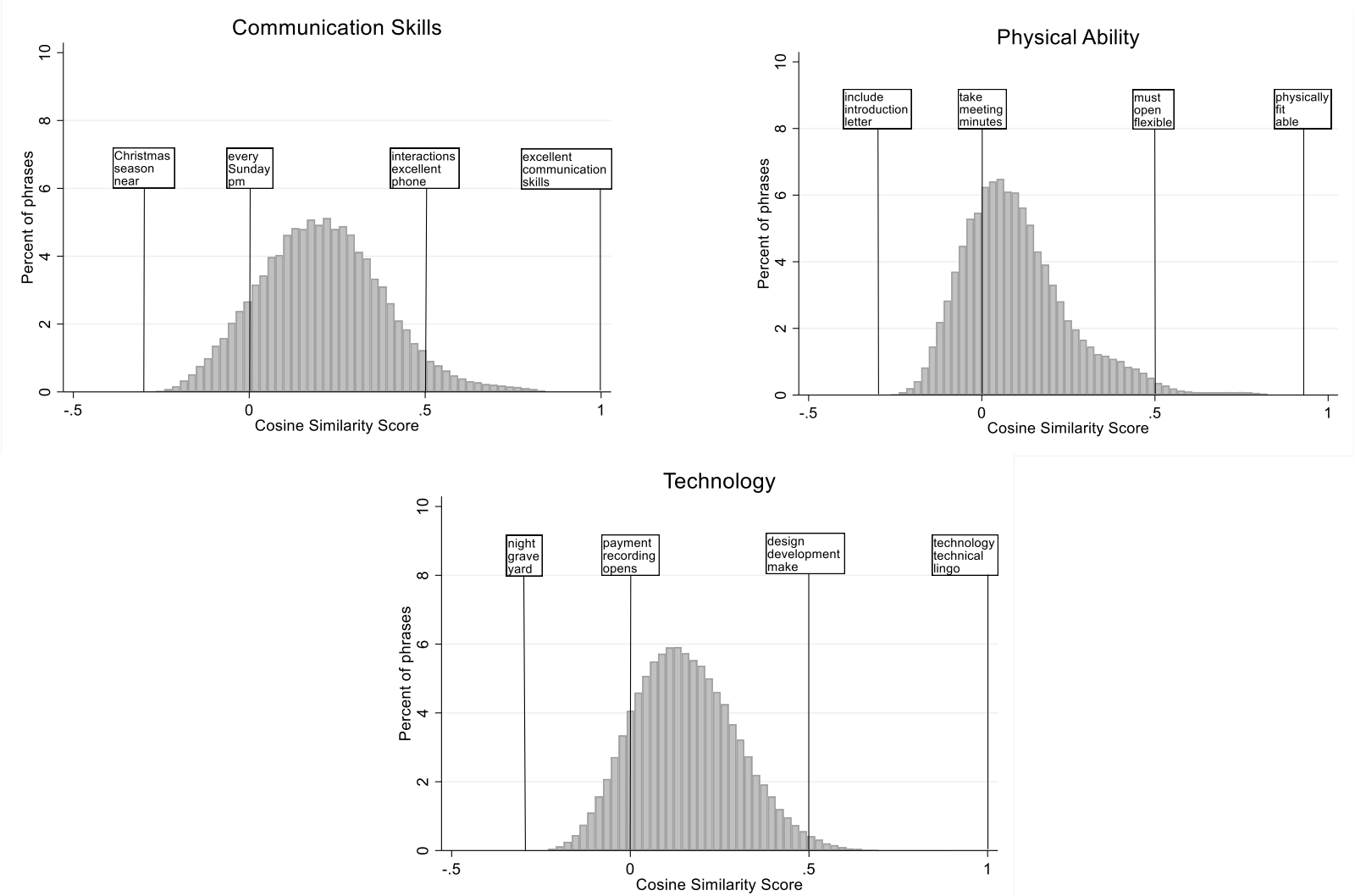
Note: MTURK participants self-reported their demographic characteristics. Respondents who selected two or more of the race and ethnicity categories were grouped into the "Two or More" group.

Table 4: Differences in Beliefs by Treatment (Negative Implies More Biased Against Older Workers)

	Self-beliefs			Beliefs about all respondents			Beliefs about respondents over 50		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	-0.886***			-1.266***†††			-1.248***		
	(0.047)			(0.050)			(0.053)		
	[-0.718]			[-0.900]			[-0.873]		
Treatment × Communications		-0.316***	0.015		-0.465***†††	0.001		-0.450***	0.004
		(0.033)	(0.022)		(0.050)	(0.046)		(0.045)	(0.038)
		[-0.256]	[0.013]		[-0.330]	[0.001]		[-0.315]	[0.003]
Treatment × Physical ability		-1.623***	-1.526***		-2.153 ***†††	-2.060 ***†††		-2.079***	-2.066***
		(0.098)	(0.106)		(0.091)	(0.099)		(0.101)	(0.108)
		[-1.315]	[-1.237]		[-1.531]	[-1.465]		[-1.454]	[-1.445]
Treatment × Technology skills		-0.895***	-0.488***		-1.403***†††	-1.040***†††		-1.432***	-1.108***
		(0.064)	(0.063)		(0.072)	(0.078)		(0.078)	(0.083)
		[-0.725]	[-0.395]		[-0.998]	[-0.739]		[-1.001]	[-0.775]
Treatment × Communications × AARP			-1.327***			-1.865***†††			-1.819***
			(0.105)			(0.106)			(0.114)
			[-1.705]			[-1.326]			[-1.272]
Treatment × Physical ability × AARP			-0.288***			-0.281***			-0.040
			(0.091)			(0.076)			(0.076)
			[-0.233]			[-0.200]			[-0.028]
Treatment × Technology skills × AARP			-1.629***			-1.455***†			-1.296***
			(0.099)			(0.090)			(0.096)
			[-1.320]			[-1.035]			[-0.906]
Physical ability		0.179***	0.179***		0.172***	0.172***		0.152***	0.152***
		(0.046)	(0.046)		(0.053)	(0.053)		(0.056)	(0.056)
		[0.145]	[0.145]		[0.122]	[0.122]		[0.106]	[0.106]
Technology skills		0.011	0.011		-0.052	-0.052		-0.079	-0.079
		(0.049)	(0.049)		(0.060)	(0.060)		(0.062)	(0.062)
		[0.009]	[0.009]		[-0.037]	[-0.037]		[-0.056]	[-0.056]
Adjusted R^2	0.149	0.230	0.353	0.211	0.332	0.452	0.205	0.321	0.422
Observations	2,718	2,718	2,718	2,718	2,718	2,718	2,718	2,718	2,718

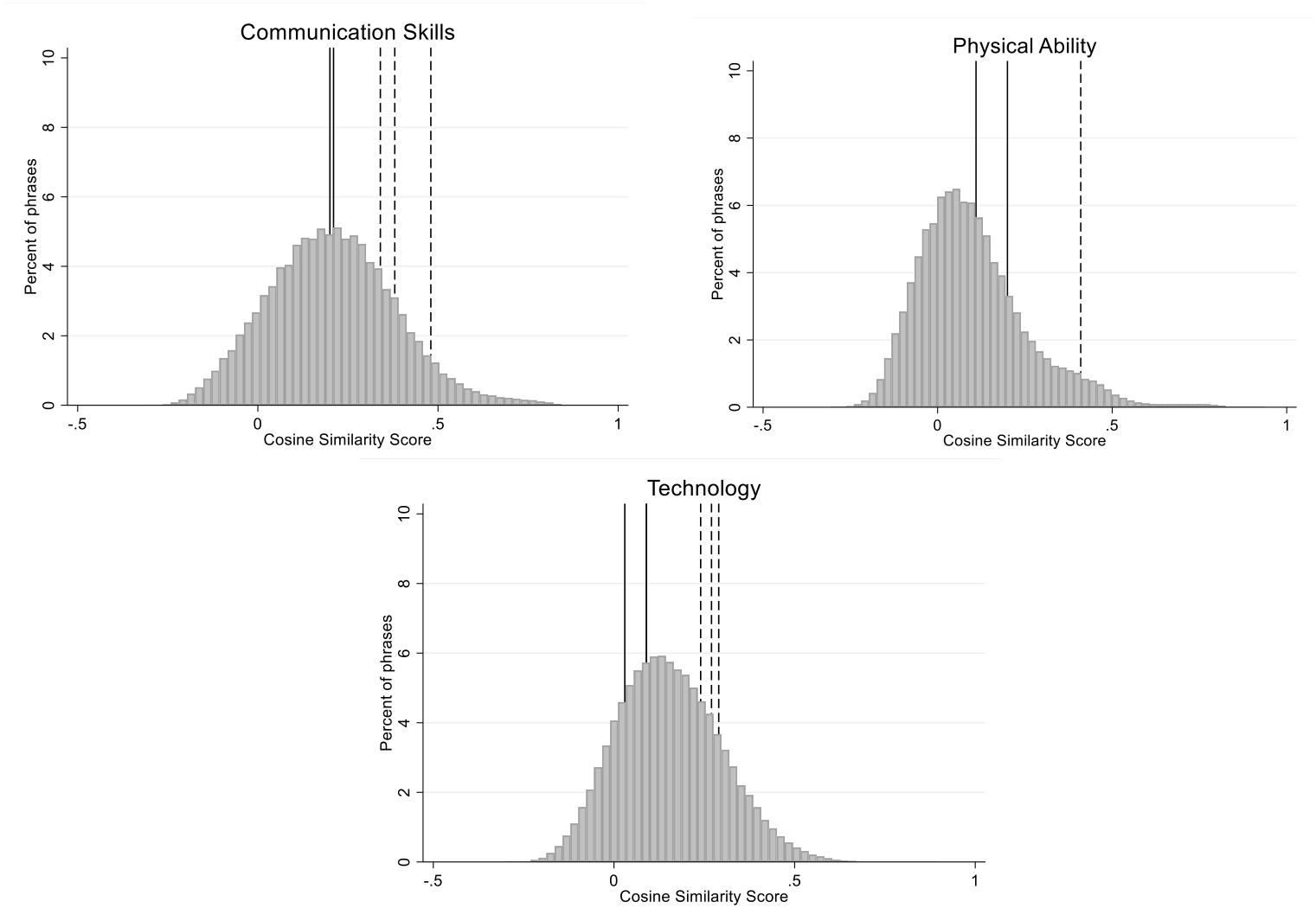
Note: * indicates statistical significance of the coefficient. * p<0.1,** p<0.05,*** p<0.01. † indicates significant differences between the same coefficients in the Beliefs about all respondents and the Self-beliefs models. † p<0.1, †† p<0.05, ††† p<0.01. In each regression, we include a constant and controls for gender, level of education, race, and age (not reported). Numbers in parentheses are robust standard errors clustered at the respondent level; numbers in brackets are coefficients normalized to standard deviations of the outcome variable. Negative numbers indicate higher levels of perceived bias against older workers, as the outcome ranges from 1 for "strongly agree" to 5 for "strongly disagree."

Figure 1A: Distributions of Cosine Similarity (CS) Scores



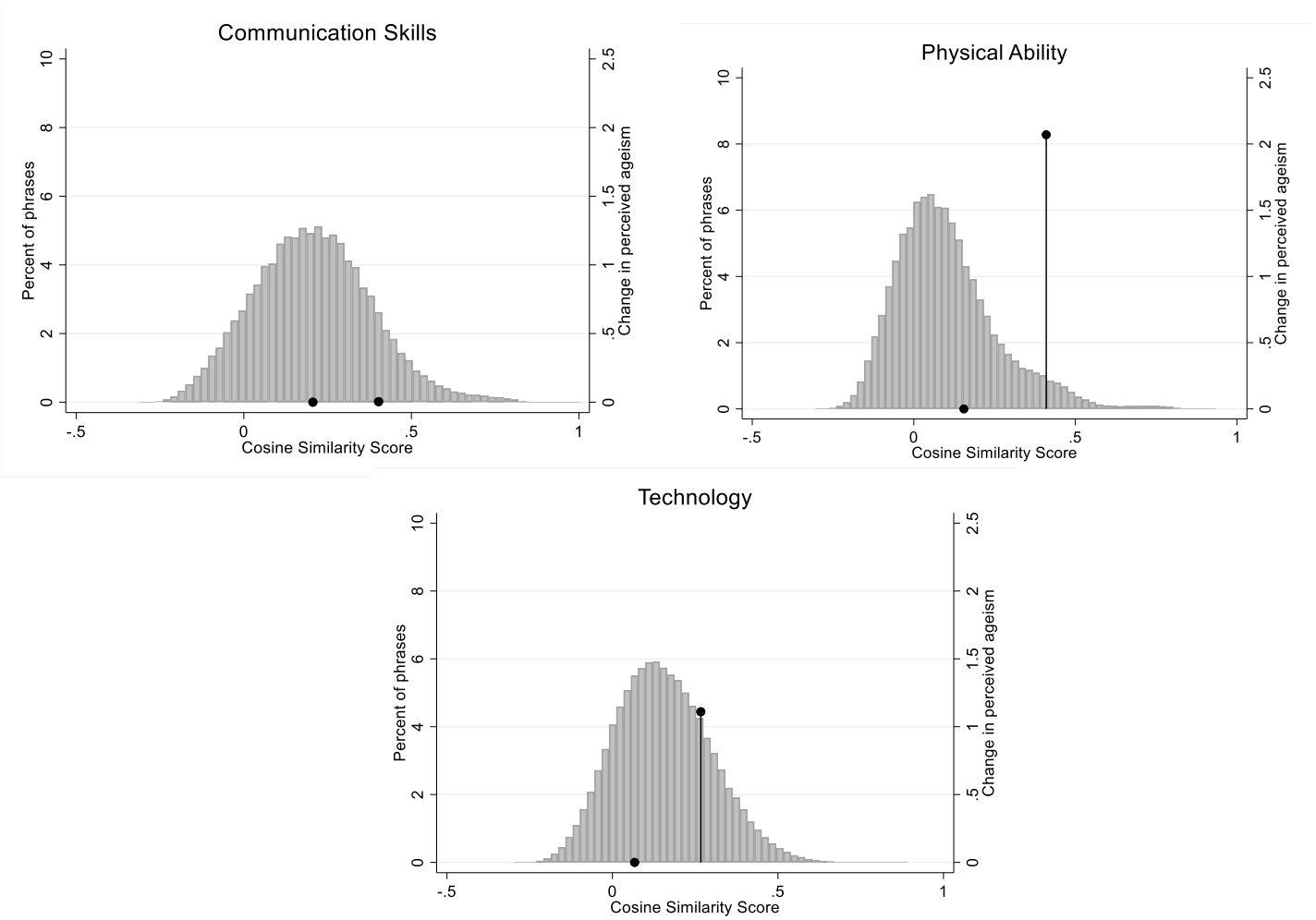
Note: Figure reports the distribution of cosine similarity scores for all trigrams from the job ads with the indicated stereotypes. The higher the cosine similarity score, the more related the trigram is to the stereotype, with a minimum of −1 and a maximum of 1. The phrases in the boxes are examples of phrases located at that point in the distribution.

Figure 1B: Locations of Treatment and Control Phrases in the CSS distribution of Job Ad Phrases



Note: See figure notes in Figure 1A for description of cosine similarity scores. 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). the

Figure 1C: Comparing the Distribution of CSS Scores and Perceived Ageism by Stereotype



Note: See Figure 1 for description of cosine similarity score. The dark points/lines are at the average cosine similarity score of the treatment and control phrases as shown in Table 2, for the indicated stereotype. The height of the right-hand dark point/line in each panel indicates the difference in the perceived ageism of the machine learning treatment phrases relative to the control phrases for individuals over 50 (Table 4, column 9).

Figure 2: Job Ad Examples

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)

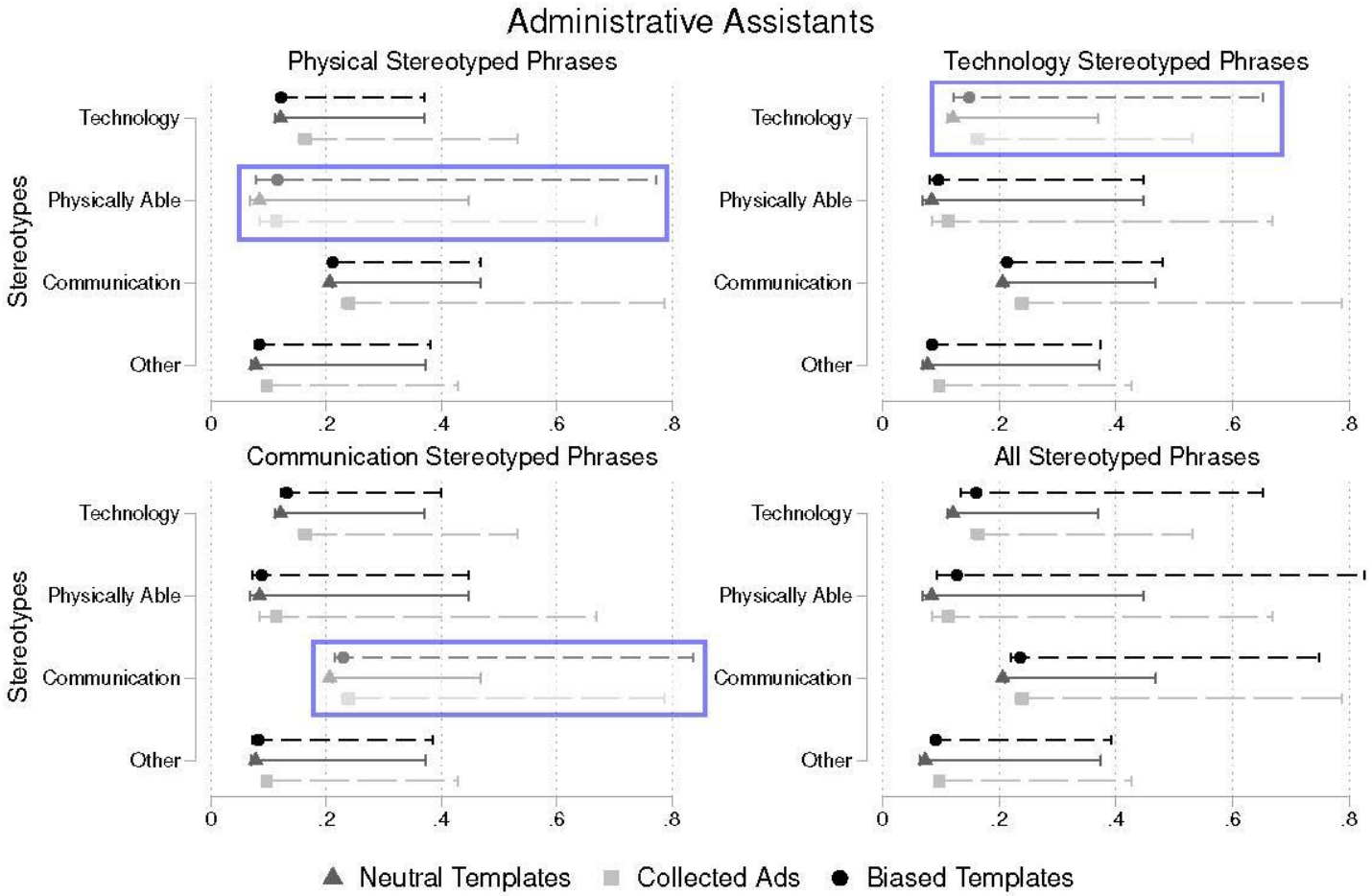
Our women's clothing store in ***City*** is looking for a sales associate to help us out weekday afternoons. We are pretty busy store and you must ***Physical Requirement*.** We are looking for someone with open to working in retail, who ***Communication Skill Requirement*.** We need you to ***Technology 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.

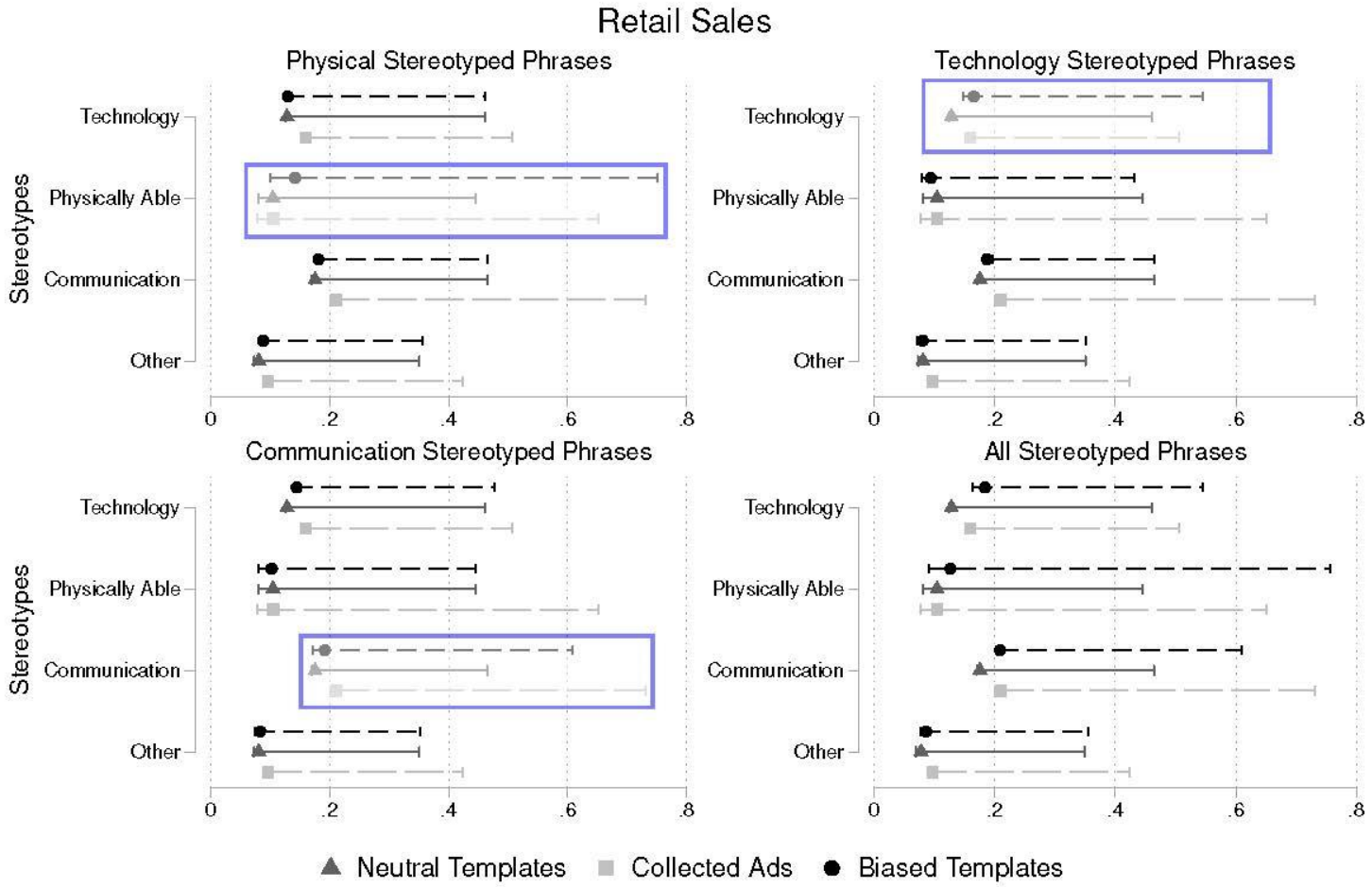
Requirements
-Professional appearance & attitude
-Detail oriented
-***Communication Skill Requirement***
-***Physical Requirement***
-***Technology Requirement***
-At least 18 years of age
-Access to transportation

Figure 3: Cosine Similarity Score of Administrative Assistant Templates



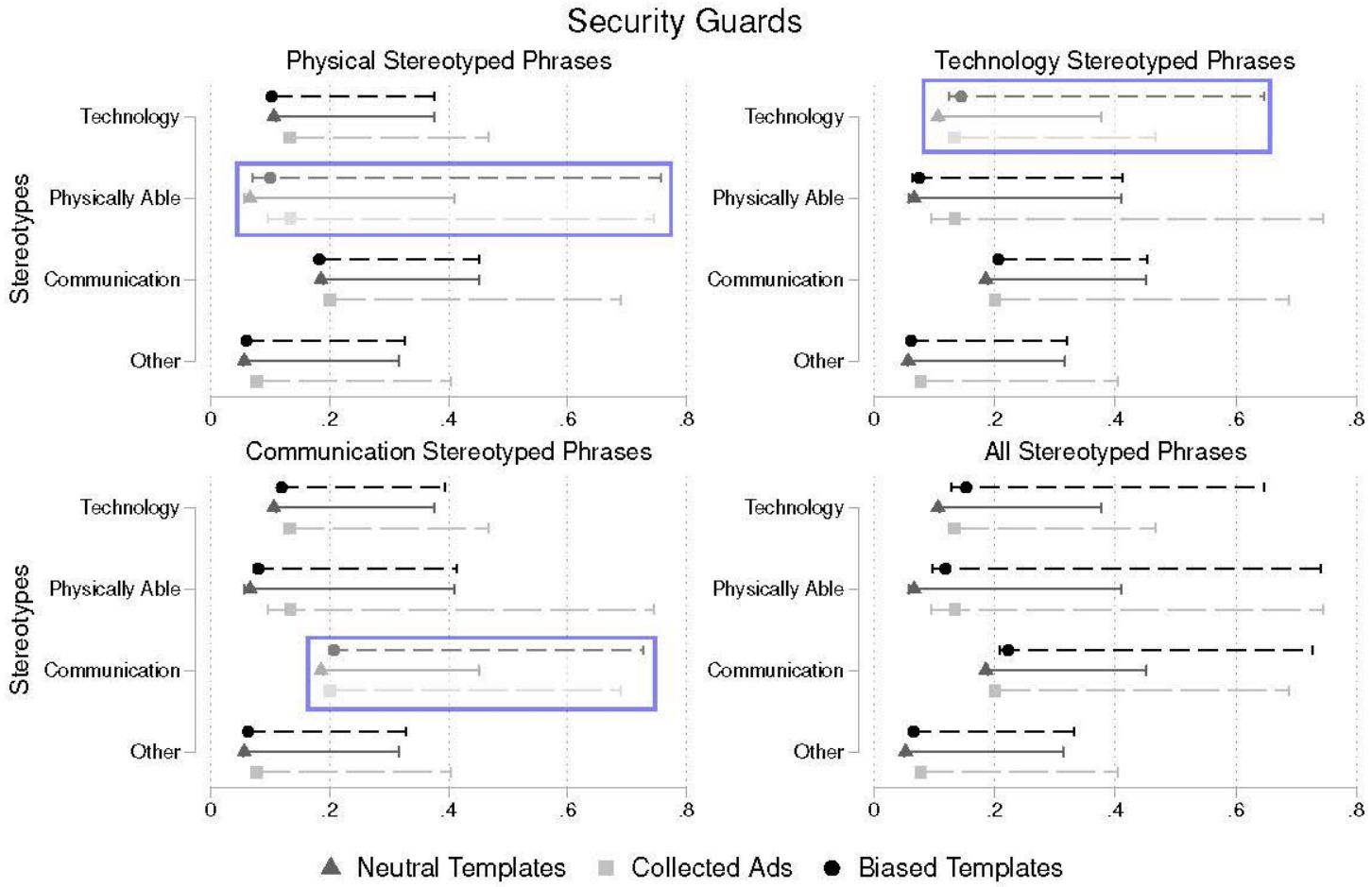
Note: Graphs display median to 99th percentile range of trigram semantic similarity scores for stereotypes for Administrative Assistant ads. The average trigram semantic similarity score for each stereotype is represented by the respective shape for each template. The category "Other" is the average of 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 4: Cosine Similarity Score of Retail Sales Templates



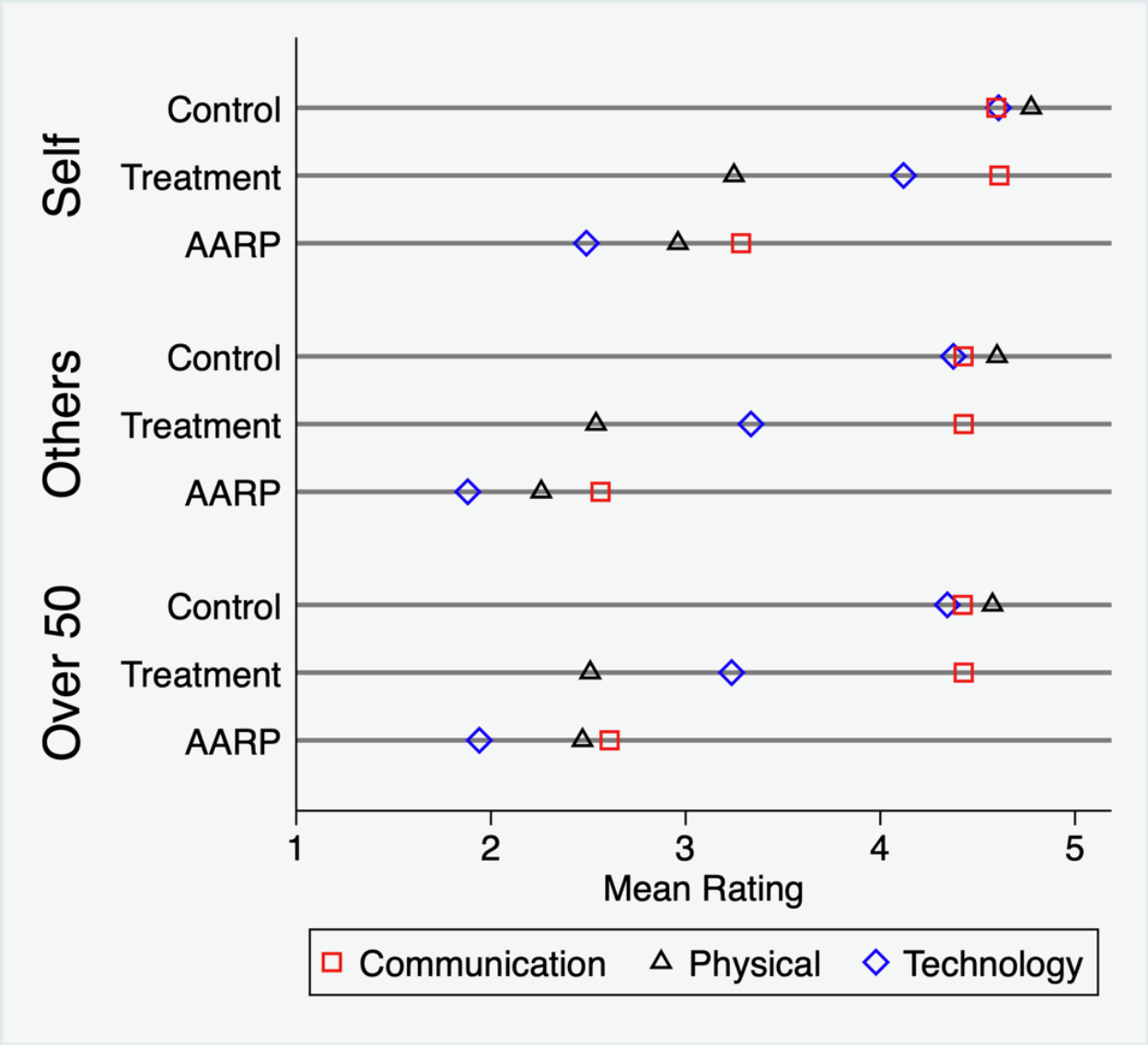
Note: Graphs display median to 99th percentile range of trigram semantic similarity scores for each stereotype for Retail Sales ads. The average trigram semantic similarity score for each stereotype is represented by the respective shape for each template. The category "Other" shows the averages for 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 5: Cosine Similarity Score of Security Guard Templates



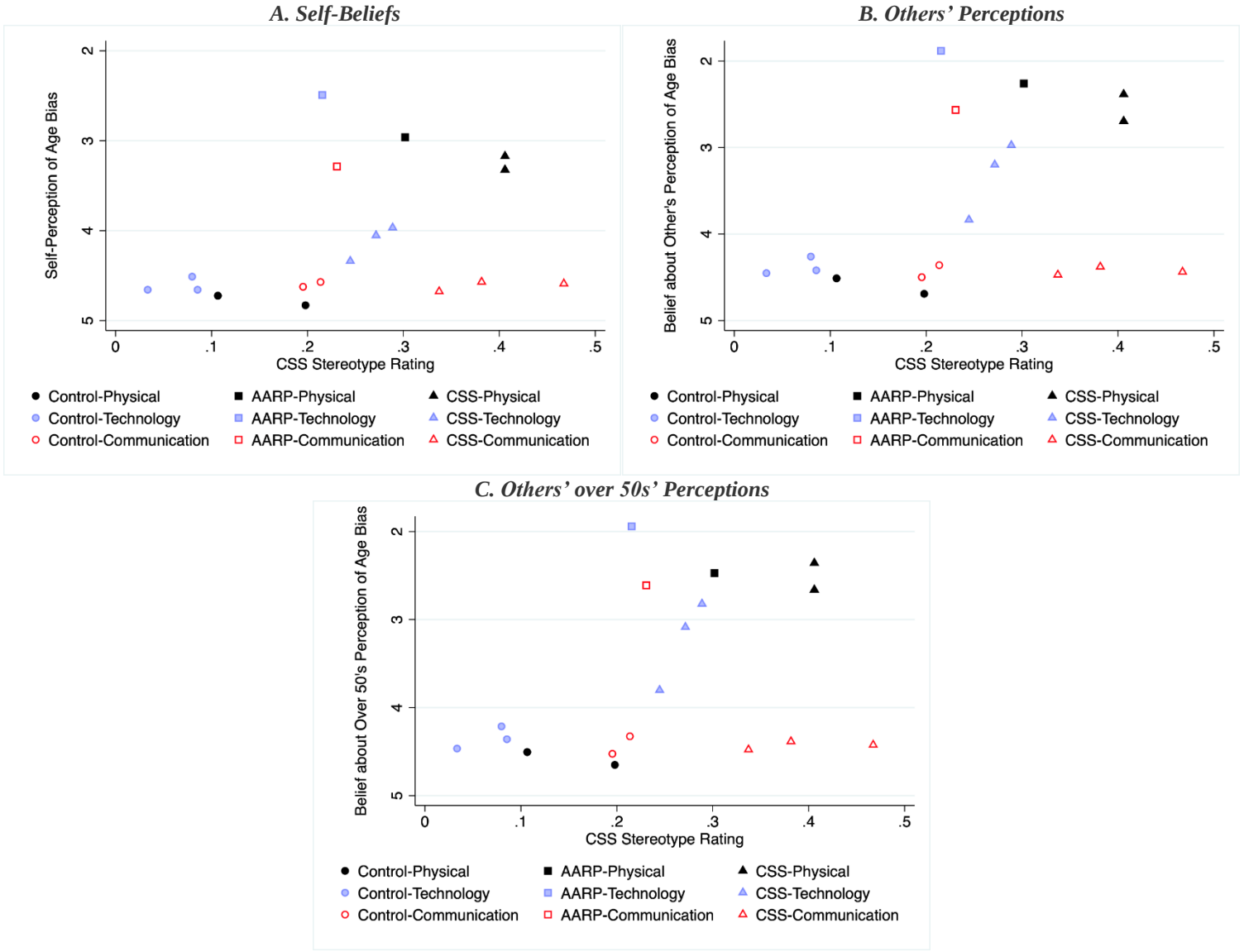
Note: Graphs display median to 99th percentile range of trigram semantic similarity scores for each stereotype for Security Guard ads. The average trigram semantic similarity score for each stereotype is represented by the respective shape for each template. The category "Other" is the average of the 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 6: 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 above 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 7: Scatterplot of Self-Beliefs of Age Bias and Cosine Similarity (CS) Scores



Note: CSS = "cosine similarity score." Figure plots MTURK respondents' average perceptions of age bias against CSS stereotype ratings from Table 2. Lower numbers on the y-axis indicate higher levels of perceived age bias. Higher CSS scores on the x-axis indicate higher average levels of semantic similarity of a phrase with its respective stereotype. Circular, triangular, and square markers represent control phrases, CSS treatment phrases, and AARP treatment phrases, respectively. Black (solid), blue (shaded), and red (unshaded) markers represent physical, technology, and communication phrases, respectively.