NBER WORKING PAPER SERIES

OLDER WORKERS NEED NOT APPLY? AGEIST LANGUAGE IN JOB ADS AND AGE DISCRIMINATION IN HIRING

Ian Burn
Patrick Button
Luis Felipe Munguia Corella
David Neumark

Working Paper 26552
http://www.nber.org/papers/w26552

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2019

We are grateful for helpful comments from seminar participants at the University of Bristol, University of Illinois, University of Liverpool, University of Maastricht, University of Tokyo, IZA, LISER, and Southern Methodist University. We thank Hayley Alexander and Emma Tran for excellent research assistance. We are especially grateful for the help of Nanneh Chehras, who assisted with the early stages of this paper. Patrick Button is thankful for generous grant support from the National Institutes of Health via a postdoctoral training grant to the RAND Corporation (5T32AG000244-23), which partly funded Patrick Button's work on this project from 2018-2019. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Older Workers Need Not Apply? Ageist Language in Job Ads and Age Discrimination in Hiring
Ian Burn, Patrick Button, Luis Felipe Munguia Corella, and David Neumark
NBER Working Paper No. 26552
December 2019
JEL No. J14,J23,J7,J78

ABSTRACT

We study the relationships between ageist stereotypes – as reflected in the language used in job ads – and age discrimination in hiring, exploiting the text of job ads and differences in callbacks to older and younger job applicants from a previous resume (correspondence study) field experiment (Neumark, Burn, and Button, 2019). Our analysis uses methods from computational linguistics and machine learning to directly identify, in a field-experiment setting, ageist stereotypes that underlie age discrimination in hiring. We find evidence that language related to stereotypes of older workers sometimes predicts discrimination against older workers. For men, our evidence points most strongly to age stereotypes about physical ability, communication skills, and technology predicting age discrimination, and for women, age stereotypes about communication skills and technology. The method we develop provides a framework for applied researchers analyzing textual data, highlighting the usefulness of various computer science techniques for empirical economics research.

Ian Burn
Swedish Institute for Social Research
ian.burn@sofi.su.se

Patrick Button
Department of Economics
Tulane University
6823 St. Charles Avenue
New Orleans, LA 70118
and NBER
pbutton@tulane.edu

Luis Felipe Munguia Corella
Department of Economics
University of California, Irvine
3151 Social Science Plaza
Irvine, CA 92697
lfmungui@uci.edu

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

Introduction

The most credible studies demonstrating the existence of discrimination in hiring are field experiments – more specifically, resume-correspondence studies (Fix and Struyk, 1993; Gaddis, 2018; Neumark, 2018). These studies have been applied to discrimination based on race, ethnicity, sex, age, and other group membership (e.g., disability). In this paper, we develop and implement methods to explore the role of stereotypes in hiring discrimination using the text of job ads, a type of evidence from resume-correspondence studies that has largely been ignored in the previous literature. We apply this method to evidence on age discrimination, although the method is applicable to resume-correspondence studies for other groups, and the techniques used are applicable to a wider range of empirical research questions in economics.

Age discrimination is of great policy interest in the United States and other countries because of rapidly aging populations. Low labor force participation rates of older individuals imply that aging populations lead to rising dependency ratios, which in turn strain the finances of many public programs targeted at older individuals, especially retirement and health care programs. As a result, there is an imperative to increase the employment of older individuals. The hiring of older individuals is likely an important part of the solution. Nearly half of older workers move to "bridge" jobs or "partial retirement" jobs (see, e.g., Johnson, Kawachi, and Lewis, 2009) before transitioning to complete retirement, or leave retirement to take jobs before retiring again (so-called "unretirement").[1] Age discrimination may hinder the ability of older individuals to move into new jobs or to re-enter the workforce.

Resume-correspondence studies of age discrimination create fictitious but realistic job applicants who are on average equivalent except for age, which is signaled through school graduation year(s). Researchers use the fictitious job applicants to apply for real job openings, and age discrimination in hiring is measured by comparing interview request rates ("callbacks") between older and younger applicants. Previous resume-correspondence studies almost always point to substantial age discrimination in hiring (Bendick, Jackson, and Romero, 1997; Bendick, Brown, and Wall, 1999; Riach and Rich, 2006, 2010; Lahey, 2008;

[1] This behavior is usually anticipated and often is not due to some adverse economic event during retirement (Maestas, 2010).

Baert et al., 2016; Farber, Silverman, and von Wachter, 2017; Farber et al., 2019; Carlsson and Eriksson, 2019; Neumark, Burn, and Button, 2016, 2019; Neumark et al., 2019).

Recently, we conducted a large-scale field experiment studying age discrimination in hiring, focusing on potential sources of bias in past studies. We found compelling evidence of age discrimination – especially against older women (Neumark et al., 2019, henceforth NBB).[2] Our goal in the present paper is to advance the experimental literature on age discrimination in a direction that helps us understand what underlies age discrimination, delving inside the black box of why or how employers discriminate based on age. Specifically, we use the text data from the job ads in NBB to explore whether – and if so which – age stereotypes are associated with actual discrimination by employers. This inquiry is motivated by research in industrial psychology (and related areas), discussed in detail below, documenting that employers and others have negative stereotypes about older workers – such as lower ability to learn, less adaptability, worse interpersonal skills, less physical ability, lower productivity, worse technological skills and knowledge, and less creativity – all of which can deter their hiring.

Little is known about which stereotypes employers act on when making actual hiring decisions. The industrial psychology literature mostly uses small surveys given to students, or a general population, who are asked about their attitudes concerning older individuals but not necessarily in employment contexts, let alone the specific context of older workers seeking new jobs. Even in the less common case in which researchers use a sample of managers with hiring experience, in their actual roles as managers they may not act on these stereotypes (or may act on only a subset of them). It may also be difficult for survey respondents to honestly reveal discriminatory preferences, stereotypes, or values if they are socially undesirable (e.g., Barnett, 1998; and Krumpal, 2013).

For these reasons, in this paper we pursue evidence on the importance of age-related stereotypes for actual labor market behavior. We provide, to our knowledge, the first study that links age stereotypes to

[2] NBB provide an extensive discussion regarding the interpretation of resume-correspondence study findings as reflecting age discrimination. Here, we simply interpret the evidence this way, and refer readers to that paper for discussion of this issue.

evidence on age discrimination in hiring.[3] We use the text data in the thousands of job advertisements from our field experiment, and explore what job-ad language related to age stereotypes predicts age discrimination in hiring. For example, one stereotype against older workers is that they are not as good with technology (McCann and Keaton, 2013). Job ads could contain language related to this stereotype (e.g., "must be a technological native"). We can then ask whether job ads containing such language are less likely to result in callbacks for older job applicants.

We find evidence that language related to stereotypes of older workers sometimes predicts discrimination against older workers.[4] For men, our evidence points most strongly to age stereotypes about physical ability, communication skills, and technology predicting age discrimination, and for women, age stereotypes about communication skills and technology.[5]

This paper makes three main contributions. First and most important, we are the first to create a detailed methodology, leveraging machine learning and textual analysis, to analyze the text data that is often available from field experiments on discrimination.[6] This can include, for example, job ads in studies of labor market discrimination or rental ads in housing audit studies. As audit and correspondence studies expand to study more markets, there are potentially more ways to leverage text data.[7]

[3] There is one study that finds a link, albeit less directly, between age discrimination in hiring and age stereotypes. Carlsson and Eriksson (2019) conduct a resume-correspondence study and ask employers about stereotypes, finding that employers in their survey think that older workers have lower ability to learn new tasks, are less flexible/adaptable, and have less ambition. But they do not directly link the hiring outcomes to these survey responses about stereotypes.
[4] Of course, our methods do not speak to the role of stereotypes held by employers that are not manifested in job ads.
[5] As discussed later in the paper, there is some evidence from industrial psychology and related research of stereotypes that are favorable to older workers, and some that stereotypes that can either favor or disfavor them. We discuss the evidence on these stereotypes as well. Generally, we find that language associated with positive stereotypes of older workers sometimes predicts less age discrimination, but we do find cases where this language predicts more age discrimination. As one would expect, we find that language associated with ambiguous stereotypes is sometimes associated with less discriminination against older workers and sometimes with more discrimination against them.
[6] Most correspondence studies do not analyze textual data, but there are some that do so on a limited basis. Hanson, Hawley, and Taylor (2011) is the most notable example; they 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. Tilcsik (2011) identifies four words in job ads related to masculine stereotypes (decisive, aggressive, assertive, and ambitious) and links those to hiring outcomes in a study of discrimination against gay men. There is research on age and gender preferences in job ads in countries, such as China and Mexico, where stating explicit preferences is not illegal (Kuhn and Shen, 2013; Hellester, Kuhn, and Shen, 2014).
[7] For example, Kugelmass (forthcoming) does a small correspondence study of discrimination in access to appointments with mental health professionals, who have on-line profiles, and Ameri et al. (2017) do a correspondence study of discrimination in access to AirBnB rentals. Both studies use platforms in which there is text data that could potentially by analyzed.

In addition to the evidence we provide from the study to which we apply this methodology, an important feature of our work is to develop a systematic way to use textual data in future studies, based on an a priori classification of the language that can be developed independently of the analysis of the relationship between the coding of language and outcomes in the experimental data. Our method of analyzing and characterizing the language from job ads or other text has been developed with the goal that researchers doing future correspondence studies or other types of studies who wish to utilize the text of the ads or other sources of information could pre-register the use and "output" from this method before collecting the data.

Our second contribution is to produce evidence on which age-related stereotypes that appear in job ads are associated with hiring discrimination against older workers – the first evidence we know of that can establish relationships between age-related stereotypes and actual employer behavior. Understanding which stereotypes underlie age discrimination can point to policy responses for reducing age discrimination. For example, job training, job coaching, or educational campaigns can focus on addressing the relevant negative stereotypes, or efforts could be focused on improving hiring practices, perhaps by increasing the information available to employers that reduces the attribution of stereotypes to older workers to whom they do not apply.

Third, our analysis provides evidence on whether employers with less intent to hire older workers – as captured in our experimental results – use ageist language in their ads. An extreme version of such language is stating maximum experience levels in job ads – as occurred recently in *Kleber v. Carefusion Corp.* – which will clearly act to exclude many older applicants.[8] More generally, the Code of Federal Regulations covering the ADEA currently state, "Help wanted notices or advertisements may not contain terms and phrases that limit or deter the employment of older individuals. Notices or advertisements that contain terms such as age 25 to 35, young, college student, recent college graduate, boy, girl, or others of a similar nature violate the Act unless one of the statutory exceptions applies" (§1625.4). Thus, our work can provide information to agencies that enforce age discrimination laws on job-ad language that may predict

[8] 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). See the discussion of the ruling in this case in Button (2019).

employer discrimination in hiring.

Background and Data from the Previous Resume-Correspondence Study

To obtain estimates of age discrimination in hiring, NBB conducted a large and comprehensive resume-correspondence study of age discrimination. The study used realistic but fictitious resumes for young (aged 29-31), middle-aged (aged 49-51), and older (aged 64-66) job applicants. Extensive details on the experimental design are provided in NBB. Here, we summarize the key features of the study so that the job-advertisement data we exploit in the present paper can be understood.

The study entailed sending 40,223 applications (resumes) to 13,371 job positions in 12 cities (in 11 states). This is by far the largest resume correspondence study of hiring discrimination to date, and the large number of job ads included in the study is critical to the methods we use in the present paper. NBB sent applications for positions in occupations that, according to Current Population Survey data, both older and younger individuals often take as new jobs (hence likely bridge jobs): administrative assistant and retail sales for women, and retail sales, security, and janitor for men. NBB sent three applications per position: always one younger applicant, and two older applicants of different ages (49-51 or 64-66) or with different work experience histories.[9] NBB tracked callbacks – interview requests or similar positive responses from employers – and compared them by age.

Figure 1 presents the main descriptive evidence from NBB. Across all occupations and genders, older applicants (age 64-66) got fewer callbacks than younger applicants. (These differences were statistically significant in all cases, except for men applying for security jobs.) As Figure 1 shows, the magnitude of the discrimination against older women was larger. NBB present a number of more sophisticated analyses, but the basic conclusion remains the same.

[9] While some of the resumes sent were on average identical to isolate the effect of age, as in the usual resume-correspondence design, NBB also sent some older worker resumes with more realistic, longer work histories; arguably these applicants are more comparable to the younger applicants because their experience is commensurate with their age – like for young applicants. This was done to avoid the possibility of upward-biased estimates of age discrimination, as older workers would not normally have the same listed work experience as younger workers. We also used the different resume types to explore whether older workers who exhibit "bridging" behavior – the movement from demanding jobs or jobs with more responsibility to jobs that are more flexible or with less responsibility – experienced more discrimination. The results showed that the measured discrimination was generally insensitive to the work experience history on the resume (NBB).

Conceptual Framework

Why might employers use stereotyped language in job ads, and what might this predict for our analysis? One hypothesis is that employers who discriminate based on age use stereotyped language to try to shape the applicant pool, to reduce the likelihood that age discrimination is detected. Using language that conveys positive stereotypes related to young workers might discourage older workers from applying for the job (or negative stereotypes related to older workers – although that seems less likely and is, in fact, less likely in our data). This would lead to the underrepresentation of older applicants in the applicant pool.

Why is this valuable to a discriminating employer? Assume that the probability of an anti-age discrimination action is positively related to how much lower the ratio of job offers to applicants is for older applicants than for younger applicants. Then for the same *number* of older and younger hires, an employer who uses stereotypes that discourage older job applicants would have a lower probability of facing an anti-age discrimination action. Thus, we can test the hypothesis that discriminating employers use ageist language in job ads by relating the measure of age discrimination in the resume-correspondence study (differences in callback rates for older versus young applicants) to the age stereotypes in the job-ad language.[10] This hypothesis does not necessarily distinguish between taste and statistical discrimination, but rather just tests whether employers who do not want to hire older workers use stereotypes in job ads to facilitate their discrimination.

A second hypothesis is more closely related to statistical discrimination. Different jobs may have different requirements, which are stated in job ads. But employers may hold stereotypes about older job applicants in relation to these job requirements – for example, assuming that older workers are less likely to be able to do the heavy lifting that a job requires. This behavior is pure statistical discrimination.

While economists are interested in the nature of discriminatory behavior, both statistical and taste discrimination are illegal under U.S. law. EEOC regulations state: "An employer may not base hiring

[10] There is, though, a potential bias against finding evidence that job ads with ageist stereotypes lead to lower callback rates for older applicants, if the ageist language lowers the share of older applicants enough so that the employer does not have to discriminate much against older applicants to get the desired younger workforce. While this may seem implausible, it would only imply that our results would be stronger without this bias.

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."[11] This text does not refer to whether the stereotypes are correct (i.e., right on average) or not, although from an efficiency perspective, economists would likely be more concerned about incorrect stereotypes.

A somewhat different and more complicated question is whether job requirements reflected in stereotyped language in job ads, to the extent they result in less hiring of older workers, are legal, which 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."[12] In other words, a job requirement that is associated with less hiring of older workers is not necessarily illegal.

Our evidence does not speak to the potential legality of job requirements that reflect age stereotypes. However, evidence that such job requirements are associated with hiring discrimination against older workers would prompt important questions about the validity of these job requirements, and more so if we think the first hypothesis – that employers put these in ads to discourage older workers from applying – has some validity.

We do not necessarily know – nor do we need to take a stand – on why employers discriminate based on age. They may want to avoid older workers because of taste-based discrimination, or because of statistical discrimination. The potential implications for the observed relationship between stereotyped language and hiring are the same.

Methods

The key task in this paper is to classify job ads by the age stereotypes that appear in their language. To do this, we scrape the text of the job ads and use language processing software to identify language that conveys or relates to age stereotypes. We then use this classification of job ads to test whether employers

[11] See http://www1.eeoc.gov//laws/practices/index.cfm?renderforprint=1 (viewed September 15, 2019).
[12] See 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).

who use language in their job ads related to negative stereotypes of older workers are less likely to hire older workers – as captured in the experimental results.[13]

Our strategy was to specify the relationships between job-ad language and age stereotypes ex ante, prior to doing any analysis of which job-ad language predicts measured discrimination, and also to make the identification of which phrases from job ads predict discrimination mechanical. This dual strategy was intended to avoid the risk of cherry picking phrases from job ads that predict age discrimination, and of ex post rationalization of the results (finding which phrases in the job ads predict discrimination and then searching for age stereotypes related to these phrases).

Our steps are as follows: First, we identify common age stereotypes from the research literature in industrial psychology. Second, we use computer science methods on semantic similarity in text data to identify and code words and phrases in the job ads that are related to specific age stereotypes (Mikolov et al., 2013a and 2013b). Third, because we have a very large number of words and phrases in the job ads, we use machine learning methods to identify the words and phrases from the job ads that predict age discrimination in hiring. Finally, we use the machine learning results to analyze statistically whether the words and phrases that reflect age stereotypes are particularly predictive of age discrimination.[14] These steps are explained in the following subsections.

Identifying Stereotypes of Older Workers

We conducted a detailed review of the industrial psychology, communications, and related literature to identify age stereotypes that this literature identifies as applying to workers in their 50s and 60s. We relied on studies that were more likely to cover the cohorts covered by the data in NBB, as there may be differences in age stereotypes across cohorts (Gordon and Arvey, 2004); hence, we avoided studies published before the 1980s and studies that focused on non-western countries. We reviewed an extensive set of both literature

[13] And, as noted earlier, we also study age-related stereotypes that are not necessarily negative with regard to older workers.

[14] To be clear, however, one could interpret our procedures in the reverse order – first estimating models for which phrases in the job ads predict age discrimination, and then studying the relationship between these phrases and age stereotypes. We present our methods in the order in the text to emphasize that we specified the relationships between job-ad language and stereotypes ex ante – prior to obtaining any information on which job-ad language predicts age discrimination.

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.

If a study met these inclusion criteria, we compiled the list of the stereotypes that the study identified as applying to older workers. We also noted how the stereotype was described or phrased. Since studies often have similar stereotypes but phrase them differently, we grouped the stereotypes that were very similar into aggregate categories in a similar manner to the literature review and meta-analysis papers (e.g., Posthuma and Campion, 2007).[15] 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 a list of 17 stereotypes of older workers, 11 of which are negative (lower ability to learn, less adaptable, less attractive, worse communication skills, less physically able, less productive, worse with technology, less creative, worse memory, hard of hearing, and negative personality) and six of which are positive (more productive, dependable, careful, more experienced, better communication skills, and warm personality). Tables 1-3 list these stereotypes. Table 1 lists stereotypes related to health. In this case, all four stereotypes about older workers are negative. The table also shows the similar phrasings across studies for the aggregate stereotypes we assign. Tables 2 and 3 present the same kind of information for stereotypes related to personality and to skills.

Among the 17 stereotypes, based on the existing studies, two pairs are contradictory: worse communication skills and better communication skills, and less productive and more productive. In our empirical analysis, therefore, we explore the effects of these age-related stereotypes in both directions, which gives us evidence on the net effect of these related stereotypes – in favor of or against older workers.

Matching Stereotypes to Words and Phrases in the Job Ads

We want to identify words and phrases in the job ads that are related to the 17 stereotypes, with the goal of capturing all the ways that the stereotypes could reasonably appear in job abs. Figure 2 gives an

[15] For example, within the aggregate category of "Less Adaptable," we include: "resistant to change" (McGregor and Gray, 2002; Weiss and Maurer, 2004); "adapt less well to change" (Warr and Pennington, 1993); and "[less] flexibility" (Levin, 1988).

example of a job ad. The job ad contains phrases that, on the surface, could be related to these stereotypes, including, for example, "experience," "social skills," and "social networking." The complication is that we do not expect age stereotypes to be expressed in the job ads exactly as they are in the research literature. Rather, there are many words and phrases that could be related to these 17 stereotypes, so that the true number of stereotyped words and phrases in the job ads could be very large.

We use methods from computational linguistics to determine the semantic similarity between phrases, as explained below. This process includes two steps. First, we use machine learning to calibrate a model to identify the semantic similarity between words and phrases. In particular, we use machine learning to train a model using textual data from English-language Wikipedia.[16] The model has a structure that relates semantic similarities among the 885,424 words used in the job ads based on their usage in Wikipedia articles.[17] Second, we use this Wikipedia model to calculate the similarity between the 17 stereotypes and phrases consisting of these words in the job ads. We now turn to a more detailed explanation of our methods.

In the first step, we train the model using the entirety of English-language Wikipedia. The method uses neural networks, which are trained to reconstruct linguistic contexts of words, to take what would otherwise appear to be a jumble of words from the job ads (as well as the age stereotypes) and to sort them such that words that are used in similar contexts, as measured by Wikipedia, are placed closer together.

We use an algorithm called *word2vec* (Mikolov et al., 2013a and 2013b) to identify the similarity of two words using the context in which the words appear.[18] The *word2vec* algorithm uses a continuous "bag of words" algorithm to use the context of a word's usage to predict other related words. The model produces a

[16] We use the English Wikipedia corpus as of November 3, 2017. This included 5.4 million articles. See https://dumps.wikimedia.org/enwiki/ (viewed November 3, 2017). As is standard in the neural networks literature, we divide each Wikipedia article into paragraphs (Adafre and De Rijke, 2006). We further split the paragraphs into single sentences. Each sentence and paragraph is used as a separate document in the machine learning algorithm. The intuition is that sentences can provide information on closer relationships between words, like "ice" and "cold," while paragraphs are needed for more general relationships, like "ice" and "Antarctica," which are related but might be less likely to appear in the same sentence.

[17] Note that in English language there are fewer than 885,424 words. For example, the *Oxford English Dictionary*, second edition, includes 171,476 words in current use (https://www.lexico.com/en/explore/how-many-words-are-there-in-the-english-language, viewed September 15, 2019). But the job ads include names, places, misspellings, verb conjugations, etc.

[18] Our application of the *word2vec* algorithm is taken from https://radimrehurek.com/gensim/models/word2vec.html (viewed September 15, 2019). Readers interested in learning more about this method are directed to http://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.XOthPIhKiUl (viewed September 15, 2019) for an overview of the implementation of the *word2vec* algorithm and alternative applications.

vector space where each unique word from Wikipedia is given a corresponding vector in a vector space created by *word2vec*, and words that are used more similarly to each other are located closer together in the vector space. This vector space is the mathematical representation of the relationships between these words.

The structure of the *word2vec* neural network begins with the inputs (the entirety of English-language Wikipedia) and then uses a series of hidden layers, which are not observed by the researcher, to create the vector space. The hidden layers help sort the inputs and shrink the dimensions. Each layer takes in multiple inputs and produces one output, reducing the dimensionality of the vector without losing valuable information, acting as linear functions that weight all the inputs to produce an output. Each input in the layer has a weight, and the layer has a bias. For instance, if a layer has three inputs, it requires three weights and one bias.[19] The output of these weights and bias is a projection function that, using the estimated weights and biases, will place words from the input that are semantically similar to each other close to each other in the output vector space – i.e., a vector that can then be used to construct the semantic similarity between any two words, as we describe below. Using this estimated projection function, we can trace a path from any word to another word and represent their similarity as a numerical measure.

Figure 3 provides an illustration. In this case, there are five inputs that are closely related, hence (hypothetically) belonging to a single layer. The *word2vec* algorithm takes the vector of input words and projects them to an output vector. The output vector is ordered such that words that are more closely related to each other are placed closer to each other (hence, e.g., "muscle" is closer to "athlete" than to "carry," based on usage in Wikipedia). This example features a 5×1 vector projected onto a 5×1 output vector. There are a total of five words and only one node (the second dimension of the output vector) to define the context. Usually, to analyze semantic similarity with massive databases like Wikipedia, the recommended vector size is between 100 and 200 nodes. The more nodes, the more precise the model will be. We picked 200 nodes to

[19] For simplicity, imagine a neural network that exists in two dimensions (rather than the actual 200 dimension vector space we use). y is the output of the hidden layer, which is a cardinal number such that two words closer together in meaning based on their usage in Wikipedia will have numbers closer together. y is a linear function of dummy variables for every word in the layer (x), with weights and a bias correction that allows the projection function to shift up or down to improve the predictive power of the model: $y = w_1x_1 + w_2x_2 + w_3x_3 + b$. The bias correction is (b), and the weights (w) are the coefficients of the model. Note that the bias correction is equivalent to the intercept in a regression.

increase the precision in the measurement of semantic similarity.[20] Hence, the actual neural network we construct takes as its input an 885,424×1 vector containing all the words and projects it into an output matrix that is 885,424×200.[21] We use this matrix – the neural network created by our *word2vec* algorithm – to calculate the semantic similarity of two words based on a "cosine similarity score" (CS score), explained below.[22]

The next step is to use these similarity scores to identify the words in the job ads with usage (in Wikipedia) that is highly related to the usage (again, in Wikipedia) of our set of stereotypes.[23] However, to this point, our explanation (and the example in Figure 3) have been based on single words. Because a single word may often fail to contain enough information to the association with a stereotype (which are typically expressed in multiple words), we instead use three-word phrases from the job ads in our analysis, called "trigrams." We create these trigrams by removing words such as "the," "and," or "a" – so-called "stopping words" in language processing – and then creating all trigrams from the remaining words. The trigrams are all sets of three consecutive words excluding these stopping words. We retain the stereotypes as the number of words in which they are expressed in the first column of Tables 1-3 after removing the words indicating the direction of the stereotype, such as "more" or "less." Then, for each of the stereotypes, we calculate the CS score between the stereotype and all trigrams used in the job ads. This requires some explanation.

Because the *word2vec* model is created using single words, we have estimated weights only for

[20] Pennington et al. (2014) show that there is a considerable gain in the accuracy from 100 to 200 nodes, but after 200, the gains are very marginal (see Figure 2 of Pennington et al., 2014).

[21] In our *word2vec* algorithm, the creation of this neural network begins by working from the input layer to the output to determine the optimal weights and bias in each layer of the network ("forward propagation"). This step consists of estimating the probability that a word is between a set of other words. We select optimal weights and bias to minimize the errors of these predictions. But when using only forward propagation, the estimated output can have a high error rate. To improve the estimation, we update the biases and weights based on the error rate in the model's prediction using a process known as "backward propagation." This process of using both forward and backward propagation iterations is counted as a training iteration. For our purposes, we use five training iterations of the *word2vec* algorithm (the default setting in the *word2vec* package). After the five training iterations, we have fully calibrated the neural network and populated the vector space. Our final vector space contains one row for each of the 885,424 words used on the job ads, and 200 columns containing the estimated weights from the linear projection functions.

[22] For more details about cosine similarity and semantic similarity and these kinds of models, see Clark (2014) and Jurafsky and Martin (2017).

[23] Note that there are two pairs of stereotypes that are mirror images of each other: worse/better communication skills and warm/negative personality. For these pairs, we just combine the stereotypes into a single phrase. Worse/better communication skills becomes communication skills and negative/warm personality becomes personality. Thus, we end up looking at cosine similarity scores with these 15 stereotypes. When we discuss the results, below, we explicitly consider the evidence on these ambiguous stereotypes.

single words. To calculate the CS score between stereotypes and trigrams, we recover the weights applied to the hidden layer in the network that corresponds to the word in question, apply these weights to generate new weights for the trigrams and stereotypes, and then use the vectors of these new weights to calculate the CS score.

The first step in this process is to estimate the vector corresponding to the three words in the trigram (or the words in a stereotype). To do this, we add the weights element-by-element for each word.[24] For example, if the model uses two hidden layers and produces two weights for each word of three words in the trigram "able lift lbs," then the total vector of weights of the trigram is computed as:

$$able\ lift\ lbs = \begin{bmatrix}0.3\\0.2\end{bmatrix} + \begin{bmatrix}0.4\\0.1\end{bmatrix} + \begin{bmatrix}0.5\\0.2\end{bmatrix} = \begin{bmatrix}1.2\\0.5\end{bmatrix} \quad [1]$$

Using these vectors for trigrams and stereotypes, our next step is to estimate the CS score between them; in particular, we estimate the CS score between every trigram and every stereotype. The CS score measures the similarity between two vectors of an inner product space. The similarity between the vector of weights of a trigram and the vector of weights of a stereotype is given by the following equation.

$$\mathrm{CS}(\mathrm{trigram}, \mathrm{stereotype}) = \frac{\mathrm{dot\ product}(\mathrm{trigram}, \mathrm{stereotype})}{\|\mathrm{trigram}\|\ \|\mathrm{stereotype}\|} \quad [2]$$

where "trigram" and "stereotype" in the equation refer to the vectors of weights.[25]

The CS score varies between −1 and 1 (Clark, 2014). A CS score of −1 means the words never appear in similar documents (i.e., the sentences and paragraphs in Wikipedia). More positive CS scores indicate there is a greater semantic similarity. If the words coincide perfectly, the CS score equals 1. As an example, Figure 4 shows the distribution of CS scores of all trigrams with a particular stereotype (communication skills). Note that it is centered above zero, which makes sense since we are looking at text from job ads. To provide some examples, trigrams at the lower end of the distribution are highly unrelated.

[24] This procedure is derived from Mikolov et al. (2013c). They demonstrate that the relationships between words captured by the methods we use also capture relationships between small numbers of words (their focus is on pairs), based on addition or subtraction of the vectors corresponding to these words. As a prime example, the representation of the word queen can be roughly recovered from the representations of "king," "man,", and "woman" – i.e., queen ≈ king − man + woman.

[25] The || notation indicates the Euclidean norm, so, e.g., $\|[x, y]^T\| = (x^2 + y^2)^{1/2}$. Thus, when the vectors are identical CS = 1, and when the vector elements have the same absolute values but opposite signs, CS = −1. (The implication is that, as is the case, weights can be negative.)

These include “christmas season near” and “hotel near seattle” (both with scores of −0.3). Trigrams with scores close to 0.0 include “every Sunday pm” and “work year round.” Trigrams at the top of the distribution with scores of 1.0 include “excellent communication skills” and “prioritizing skills communication.”

Testing which Phrases Predict Callback Differences by Age: Bag of Words and Elastic Net

As explained earlier, our goal is to estimate the relationship between age stereotyped language in a job ad and the likelihood that older or younger applicants received callbacks. We hypothesize that job ads with negative age stereotypes will have relatively lower callback rates for older workers (while job ads with positive age stereotypes will have relatively higher callback rates for older workers). Thus, our next step is to use a machine learning algorithm to identify which trigrams predict differential treatment of older applicants. We use a “bag of words” machine learning method to capture meaningful phrases in the job ads, combined with an elastic net regression (first proposed by Zou and Hastie, 2005) to identify the phrases that best predict the probability of discrimination.[26]

Our data set includes all responses to the triplet of job applications sent in response to each job ad that could be matched to an employer and their job advertisement. It is possible to match 34,260 job applications to 11,420 job advertisements, corresponding to 22,840 observations for older and middle-aged applicants.[27] Our outcome is a dichotomous variable equal to one if the older applicant did not receive a callback but the younger applicant did, and zero otherwise. That is, if both applicants are called back, neither applicant is called back, or only the older applicant is called back (which is less common than the reverse case), we do not consider the outcome to reflect age discrimination.[28] In these cases, we code the outcome

[26] We use the Python program *sklearn*. See https://scikit-learn.org/ (viewed September 15, 2019).
[27] There are 4,266 applications that cannot be matched to a saved job ad. This can be due to a number of reasons; the most common was that an ad was not saved. In some cases, the ad was saved in the incorrect format and often cannot be scraped. (Research assistants were instructed to save all job advertisements as an HTML file, but there were instances of advertisements being saved as a PDF or a PNG file.) In total, 87% of applications are matched to a job ad.
[28] In theory, it is possible to impose an even stronger definition of discrimination on the data, defining discrimination as occurring if the younger applicant is called back but neither older applicant is. The challenge in using this definition is in the construction of the triplets. All triplets had one younger applicant and two older applicants, but the older applicants could either be middle-aged or older. So in some triplets the older workers will be a mixed pair, one old and one middle-aged. In these cases, the stronger definition of discrimination would require discrimination to occur against the applicant aged 49 to 51 and the applicant aged 64 to 66. However, for some of the occupations studied in NBB, we observed stronger evidence of discrimination against older applicants than middle-aged applicants, and sometimes no discrimination against middle-aged applicants. Thus, the way we define discrimination in this paper is better, as it results in separate estimates for middle-aged vs. younger applicants and older vs. younger applicants. This issue could

variable as zero. In 76% of cases, neither applicant was called back, while in 6% of cases, both applicants were called back. In 11% of cases, the older applicant was not called back and the younger applicant was, whereas the reverse occurred in 7% of cases.[29]

We use a Python program called Natural Language Tool Kit (NLTK) to turn the text of the 11,420 job ads into quantitative data by splitting the ads into trigrams (this is the "bag of words" method). There are 210,672 unique trigrams. For each job ad, we code a set of dummy variables corresponding to each unique trigram, equal to 1 if the trigram appears in the job ad and 0 otherwise. Because we have 22,840 observations (two older applicants per job ad) and 210,672 trigrams, there is a dimensionality problem of more independent variables than observations. Even absent this dimensionality problem, with a large number of trigrams we would face a challenge in determining which ones predict discrimination against older workers. With a very large number of regressors, a traditional regression model might provide a good fit to the data but would not necessarily predict well.

To explain our approach, suppose we have a model that predicts whether the older worker was discriminated against (y_i) as a function of the vector of trigram dummy variables based on the job ads (x_i). We denote the prediction model $Y(x)$. Suppose we get new data from a second experiment, where z_i is whether the older worker was discriminated against, and x_i is again the vector of trigrams on the job ad, with prediction model $Z(x)$. If our initial model $Y(x)$ is a good model, then $Z(x)$ should be very close to the new target z_i (i.e., the observed value of z_i given x_i.) Good estimators should have a small mean-squared prediction error (*PE*), defined as

$$\begin{aligned} PE(x_0) &= E\left[\left(Z(x) - Y(x)\right)^2 | x = x_0\right] \\ &= \sigma_\varepsilon^2 + Bias^2\left(Z(x_0)\right) + Var\left(Z(x_0)\right). \end{aligned} \quad [3]$$

Because we have a large number of regressors due to the large number of trigrams in our job ads, it

be avoided in future studies by simply sending pairs of applicants in response to each job ad.

[29] In each triplet sent to a job opening, there was one young worker and two older workers (randomly selected to be either middle-aged or old). Our unit of observation is each older applicant, so that each triplet produces two observations. Thus, discrimination against an older applicant is measured independently of whether the other older worker was called back.

is not possible to include all of them in the model.[30] Including too many regressors in the model increases the variance and leads to overfitting. When a model is overfitted, it will perform poorly when given new data and result in a high prediction error. If the goal of the model is a low prediction error, the way to optimize the prediction error of our model is to reduce variance at the cost of introducing more bias. In computer science, statistics, and machine learning, this approach is called regularization.[31] The method of regularization we use in this paper is called "elastic net." The elastic net algorithm weighs the benefits of adding more variables in order to pick up more local curvature in the model (i.e., fitting the outcome better) against the increased variance in the estimated coefficients due to the additional variables in the model. The prediction model is a linear function of the trigrams ($x\beta$), and the estimate is calculated as follows:

$$\widehat{\beta_{en}} = min_{\beta}\left(\|y - x\beta\|^2 + \lambda\big((1-\alpha)\beta^2 + \alpha|\beta|\big)\right). \quad [4]$$

The elastic net is a linear combination of the ridge and LASSO regularization processes. The advantage of the elastic net over ridge and LASSO methods is that it enforces sparsity (dropping irrelevant variables), but at the same time encourages grouping effects in the presence of highly correlated predictors. In most tests, elastic net overperforms LASSO and ridge in prediction error (Zou and Hastie, 2005). If $\alpha = 0$ we have a ridge regression, and if $\alpha = 1$ we have a LASSO regression.[32] The parameter λ is the penalization term. When $\lambda = 0$, the elastic net will produce the same coefficients as OLS (with all variables included). As λ grows larger, it drops more variables from the regression.

[30] When including dummy variables for each trigram, we will also encounter issues with respect to multicollinearity.
[31] For a full review of these methods see Bühlmann and Van De Geer (2011). For an introduction to regularization see http://uc-r.github.io/regularized_regression (viewed September 15, 2019).
[32] The elastic net optimization process is a combination of the ridge and the LASSO regularization processes (Zou and Hastie, 2005). These processes have the same goal of improving prediction by introducing bias and reducing the variance in our estimates, but they achieve this in two different ways. The ridge regression penalizes the size of parameter estimates. If the ridge regression sets the penalization parameter to 0, the coefficients are the same as OLS. As the penalization parameter grows larger, the coefficient size is penalized more. For high penalization parameters close to infinity, the value of the coefficient falls to zero. The problem for our setting is that ridge regression does not perform variable selection; while the coefficients may fall close to zero, the variables remain in the model (Breiman, 1996; and Zou and Hastie, 2005). Similar to ridge regression, LASSO achieves regularization by adding a penalty for non-zero coefficients. LASSO is harsher than ridge because it penalizes the sum of the absolute value of the coefficients. This leads many coefficients to be set to zero under LASSO, compared to ridge. The drawback of LASSO is that it will struggle if variables are highly collinear. It will load all of the effects of the correlated parameters onto one of the parameters and drop the rest. Therefore, it can perform quite poorly at group variable selection (Tibshirani, 1996; Zhang, 2004; and Zou and Hastie, 2005).

In our analysis, we split our data by age group (middle versus old), gender, and occupation, and run the elastic net separately with each subsample to allow the phrases that predict discrimination to vary.[33] The elastic net then selects the trigrams (three-word phrases) that are most predictive of discrimination against older workers – as measured by a callback to the younger worker but not the older worker. We use a value of $\alpha = 0.5$, which implies that we place equal weight on the ridge and LASSO penalizations.[34] The λ parameter of our models (which is selected through cross-validation, conditional on $\alpha = 0.5$) ranges between 0.002 and 0.035, implying that the models utilize the machine learning penalization. Therefore, our results are very different than what one would find just running OLS using our data. Given that we have over 200,000 independent variables, even at small values of λ our models drop many variables.

Testing the Relationship Between Stereotyped Phrases and Callback Differences by Age

Our final step is to test whether the words and phrases selected by the elastic net as predicting discrimination are more strongly related to the age stereotypes. The elastic net estimation produces a list of the trigrams that predict discrimination against older workers. And from the earlier analysis, we have a measure of the semantic similarity of these trigrams to the stereotypes, as measured by the CS scores. To judge whether there is a statistically significant relationship between negative age stereotypes in job-ad language and a lower likelihood of a callback to older applicants, using these two types of information, the counterfactual needs to be specified carefully, because even if there was no systematic relationship between this job-ad language and hiring outcomes, some trigrams associated with age stereotypes might randomly be expected to predict discrimination.

For example, suppose the elastic net randomly selected 100 trigrams as predicting discrimination. Then we would expect 10% of these 100 trigrams (10 trigrams) to be in the top decile of the distribution of CS scores for a particular age stereotype. (The same would hold true for other percentages and centile ranges,

[33] In addition to the trigram dummy variables, our elastic net models include the same controls for resume characteristics as used in NBB.

[34] In our baseline estimation, we selected α = 0.5 (which is the default of *sklearn*) because it saves computational time. It is possible to use cross-validation to estimate the optimal α. When we do this, we find that the elastic net prefers to place even more weight on the LASSO, and selects an α between the values 0.7 and 1. Most of the time, the optimal value of α is 0.7. Picking α based on cross-validation has almost no effect on the number of selected trigrams or which trigrams are selected. The only case where we find that an optimally selected α changes more than one or two trigrams is for the group of old male sales, where the number of trigrams falls from 89 to 42 (see Table 4, discussed below).

of course.) If the elastic net algorithm selects more than 10 trigrams that are in the top decile of CS scores, this would be evidence that trigrams related to that stereotype are predictive of discrimination against older workers.

To test if the elastic net is selecting more trigrams in the top decile than expected based on random chance, for each stereotype we first calculate the CS score for each of the trigrams selected by the elastic net. We then count the number of trigrams selected that were in the top 10%, 5%, and 1% of CS scores for each stereotype. For example, in the case of the elastic net algorithm for male retail sales applicants between the ages of 64 and 66, the elastic net selected 89 trigrams, and four of them were in the top 1% of CS scores for "communication skills." Thus, the share of trigrams in the top 1% of the CS score distribution was 4.49% (4/89). We conduct binomial tests of proportions to determine if the share of trigrams selected by the elastic net is greater than what would be expected from the elastic net choosing randomly. In this case, we would expect the elastic net to select 0.89 trigrams in the top 1% of the CS score distribution, and the difference between the share of selected trigrams in the top 1% is significantly different from 1% at the 5% level (the p-value of the binomial test is 0.0123).

Results

General Association between Job-Ad Trigrams, Age Stereotypes, and Age Discrimination

For each age-gender-occupation grouping, we run an elastic net to determine which phrases predict discrimination. Table 4 summarizes the results from our elastic net models.[35] The elastic net selects the trigrams that are most predictive of discrimination against older workers. In four cases, the elastic net algorithm produced very few selected trigrams (either zero, one, or two trigrams). Security guard job ads were particularly uninformative for men, with one trigram selected for older applicants and zero trigrams selected for middle-aged applicants.[36] Among women, older applicants for administrative assistant jobs had only one trigram selected, and middle-aged applicants for retail sales jobs had two trigrams selected. Not

[35] The variation in the number of observations in each cell is due to the randomization procedures described in NBB and the number of job ads in each occupation.
[36] For the latter group, we might not expect any trigrams to predict discrimination given we do not observe any discrimination against this group in NBB.

surprisingly, as described below for these cases, the number of trigrams chosen is too small to conclude that there is a significant difference between the elastic net trigrams and a randomly selected set of trigrams.[37]

When we do find a sizable number of trigrams that predict age discrimination, trigrams related to age stereotypes are frequently significantly overrepresented. Table 5 summarizes the results. We report the cases, for the top 10%, 5%, and 1% centiles, for which we find statistically significant evidence – at the 10% level or better – of an overrepresentation of job-ad trigrams related to ageist stereotypes among the trigrams that predict discrimination. Table 5 does not provide information on the signs or magnitudes of the estimated relationships; we discuss these below. Briefly, though, in general we find that job-ad language associated with age stereotypes predicts discrimination against older workers.

In Figures 5a and 5b, we present a subset of results graphically – for two stereotypes – by comparing the CDFs of semantic similarity scores for all the trigrams in the ads to the CDFs of semantic similarity scores of the trigrams selected by the elastic net. If the elastic net is selecting trigrams that are more related (compared to a random selection of trigrams), then the CDFs of the semantic (cosine) similarity scores of the selected trigrams should lie to the right of the CDFs of the semantic similarity scores for all trigrams. (This is the same idea of overrepresentation of the selected trigrams in the upper tail of the distribution of semantic similarity scores, which we consider for specific centiles in Table 5.) Moreover, the evidence would be more compelling if the deviation is largest in the highest parts of the CS score distribution, where the trigrams are more related to the stereotype. Figure 5a presents a case where we see clear differences between the distribution of trigrams selected by the elastic net and the full distribution for "less physically able," for men. Note that this corresponds to the large number of significant differences in the "less physically able" row (for men) in Table 5. In contrast, in Figure 5b, we present a case where we observe no difference in the distribution of trigrams selected by the elastic net and the full distribution for "less creative," corresponding to no evidence of significant differences in the corresponding row of Table 5. While Table 5 lists the significant differences, Tables 6a and 6b list the full set of estimated proportions for these two stereotypes.

[37] For these cases none of the trigrams selected appear to be related to our stereotypes. See Appendix Table A1.

Figures 6a and 6b, and Table 6a and 6b, help explain how we analyze the data. But we find the statistical summary in Table 5 easiest to digest.[38]

To see the estimated relationships for the trigrams that predict age discrimination, Tables 7-12 present the elastic net regression results for the age-gender-occupation triplets for which we get more than a trivial number of trigrams selected (more than two – as reported in Table 4). The first column reports the elastic net coefficients for the variables selected for the age-gender-occupation triplet to which the table corresponds; these can be interpreted in the same way as coefficients from a linear probability model. The remaining columns are for the stereotypes for which we observe at least one instance of statistically significant overrepresentation in the top 10%, 5%, or 1% of the distribution of similarity scores between the trigrams and that stereotype (as reported in Table 5). In these columns, we report the semantic similarity score with the stereotype and the percentile in the distribution. In these tables, we highlight (in each column) the trigrams that are in the top 10% of the semantic similarity scores for the corresponding stereotype (which, as we already established, are overrepresented in the top decile).

Overall, our results suggest that stereotypes play a larger role in explaining discrimination against men than women and explain more of the discrimination against older men than middle-aged men. One can see this from the greater preponderance of significant differences for male jobs and applicants in Table 5, and similarly in the greater number of shaded entries for male jobs and applicants in Tables 7-12, and the same holds for older vs. middle-aged men. To list many examples, we find that trigrams related to stereotypes about the health (physical ability), personality, and skills of older workers are predictive of discrimination against older men. For middle-aged men, we find only that for janitors the trigrams related to the stereotype about physical ability predict discrimination, and in sales trigrams related to stereotypes about experience and technology predict discrimination. For older women, we find that trigrams related to stereotypes regarding their communication skills predict discrimination. For middle-aged women, we find some evidence that trigrams related to stereotypes about health, communication skills, dependability, and technology predict discrimination. To provide a more detailed discussion of these results, we consider in more detail the

[38] Tables like 6a and 6b for the other stereotypes are available from the authors upon request.

evidence on the stereotypes in the subgroups of stereotypes used in Tables 1-3.

Results for Stereotypes Related to Health

We first look at the stereotypes related to the health of older workers (less attractive, hard of hearing, worse memory, less physically able), listed in Table 1. We find little or no evidence that the trigrams that predict callback differences by age are disproportionately related to attractiveness, hearing, or memory; see Table 5.

For middle-aged women applying for administrative assistant jobs, there is one significant difference. The elastic net selected 23 trigrams as predicting discrimination against middle-aged women applying for administrative assistant positions (Table 4). Of these 23 trigrams, five were in the top 10% of the semantic similarity distribution for "memory." This overrepresentation of selected trigrams in the top 10% is significant at the 10% level (Table 5).[39]

The literature on age stereotypes summarized in Table 1 establishes that employers and others view older workers as possessing a worse memory than younger workers. This negative stereotype of older workers suggests that job ads that use language related to memory may be more likely to be associated with employers who discriminate against older workers. In Table 12 (corresponding to middle-aged female applicants to administrative assistant jobs), some of the trigrams that predict discrimination and are related to memory are in fact related to computers. This result appears entirely driven by "computer memory," rather than human memory. A result of this type indicates that we still must exercise caution in interpreting results from our empirical procedures, as some of the trigrams may have a high CS score with a stereotype but not reflect the content of the stereotype.

In contrast, Table 5 provides strong evidence that language related to physical ability is overrepresented among the trigrams that predict callback differences by age. The elastic net for old-male-janitor selected 21 trigrams as predicting discrimination (Table 4); 23.8% are in the top 10% of the semantic similarity distribution of "physically able" and 14.3% are in the top 5% (Table 5). These differences are

[39] One might still be concerned that we randomly get significant results. However, in this case the significant results would not be concentrated for specific stereotypes. Table 5 shows, however, that the significant results tend to show up for a subset of stereotypes, and the same result appears in analyses described below.

significant at the 10% level. The elastic net for old-male-sales selected 89 trigrams as predicting callback differences by age (Table 4): 16.9% in the top 10%, 11.2% in the top 5%, and 4.5% in the top 1%. These differences are significant at the 5% level (Table 5). For middle-male-janitor, the elastic net selected eight trigrams (Table 4), 12.5% of which were in the top 1%. This difference is significant at the 10% level (Table 5).

The literature on age stereotypes summarized in Table 1 establishes that employers and others view older workers as having worse physical ability and physical fitness than younger workers. This negative stereotype of older workers suggests that job ads that use language related to physical ability may be more likely to be associated with employers who discriminate against older workers. In fact, we find that ads with trigrams related to "physical ability" are associated with higher measured rates of hiring discrimination against older applicants.

For older applicants to janitor positions – reported in Table 7 – we estimate discrimination rates associated with these trigrams (the shaded cells in the "Less Physically Able" column) that are between 4.1 percentage points and 41.0 percentage points higher. For older men applying for sales positions (Table 8), trigrams highly related to "physical ability" predict discrimination rates that are between 0.0 to 22.3 percentage points higher. However, only one of the stereotypes ("able lift lbs") appears clearly related to physical ability. For middle-aged men applying for janitor jobs (Table 10), only one trigram that predicts discrimination is related to "physical ability," and ads with this trigram – "able lift lbs," which is clearly related to the stereotype – have estimated discrimination that is 21.7 percentage points higher.

Results for Stereotypes Related to Personality

Next, we turn to the same types of evidence for the second grouping of age stereotypes, for personality (less adaptable, careful, less creative, dependable, negative/warm personality; see Table 2). Table 5 shows that trigrams in the top 10% of the distribution of semantic similarity to adaptable make up 23.8% of the trigrams selected by the elastic net algorithm for older men applying for janitor positions. This difference is significant at the 5% level.

The research literature on stereotypes regarding personality, summarized in Table 2, establishes that

older workers are viewed by employers and others as less adaptable in the workplace. Given the negative stereotyping of older workers, we expect ads that use phrases related to this stereotype to be associated with higher discrimination against older workers. In the corresponding elastic net regression results for this age-gender-occupation cell (Table 7), we find that trigrams related to adaptability have discrimination rates that are 4.1 to 41.0 percentage points higher (Table 7).

We find, as reported in Table 5, that trigrams in the top 5% of the semantic similarity distribution for "careful" make up 14.3% of the trigrams selected by the elastic net for old-male-janitors (significant at the 10% level). Trigrams in the top 1% of the distribution make up 4.5% of the trigrams selected by the elastic net for old-male-sales (significant at the 5% level). The research literature on personality stereotypes (Table 2) establishes that older workers are viewed as more careful, and this is a positive trait, so we would expect trigrams associated with being careful to reduce the rate of discrimination against older workers. For older male applicants to janitor positions, this is not what we find. Trigrams related to being careful increase discrimination against older men by 4.1 to 10.5 percentage points (Table 7).

The results are a shade more mixed for older men applying to sales jobs, in Table 8 (the other age-gender-occupation triplet for which trigrams related to being careful are selected). Different trigrams related to being careful are associated with either higher or lower age discrimination – although more of the estimates point to higher discrimination. The effects of these trigrams range from a 3.4 percentage point decrease in discrimination to a 20.6 percentage point increase in discrimination; only one estimate ("strong work ethic") is in the direction of decreasing age discrimination.

Trigrams related to dependability are selected by the elastic net at statistically significantly higher rates for older male applicants, but not middle-aged men. Trigrams in the top 10%, 5%, and 1% are all overrepresented in the trigrams selected by the elastic net for older men applying to janitor positions (Table 5). Trigrams in the top 10% (16.7%) and the top 5% (10.1%) are overrepresented in the selected trigrams for older men applying for sales positions (significant at the 5% level; Table 5). For women, trigrams in the top 10% (21.7%) and top 5% (17.4%) are overrepresented in the selected trigrams for middle-aged women applying for administrative assistant positions (significant at the 10% and 5% level respectively).

As Table 2 shows, the research literature establishes that employers and others view older workers as more dependable. Do we find that trigrams related to dependability reduce discrimination? The evidence is mixed but leans in the opposite direction. For older men in janitor positions (Table 7), the coefficients on the trigrams relating to dependability from the elastic net are positive (consistent with more age discrimination), except for one trigram ("customer service skills"), which has a negative coefficient (a 0.7 percentage point decrease in discrimination). Among older male applicants in sales (Table 8), we find mixed evidence that job-ad trigrams related to dependability are associated with less discrimination. The coefficients of trigrams related to dependability vary from −3.9 to 5.2 percentage points. Among middle-aged female applicants to administrative assistant jobs (Table 12), we find evidence that some job ad trigrams related to dependability are associated with less discrimination, as predicted by the literature. The coefficients of the trigrams related to dependability range from −2.3 percentage points to 1.4 percentage points.

We find that trigrams in the top 1% of the semantic similarity distribution for "personality" (negative/warm) make up 3.4% of the trigrams selected by the elastic net for old-male-sales (significant at the 10% level; Table 5). The research literature on personality stereotypes (Table 2) establishes that older workers are viewed as having either warm or negative personalities, so there is no clear prediction as to whether trigrams associated with personality should reduce the rate of discrimination against older workers. For older male applicants to sales positions, we find that trigrams related to personality have estimated effects on discrimination against older men ranging from −3.4 to 7.9 percentage points, so the evidence is mixed, with more negative coefficients (indicating less discrimination) than positive coefficients (Table 8).

Results for Stereotypes Related to Skills

Finally, we turn to the stereotypes related to skills (lower ability to learn, better/worse communication skills, more experienced, more/less productive, worse with technology). We find evidence that job-ad trigrams that predict callback differences by age are disproportionately related to the stereotypes regarding communication skills, experience, productivity, and technology (Table 5). We find no evidence that stereotyped language related to the ability to learn is particularly predictive of different hiring rates by age (Table 5).

For communications skills, trigrams in the top 5% and the top 1% of the distribution are more likely to be selected by the elastic net trigram for older men applying for janitor positions (14.3% and 9.5%), significant at the 10% and 5% levels, respectively. For older women in sales, the elastic net selects 6.7% of its trigrams from the top 1%. This difference is significant at the 5% level. The elastic net for middle-aged women in administrative assistants selects 21.7% of its trigrams from the top 10% (significant at the 10% level) and 8.7% of its trigrams from the top 1% (significant at the 5% level).

The literature on stereotypes about older workers' communication skills is mixed, with some research suggesting that older workers are viewed as having worse communication skills, and some research suggesting the opposite (Table 3). Our results suggest that for older workers, the negative association is more prominent among the employers in our sample. Trigrams highly related to communication skills predict higher levels of discrimination for older men applying for janitor positions (Table 7), and older women applying for sales positions (Table 9). For middle-aged women applying for administrative assistant positions, we find more mixed results (Table 12); trigrams related to communication skills predict from 5.8 percentage points higher discrimination to 2.3 percentage points lower discrimination.

Turning to experience, we do not find that much evidence that the trigrams selected by the elastic net are strongly related to the stereotype that older workers are more experienced (Table 5). We find this only for middle-aged male applicants to sales jobs. Somewhat surprisingly, we find – in Table 11 – that job ads with trigrams that emphasize experience (e.g., "experience plus must" and "prior experience retail") are associated with higher discrimination against older workers.

Turning to productivity – where, like for communications skills, the literature on stereotypes is mixed (Table 3) – we find little evidence that trigrams selected by elastic net are disproportionately related to this stereotype; the one case is for older male applicants to janitor positions. In Table 7, we find that these trigrams are associated with more discrimination against older workers, suggesting that – at least in relation to hiring in our sample – this is a negative stereotype.

Finally, we examine the evidence on the stereotype that older workers are worse with technology. The literature on stereotypes is universally negative about the technological skills of older workers (Table 3).

However, we do not find so much evidence that the selected trigrams are disproportionately related to this stereotype. We do find some significant evidence of this overrepresentation for middle-aged male applicants for sales jobs and female applicants for administrative assistant jobs. As shown in Table 11, for middle-aged male sales applicants, the trigrams related to this stereotype are mostly related to higher discrimination against older job applicants, with one exception. The coefficients on these trigrams range from a 0.2 percentage point decline in discrimination to a 4.8 percentage point increase, suggesting that employers who emphasize technology in their job ad are less likely to callback middle-aged men. For middle-aged women applying for administrative assistant jobs, among these trigrams, there are more negative coefficients compared to middle-aged men in sales (Table 12 versus Table 11). The effect of these trigrams ranges from a 2.3 percentage point decline in discrimination against older workers to a 5.8 percentage point increase in discrimination.

False Positives

The *word2vec* algorithm could sometimes identify trigrams from the job ads that are not meaningfully related to our age stereotypes – which we might think of as false positives – and if these trigrams happen to predict lower relative callback rates for older job applicants, these false positives could generate bias towards concluding that job-ad language related to age stereotypes predicts age discrimination. Thus, in this subsection, we provide an admittedly subjective assessment of whether false positives from our *word2vec* neural network are a concern for our results by examining the trigrams selected by the elastic net from the perspective of a human reader rather than a machine.[40]

There is much evidence that the algorithm works well. For many of the trigrams selected by the elastic net algorithm, our *word2vec* algorithm assigns high CS scores to trigrams that appear related to the stereotype. In addition, among the selected trigrams, the *word2vec* algorithm often assigns a trigram to the correct stereotype. And trigrams that are given a higher CS score do in fact appear to be more related than

[40] One could also be concerned above false negatives – trigrams related to age stereotypes that *word2vec* fails to identify as semantically similar, and hence for which we do not obtain elastic net evidence on whether the trigrams predict age discrimination. These false negatives could generate bias in the opposite direction. However, there is no realistic way to assess the *word2vec* results for the far larger set of trigrams not selected by the elastic net that could represent false negatives.

those with lower CS scores.

Consider, for example, the results for the stereotype typed words related to "communication skills." In Table 7, for older male applicants to janitor jobs, all of the trigrams that are selected by the elastic net and are in the upper part of the distribution of semantic similarity scores with this stereotype – i.e., the shaded cells in the corresponding column – appear to be related to communication skills. These include "ability communicate effectively," "good communication skills," and "customer service skills." All three trigrams are in the top 1% of the semantic similarity distribution. For older female applicants to sales jobs (Table 9), the elastic net selects two trigrams that the *word2vec* algorithm says are in the top 1% of the semantic similarity distribution – "interpersonal communication skills" and "customer service skills" – which are both clearly related to "communication skills." (Of course, "interpersonal communication skills" is obviously a nearly one-to-one match for "communication skills.") Again, we find *word2vec* equating customer service with communication, an association that makes sense. For middle-aged women applying for administrative assistant jobs (Table 12), many of the trigrams selected by the elastic net and to which the *word2vec* algorithm assigns a high CS score with "communication skills" again feature customer service ("customer service experience," "customer service skills"). We also find that "written verbal communication" is a selected trigram. Perhaps the only potential false positives selected by the elastic net are "knowledge Microsoft office" in the top 10% of trigrams and "excellent computer skills" in the top 1%. ("Excellent computer skills" is likely ranked higher than "knowledge Microsoft office" because it includes "skills.")

On the other hand, we begin to notice selection of trigrams less clearly related to age stereotypes (false positives) when the threshold for the top decile of the CS score is low. Comparing potential false positives generated by *word2vec* for old-male-janitors (Table 7), we found none for "communication skills," where the top 10% of CS scores begins at 0.424. In contrast, for the stereotype "dependable" the threshold for the top 10% of CS scores is 0.240. And in Table 7, *word2vec* indicates that there are many trigrams that are among the most related to dependable, but in absolute terms they are fairly unrelated, and this is reflected in the lack of correspondence between the trigrams and the stereotype. As an example, the trigram "cleaning supplies equipment" has a score in the 92nd centile of the distribution, but a semantic similarity score of only

0.257.

This pattern seems to be stronger when a larger number of trigrams is selected by the elastic net. If we look at trigrams related to "dependable" from the old-male-sales elastic net (Table 8), many trigrams in the top 10% of the semantic similarity distribution do not appear to have such obvious connections to the stereotype. While a few are tangentially related ("must reliable transportation," "reliable transportation outgoing," "iPhone must reliable," and "strong work ethic") due to mentioning either situations where dependability would be important or using the word reliable in the trigram, there are obvious false positives.[41] Many of these trigrams have CS scores between 0.250 and 0.400.

These patterns suggest that higher absolute cutoffs produce fewer false positives, which makes sense, since the CS score is an absolute measure. This appears to be especially true when the top 10% (as opposed to a smaller percent) is a low threshold. Therefore, we would caution against interpreting the results of our analysis too strongly for stereotypes for which only the top 10% of trigrams are overrepresented but not the higher thresholds, or in cases where the average CS score in the top 10% of trigrams is low. We are especially concerned about distributions where the top 10% begins below about 0.3. In cases where we find an overrepresentation of trigrams in the top 1% or when the average CS score is high, we are more confident that these contain few false positives and represent a strong relationship with ageist stereotypes.

Table 13 summarizes our interpretation of the results using this rule-of-thumb on false positives. The table has the same structure as Table 5, and shows results for the same cells as Table 5. But Table 13 reports the threshold for the top decile and the mean of the CS score distribution, and bolds the estimates that are more reliable based on the false-positive criterion just discussed – CS score thresholds for the top decile of trigrams that are 0.300 or higher. Based on this rule of thumb, we have concerns about the trigrams related to memory, adaptability, dependability, personality (negative/warm), and experience. The distributions of the semantic similarity scores for these stereotypes have thresholds for the top decile that are low.[42] Conversely,

[41] These false positives include "transportation outgoing friendly," "outstanding customer service," "salaried positions strong," "customer service skills," "generous employee discount," and "friendly personality marketsource."
[42] The top decile for memory begins at a CS score of 0.215, the top decile for adaptability begins at 0.292, the top decile for dependable begins at 0.253, the top decile for personality begins at 0.244, and the top decile for experienced begins at 0.180.

we have more confidence in our results for less physical ability and careful (for men) and communication skills and worse with technology (for men and women).[43]

Supplemental Analyses

In this section, we test how robust our results are to the choices we made in designing the elastic net algorithm, and present other evidence assessing the validity of our results.

Number of Words in a Phrase

One choice we made that could influence our results was how many words to use in a phrase. As our baseline, we chose to use trigrams. When choosing the number of words in phrases, we balanced two competing factors: the predictive power of a phrase and the frequency with which a phrase appears in our job ads. These factors influence how many variables the elastic net will select, because the elastic net tries to be parsimonious. Assuming the same frequency of usage, when phrases are highly predictive, the model requires only a handful of phrases to predict discrimination against older workers. As the predictive power decreases, the model requires an increasing number of phrases to predict discrimination.[44]

Conversely, assuming the same predictive power, the frequency with which a phrase is used on the job ads follows a similar pattern. When the frequency of usage is low, more phrases are selected to predict discrimination because each phrase allows the model to classify only a handful of job ads. As the frequency of usage increases, the model requires fewer phrases to classify the same number of job ads as discriminating against older workers or not.[45]

If the gain to our model's performance due to the increase in predictive power of larger phrases is larger than the decline in predictive power due to lower frequency of usage, we would expect the elastic net

[43] The stereotypes for which the top decile begins above 0.300 have higher average CS scores than the stereotypes for which the top decile begins below 0.300, which further bolsters the evidence that these trigrams are not false positives.
[44] Consider the case where 50% of the times "worker" appears, it is paired with the word "efficient," and the other 50% it is paired with the word "need." Suppose further that every time "efficient" and "worker" are paired, the older worker is discriminated against, but when "worker" and "need" appear together the older and younger worker are treated identically. If we use one-word phrases to predict discrimination, the elastic net may not select "worker" as predicting discrimination because "worker" only predicts discrimination 50% of the time (as good as random). If we use two-word phrases, the elastic net will likely identify "efficient worker" as predicting discrimination.
[45] Consider the case where we have two phrases which both predict discrimination with 100% accuracy, but one phrase appears on 100 ads and the other appears on 10 ads. The elastic net is likely to select only the dummy variable for the phrase which appears 100 times, since it is 10 times more efficient at predicting discrimination. For the elastic net to be equally likely to prefer the phrase that appears only 10 times, that phrase would have to be 10 times more predictive than the more-frequent phrase.

to select more phrases as predicting discrimination when we increase the number of words in a phrase. If the increases in predictive power as the size of the phrases is increased is not large, then the decline in performance due to the decline in the frequency of usage will dominate. In this case, increasing the length of our phrases will decrease the number of phrases the elastic net uses to predict discrimination.

Consistent with our strategy of specifying the relationships between job-ad language and age stereotypes ex ante, we chose to use trigrams before doing any analysis of the relationship between the selected phrases, stereotypes, and measured discrimination. This avoided the risk of cherry picking – choosing the number of words to use in phrases (three, or something else) to obtain a particular set of results.

Nonetheless, after the fact, we also ran the elastic net using one, two, four, and five-word phrases, to examine the sensitivity of the results to changes in the number of words in a phrase. Table 14 reports the number of phrases the elastic net selected for each age-gender-occupation cell. The correlation between the words in a phrase and the number of phrases selected by the elastic net varies by occupation. For retail sales (both men and women) and administrative assistants, increasing the number of words in a phrase generally decreases the number of phrases selected by the elastic net. For janitors and security guards, increasing the number of words in a phrase generally increases the number of phrases selected by the elastic net, and the increase is sometimes large; when using four- or five-word phrases, the elastic net selects hundreds of phrases as predicting discrimination in some instances.

The results in Table 14 suggest that three-word phrases were a good choice, for a few reasons. First, they produce the most cells where we have enough phrases to test for the overrepresentation of highly related phrases (6 out of 10 age-gender-occupation cells). Second, trigrams produce the evenest distribution of cells with enough selected phrases across gender (4 of 6 male cells and 2 of 4 female cells). And third, the average number of selected phrases on average is small enough (20 phrases versus 93 phrases when using four words or 73 when using five words) that one can more plausibly consider the phrases individually (as, e.g., in the prior discussion of false positives). Also note that as the number of phrases selected increases, the magnitudes of the elastic net coefficients shrink. For example, we find that the elastic net selects 89 trigrams in Table 8, but one-third of these trigrams have coefficients with an absolute value of less than 0.0005. This

suggests that while these trigrams do predict discrimination, their effects are very often negligible. In Tables 7, 9, 10, and 11, when the elastic net selects far fewer trigrams, we rarely find trigrams with coefficients very close to zero. Moreover, we saw that when a large number of phrases is selected (for older male applicants to sales jobs, in Table 8), many had meanings less clearly related to the stereotypes.

To see how much the decision to use trigrams influenced our evidence on job-ad language and age stereotypes, Table 15 reports the results paralleling Table 5 – on the overrepresentation of the selected phrases in the upper tails of the distributions of CS scores – when we vary the number of words used to define phrases. We report these results for the findings shown to be significant and strong (in terms of a high absolute CS score) in Table 13; these correspond to the bold-faced cells in Table 13. When the elastic net selects enough phrases for us to perform the binomial test of proportions (say, more than three), we usually observe significant over-representation of selected phrases in the uppers tails of the distributions of the CS scores for phrases with different numbers of words as well. This occurs for "less physically able" for older male applicants in sales, "worse/better communication skills" for middle-aged female administrative assistant applicants, older male janitor applicants, and (mostly) older female sales applicants, and to a lesser extent "worse with technology" for middle-aged male applicants in sales – and hence reinforces our main conclusions about which stereotypes matter the most for men and for women.[46] The exception where this does not happen for any threshold for a stereotype is for the "careful" stereotype for older men in sales.[47]

A lesson from this analysis is that it would be valuable to run the type of analysis in Table 14 (and additional analyses studying the relationships between phrases and stereotypes) prior to doing any analysis of the associations with discrimination. In another context – such as ads with a template that restricts the language or allows more expansive language, or with different stereotypes (say, in a study of race discrimination) – a different number of words in phrases may better match to stereotypes. It can also, of

[46] For example, we find that 16.85% of phrases that predict discrimination against men aged 64-66 applying to sales jobs are in the top 10% of the physical ability semantic similarity score distribution. When we use four-word phrases, 21.82% of the phrases are in the top 10%, and when we use five-word phrases, 25.53% are in the top 10%.
[47] The fact that the elastic net fails to select more than three phrases in many instances as we vary the length of the phrases does indicate that our results are due in part to the decision to use three-word phrases. Specifically, men applying for security guard jobs are not analyzed because the elastic net did not select enough trigrams, but if we use five-word phrases we would have enough N-grams to analyze them.

course, be useful to explore the overall sensitivity of the findings to the number of words used.

Definition of Discrimination

A second choice we made was how to define discrimination. Our main focus has been to understand discrimination against older applicants, and hence we defined discrimination as a callback to younger applicants but not older applicants. It is possible that studying discrimination in favor of older applicants (against younger applicants) would detect more evidence of positive stereotypes reducing discrimination against older workers. We therefore also did analyses where we redefined the outcome variable in our elastic net estimation to be one if the younger applicant was not called back but the older applicant was called back. For this analysis, we created separate pairs of each older applicant in the pair combined with the corresponding younger applicant (even though this means younger applicants get used in two pairs). We did this because otherwise we would have to use a more stringent definition of favoring the older applicants entailing callbacks to *both* older applicants but not the younger applicant.

Table 16 presents the summary of our results if we redefine the outcome variable to be discrimination against *younger* workers (paralleling Table 5). There appears to be very little evidence that employers' preference for older workers is correlated with the words and phrases used on the job ads. As reported in the top row, in only two cases does the elastic net select more than one trigram. (Results are reported only for age-gender-occupation cells with at least one trigram selected.) The elastic net for older women in administrative assistant positions selects five trigrams as significant. Rarely are any of these five trigrams in the top 10% of the semantic similarity score. When they are, in the case of hearing and memory, we do not find statistical evidence they are overrepresented for hearing, but we do find evidence they are overrepresented for memory.[48] The elastic net for middle-aged women in administrative positions selects ten trigrams as significant. These trigrams are overrepresented in the top decile for a number of stereotypes, but in no case is this significantly different at the 5%-level or higher from what we would expect if we drew ten trigrams at random.

[48] We are hesitant to call this evidence strong because memory is a problematic stereotype given that the top decile of semantic similarity scores begins at a point where words are very unrelated. Thus, overrepresentation in the top decile does not indicate that words highly related to the stereotype predict discrimination. Indeed, the semantic similarity score of the two trigrams selected are only 0.25 and 0.27.

Though we do not find any significant results in Table 16, the results are important for what they tell us about stereotypes favoring older workers. We do not find a significant overrepresentation of trigrams in the top 10% for any stereotype where the literature was split as to whether the stereotype favored older workers. This bolsters our evidence that stereotypes about older workers that matter in the labor market – at least as reflected in our data and approach – are mostly negative, and that age-stereotyped language will, if anything, predict discrimination *against* older workers.

Placebo Analysis

The final analysis we report is, in a sense, a different way of thinking about the false positive problem. In particular, we have estimated models for measured discrimination that include massive numbers of potentially explanatory variables. We then showed that some of the chosen predictors of age discrimination are associated with age stereotypes, and this happens more than would occur randomly. The elastic net is designed to reduce false positives by utilizing cross-validation, but our setting features fewer observations than is common in the computer science literature. So, given our sample size and our decision about how many k-folds to use for the cross-validation (we use five), it is possible that the elastic net has not rooted out all the false positives. To test how well our method works, we assign a placebo outcome to see if our parameter and estimation choices lead to the selection of trigrams as predicting our placebo outcome.

Recall that we had 11% of cases where the older applicant was not called back and the younger applicant was – indicating age discrimination. Thus, we now randomly assign a placebo discrimination outcome to 11% of the triplets in each age-gender-occupation cell, and run the elastic net using these placebo assignments. Table 17 compares the number of trigrams selected by the elastic net using the true data to the number of trigrams selected when using a treatment that is randomly assigned. In almost all cases, we observe the elastic net selecting zero trigrams as predicting the discrimination outcome. The differences can be stark. For example, the elastic net for older men in sales returned 89 trigrams as predicting discrimination. Under the random assignment of the treatment, only five trigrams are predictive. We see similarly sharp declines for the other elastic net results. Indeed, most of them return no predictive trigrams under the placebo treatment. These results further reinforce the conclusion that the results we obtained using actual measured

discrimination do not reflect false positives.[49]

Conclusion

In this paper, we have developed a new methodology for analyzing the job ads collected during a resume-correspondence study. By combining different methods of machine learning, we are able to determine which words and phrases in those job ads predict discrimination, and to determine how related these words and phrases are to ageist stereotypes.

A key contribution of our methodology is that it can be adapted to other contexts. In audit or correspondence studies of labor market discrimination, regardless of the group studied, textual data is or can be collected. It may also be possible to apply our methods to studies of discrimination in other markets – such as housing or health care – depending on what kind of information is included in the ads or postings used in the market. With relatively few changes to our methods, researchers could test for relationships between the usage of stereotyped language and the discrimination these studies measure. Moreover, language processing techniques may be useful in studying discrimination in different parts of the process of hiring or other employment decisions, such as recommendation letters or employee evaluations.[50]

In our context of age discrimination, the evidence suggests that ageist stereotypes in job ads are related to employers' decisions not to call back older applicants. For both men and women, and across different occupations, we find evidence that employers who do not call back older applicants but do call back younger applicants use phrases in their job ads that are related to ageist stereotypes.

For men, the stereotypes that matter depend on the age and occupation of the applicant. Stereotyped language related to an older man's physical ability predicts age discrimination against older workers applying to be janitors (applicants aged 64-66 and those aged 49-51) and against male applicants aged 64-66 applying for sales positions.[51] Language related to stereotypes about an older worker's personality (careful)

[49] Another way to think about this is that elastic net algorithm uses cross-validation to determine the optimal value of the elastic net parameters. By repeatedly splitting the sample in half and testing how predictive the model is, the elastic net ensures that the results are not driven by the partitioning of the data. Therefore, if we assign a random placebo treatment that is uncorrelated with the words used in an ad, we would expect the elastic net to return zero predictive trigrams if the number of cross-validations is large enough. Thus, these results suggest that five iterations of the cross-validation are enough to prevent the elastic net from selecting trigrams with no true predictive power in most cases.
[50] For a discussion of research on letters of recommendation, see Madera et al. (2009).
[51] Here, we refer to Table 13.

predicts more discrimination against male applicants aged 64-66 applying for janitor or sales jobs. Stereotypes about an older applicant's skills predict discrimination against male applicants aged 64-66 for janitor positions (communication skills), and male applicants aged 49-51 applying for retail sales jobs (technological skills). For women, we find that stereotypes about an older applicant's skills predict discrimination against females aged 64-66 applying for sales jobs (communication skills) and female applicants aged 49-51 applying to be administrative assistants (communication skills and technological skills).

Importantly, we find virtually no evidence that positive stereotypes of older workers are correlated with less hiring discrimination. The results are much more suggestive that when phrasing related to positive stereotypes is present, there is either no change in discrimination or an *increase* in discrimination against older workers. This suggests that surveys of employers may overstate how positively employers view older workers and that these surveys do not reflect actual hiring behavior.

Our findings provide a much more nuanced view of the kind of evidence we get just from comparing callback rates in correspondence studies. The evidence from the job ads suggests that discrimination against older workers occurs for different reasons in different occupations. It may even be different for older workers in different age ranges. Therefore, the policy responses to the age discrimination in hiring documented in NBB and other resume-correspondence studies need to be more nuanced. For example, if older workers are aware of the relationship between ageist language in job ads and hiring discrimination, they may alter their job search behavior, complicating efforts to prosecute age discrimination by policymakers. More work is needed to understand the effect of ageist stereotypes on older workers, especially at points further along in the hiring process.

The evidence provided in this paper has important implications for policy. Our results can provide guidance to the Equal Employment Opportunity Commission and state agencies that enforce age discrimination laws. If employers use ageist language to discourage older workers from applying to jobs, then applicant pools may be shaped to make age discrimination in hiring harder to detect. Barring such language may reduce employer efforts to shape the applicant pool, and testing for age stereotypes in job ads

could be used to detect firms that may discriminate based on age in hiring decisions.[52] And of course, the methodology we develop could be applied to evidence on discrimination against other groups.

One limitation of our work is that we can only learn about the role of age stereotypes that appear in the job ads studied. This could imply that there are stereotypes employers have about older workers that affect hiring, but on which our evidence is silent. On the other hand, thinking back to our two key hypotheses, we may well be most interested in the stereotypes expressed in job ads. Certainly, if age-related stereotypes in job ads are being used to shape the applicant pools, it is the stereotypes in job ads that are of interest. And if age-related stereotypes in job ads signal the dimensions along with employers statistically discriminate in hiring, then these are the stereotypes that need to be assessed against the RFOA criterion. Moreover, if some stereotypes are identified in the lab, but not expressed in real-world job ads, they may simply not be very relevant to real-world labor market decisions.

[52] Shaping the applicant pool can help employers ward off claims of discrimination in hiring. In legal cases, the most compelling data on hiring discrimination comes from comparing hiring rates of the group in question (older workers, in our case) relative to the applicant pool. In the absence of data on applicants, the analysis of a firm's workforce relative to the age structure of the relevant workforce in the population is sometimes used, but such analyses pose a greater challenge to establishing evidence consistent with age discrimination.

References

AARP. 2000. *American Business and Older Employees*. AARP: Washington, DC.

Adafre, Sisa F., and Maarten de Rijke. 2006. "Finding Similar Sentences Across Multiple Languages in Wikipedia." In *Proceedings of the EACL Workshop on New Text*. Trento, Italy

Ameri, Mason, Sean Edmund Rogers, Lisa Schur, and Douglas Kruse. "No Room at the Inn? Disability Access in the New Sharing Economy." Forthcoming in *Academy of Management Discoveries.*

Armstrong-Stassen, Marjorie, and Francine Schlosser. 2008. "Benefits of a Supportive Development Climate for Older Workers." *Journal of Managerial Psychology* 23(4): 419–437.

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

Barnett, Julie. 1998. "Sensitive Questions and Response Effects: An Evaluation." *Journal of Managerial Psychology* 13(1/2): 63–76.

Bendick, Marc, Jr., Lauren E. Brown, and Kennington Wall. 1999. "No Foot in the Door: An Experimental Study of Employment Discrimination Against Older Workers." *Journal of Aging & Social Policy* 10(4): 5–23.

Bendick, Marc, Jr., Charles W. Jackson, and J. Horacio Romero. 1997. "Employment Discrimination Against Older Workers: An Experimental Study of Hiring Practices." *Journal of Aging & Social Policy* 8(4): 25–46.

Breiman, Leo. 1996. "Heuristics of Instability and Stabilization in Model Selection." *The Annals of Statistics* 24(6): 2350–2383.

Bühlmann, Peter, and Sara Van De Geer. 2011. "Statistics for High-Dimensional Data." *Springer Series in Statistics*. Berlin: Springer.

Button, Patrick. 2019. "Population Aging, Age Discrimination, and Age Discrimination Protections at the 50th Anniversary of the Age Discrimination in Employment Act." In S. Czaja, J. Sharit, and J. James (Eds.), *Current and Emerging Trends in Aging and Work*, 163–188. New York, NY: Springer.

Carlsson, Magnus, and Stefan Eriksson. 2019. "The Effect of Age and Gender on Labor Demand – Evidence from a Field Experiment." *Labour Economics* 59: 173–83.

Clark, Stephen. 2014. "Vector Space Models of Lexical Meaning." In Lapin, S., and C. Fox (Eds.), *Handbook of Contemporary Semantics.* Oxford: Blackwell.

Crew, James C. 1984. "Age Stereotypes as a Function of Race." *Academy of Management Journal* 27(2): 431–35.

Dedrick, Esther J., and Gregory H. Dobbins. 1991. "The Influence of Subordinate Age on Managerial Actions: An Attributional Analysis." *Journal of Organizational Behavior* 12(5): 367–77.

Farber, Henry S, Dan Silverman, and Till M. von Wachter. 2017. "Factors Determining Callbacks to Job Applications by the Unemployed: An Audit Study." *RSF: The Russell Sage Foundation Journal of the Social Sciences* 3(3): 168–201.

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

Finkelstein, Lisa M., Kelly D. Higgins, and Maggie Clancy. 2000. "Justifications for Ratings of Older and Young Job Applicants: An Exploratory Content Analysis." *Experimental Aging Research* 26(3):

263–83.

Finkelstein, Lisa M., Michael J. Burke, and Nanbury S. Raju. 1995. “Age Discrimination in Simulated Employment Contexts: An Integrative Analysis.” *Journal of Applied Psychology* 80(6): 652–63.

Finkelstein, Lisa M., and Michael J. Burke. 1998. “Age Stereotyping at Work: The Role of Rater and Contextual Factors on Evaluations of Job Applicants.” *Journal of General Psychology* 125(4): 317–45.

Finkelstein, Lisa M., Katherine M. Ryan, and Eden B. King. 2013. “What Do the Young (Old) People Think of Me? Content and Accuracy of Age-Based Metastereotypes.” *European Journal of Work and Organizational Psychology* 22(6): 633–57.

Fiske, Susan T., Amy J.C. Cuddy, Peter Glick, and Jun Xu. 2002. “A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow from Perceived Status and Competition.” *Journal of Personality and Social Psychology* 82(6): 878–902.

Fix, Michael, and Raymond J. Struyk, Eds. 1993. *Clear and Convincing Evidence: Measurement of Discrimination in America*. Washington, D.C.: The Urban Institute Press.

Gaddis, S. Michael. 2018. “An Introduction to Audit Studies in the Social Sciences.” In Gaddis, S. M. (Ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. New York: Springer.

Gordon, Randall A., and Richard D. Arvey. 2004. “Age Bias in Laboratory and Field Settings: A Meta-Analytic Investigation.” *Journal of Applied Social Psychology* 34(3): 468–92.

Krumpal, Ivar. 2013. “Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review.” *Quality and Quantity* 47(4): 2025–47.

Hanson, Andrew, Zachary Hawley, and Aryn Taylor. 2011. “Subtle Discrimination in the Rental Housing Market: Evidence from e-mail Correspondence with Landlords.” *Journal of Housing Economics* 20(4): 276–84.

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

Hellester, Miguel D., Peter Kuhn, and Kailing Shen. 2014. “Employers’ Age and Gender Preferences: Direct Evidence from Four Job Boards.” Unpublished paper.

Hendrick, Jennifer J., V. Jane Knox, William L. Gekoski, and Kate J. Dyne. 1988. “Perceived Cognitive Ability of Young and Old Targets.” *Canadian Journal on Aging* 7(3): 192–203.

Hummert, Mary Lee, Teri A. Garstka, Jaye L. Shaner, and Sharon Strahm, S. 1994. “Stereotypes of the Elderly Held by Young, Middle-aged, and Elderly Adults.” *Journal of Gerontology* 49(5): P240–9.

Hummert, Mary Lee, Teri A. Garstka, and Jaye L. Shaner. 1995. “Beliefs About Language Performance: Adults’ Perceptions About Self and Elderly Targets.” *Journal of Language and Social Psychology* 14(3): 235–59.

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

Jurafsky Daniel, and James H. Martin. 2017. “Vector Semantics.” In *Speech and Language Processing, Third Edition* (draft), https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Karpinska, Kasia, Kène Henkens, and Joop Schippers. 2013. “Retention of Older Workers: Impact of Managers’ Age Norms and Stereotypes.” *European Sociological Review* 29(6): 1323–35.

Kite, Mary E., Kay Deaux, and Margaret Miele. 1991. “Stereotypes of Young and Old: Does Age Outweigh Gender?” *Psychology and Aging* 6(1): 19–27.

Krings, Franciska, Sabine Sczesny, and Annette Kluge. 2011. "Stereotypical Inferences as Mediators of Age Discrimination: The Role of Competence and Warmth." *British Journal of Management* 22(2): 187–201.

Kroon, Anne C., Martine Van Selm, Claartje L. ter Hoeven, and Rens Vliegenthart. 2016. "Reliable and Unproductive? Stereotypes of Older Employees in Corporate and News Media." *Ageing and Society* 38(1): 166–91.

Kugelmass, Heather. "Just the Type with Whom I Like to Work: Two Correspondence Field Experiments in an Online Mental Health Care Market." Forthcoming in *Society and Mental Health.*

Kuhn, Peter, and Kailing Shen. 2013. "Gender Discrimination in Job Ads: Evidence from China." *Quarterly Journal of Economics* 128(1): 287–336.

Lahey, Joanna. 2008. "Age, Women, and Hiring: An Experimental Study." *Journal of Human Resources* 43(1): 30–56.

Lawrence, Barbara S. 1988. "New Wrinkles in the Theory of Age: Demography, Norms, and Performance Rating." *Academy of Management Journal* 31(2): 309–37.

Levin, William C. 1988. "Age Stereotyping: College Student Evaluations." *Research on Aging* 10(1): 134–48.

Maurer, Todd J., Frank G. Barbeite, Elizabeth M. Weiss, and Micheal Lippstreu. 2008. "New Measures of Stereotypical Beliefs about Older Workers' Ability and Desire for Development: Exploration among Employees Age 40 and Over." *Journal of Managerial Psychology* 23(4): 395–418.

McCann, Robert M., and Shaughan A. Keaton. 2013. "A Cross Cultural Investigation of Age Stereotypes and Communication Perceptions of Older and Younger Workers in the USA and Thailand." *Educational Gerontology* 39(5): 326-41.

McGregor, Judy, and Lance Gray. 2002. "Stereotypes and Older Workers: The New Zealand Experience." *Social Policy Journal of New Zealand* 18: 163–77.

Madera, Juan M., Michelle R. Hebl, and Randi C. Martin. 2009. "Gender and Letters of Recommendation for Academia: Agentic and Communal Differences." *Journal of Applied Psychology* 94(6): 1591–99.

Maestas, Nicole. 2010. "Back to Work: Expectations and Realizations of Work after Retirement." *Journal of Human Resources* 45(3): 718–48.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. "Efficient Estimation of Word Representations in Vector Space." Unpublished paper, ICLR Workshop.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. "Distributed Representations of Words and Phrases and their Compositionality." In *Advances in Neural Information Processing Systems* 26: 3111–19.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013c. "Linguistic Regularities in Continuous Space Word Representations." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*: 746–51.

Neumark, David. 2018. "Experimental Research on Labor Market Discrimination." *Journal of Economic Literature* 56(3): 799-866.

Neumark, David, Ian Burn, and Patrick Button. 2019. "Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment." *Journal of Political Economy* 127(2): 922–70.

Neumark, David, Ian Burn, and Patrick Button. 2017. "Age Discrimination and Hiring of Older Workers." *Federal Reserve Board of San Francisco Economic Letter* #2017-06.

Neumark, David, Ian Burn, and Patrick Button. 2016. "Experimental Age Discrimination Evidence and the

Heckman Critique." *American Economic Review Papers & Proceedings* 106(5): 303–08.

Neumark, David, Ian Burn, Patrick Button, and Nanneh Chehras. 2019. "Do State Laws Protecting Older Workers from Discrimination Reduce Age Discrimination in Hiring? Evidence from a Field Experiment." *Journal of Law and Economics* 62(2): 373–402

Newson, Roger, and The ALSPAC Study Team. 2003. "Multiple-Test Procedures and Smile Plots." *The Stata Journal* 3(2): 109–32.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pitt-Catsouphes, Marcie, Micheal A. Smyer, Christina Matz-Costa, and Katherine Kane. 2007. "The National Study Report: Phase II of the National Study of Business Strategy and Workforce Development." *The Center on Aging & Work at Boston College*. http://www.bc.edu/content/dam/files/research_sites/agingandwork/pdf/publications/RH04_NationalStudy.pdf.

Posthuma, Richard A., and Michael A. Campion. 2007. "Age Stereotypes in the Workplace: Common Stereotypes, Moderators, and Future Research Directions." *Journal of Management* 35(1): 158–88.

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

Ryan, Ellen Bouchard. 1992. "Beliefs About Memory Changes Across the Adult Life Span." *Journal of Gerontology: Psychological Sciences* 47(1): 41–6.

Ryan, Ellen Bouchard, Sheree Kwong See, W. Bryan Meneer, and Diane Trovato. 1992. "Age-Based Perceptions of Language Performance Among Younger and Older Adults." *Communication Research* 19(4): 423–43.

Ryan, E. B., and Sheree Kwong See. 1993. "Age-Based Beliefs about Memory Changes for Self and Others across Adulthood." *Journals of Gerontology* 48(4): 199–201.

Schmidt, Daniel F., and Susan M. Boland. 1986. "Structure of Perceptions of Older Adults: Evidence for Multiple Stereotypes." *Psychology and Aging* 1(3): 255–60.

Singer, M. S. 1986. "Age Stereotypes as a Function of Profession." *Journal of Social Psychology* 126(5): 691–92.

Stewart, Mark A., and Ellen Bouchard Ryan. 1982. "Attitudes toward Younger and Older Adult Speakers: Effects of Varying Speech Rates." *Journal of Language and Social Psychology* 1(2): 91–109.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B* 58(1): 267–288.

Tilcsik, András. 2011. "Pride and Prejudice: Employment Discrimination against Openly Gay Men in the United States." *American Journal of Sociology* 117(2): 586-626.

Truxillo, Donald M., Elizabeth A. McCune, Marilena Bertolino, and Franco Fraccaroli. 2012. "Perceptions of Older Versus Younger Workers in Terms of Big Five Facets, Proactive Personality, Cognitive Ability, and Job Performance." *Journal of Applied Social Psychology* 42(11): 2607–26.

van Dalen, Harry, Kene Henkens, and J.J Schippers. 2009. "Dealing with Older Workers in Europe: A Comparative Survey of Employers' Attitudes and Actions." *Journal of European Social Policy* 19(1): 47–60.

Warr, Peter, and Janet Pennington. 1993. "Views about Age Discrimination and Older Workers." In P.

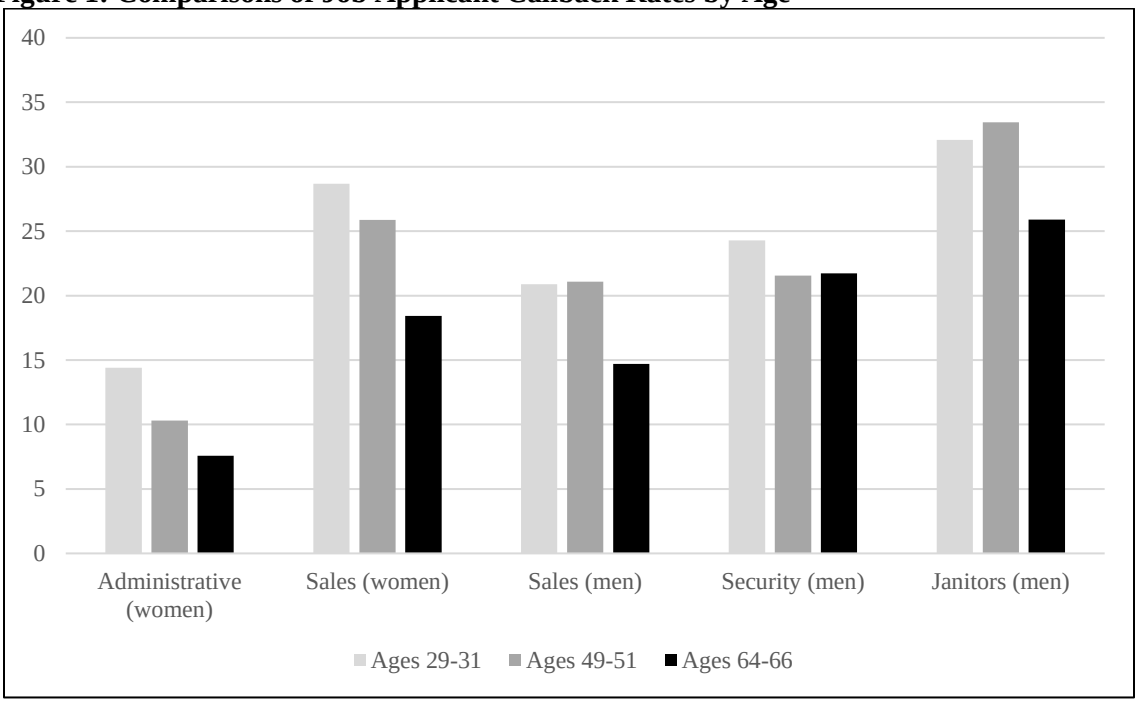
Taylor, et al. (Eds.), *Age and Employment: Policies, Attitudes, and Practice*, 75–106. London: Institute of Personnel Management.

Weiss, Elizabeth M, and Todd J. Maurer. 2004. "Age Discrimination in Personnel Decisions: A Reexamination." *Journal of Applied Social Psychology* 34(8): 1551–62.

Zhang, Tong. 2004. "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization." *The Annals of Statistics* 32(1): 469–75

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society Series B* 67(2): 301–20.

Figure 1: Comparisons of Job Applicant Callback Rates by Age



Note: A callback is defined as a positive invitation to come in for an interview. Figure is reproduced from Neumark, Burn, and Button (2017) using data from NBB.

Figure 2: Example Job Advertisement

◀ prev ▲ next ▶

Boutique Sales Associate (Larchmont Village in Hancock Park)

Beverly Blvd at Larchmont Blvd

(google map) (yahoo map)

compensation: **Commensurate with experience or performance.**

Larchmont Village/Hancock Park European-style boutique seeks a full-time sales person. Required: experience in retails sales. Above all, this is a sales position. Must have excellent social skills. A plus: experience with visual merchandising, administrative tasks, and social networking, including Facebook. Please email résumé in a PDF or Word document or in the email itself.

- Principals only. Recruiters, please don't contact this job poster.
- do NOT contact us with unsolicited services or offers

post id: 4959905041

posted: 2015-04-01 6:26pm

updated: 2015-04-01 6:26pm

Note: An example of a saved job advertisement from Neumark, Burn, and Button (2019). Each ad contained a "post id," which we use to match to the resumes that were sent out. The text of the ad, including both the title and the body of the ad, were scraped to identify all words used.

Figure 3: Visual Representation of a Hypothetical *Word2Vec* Neural Network

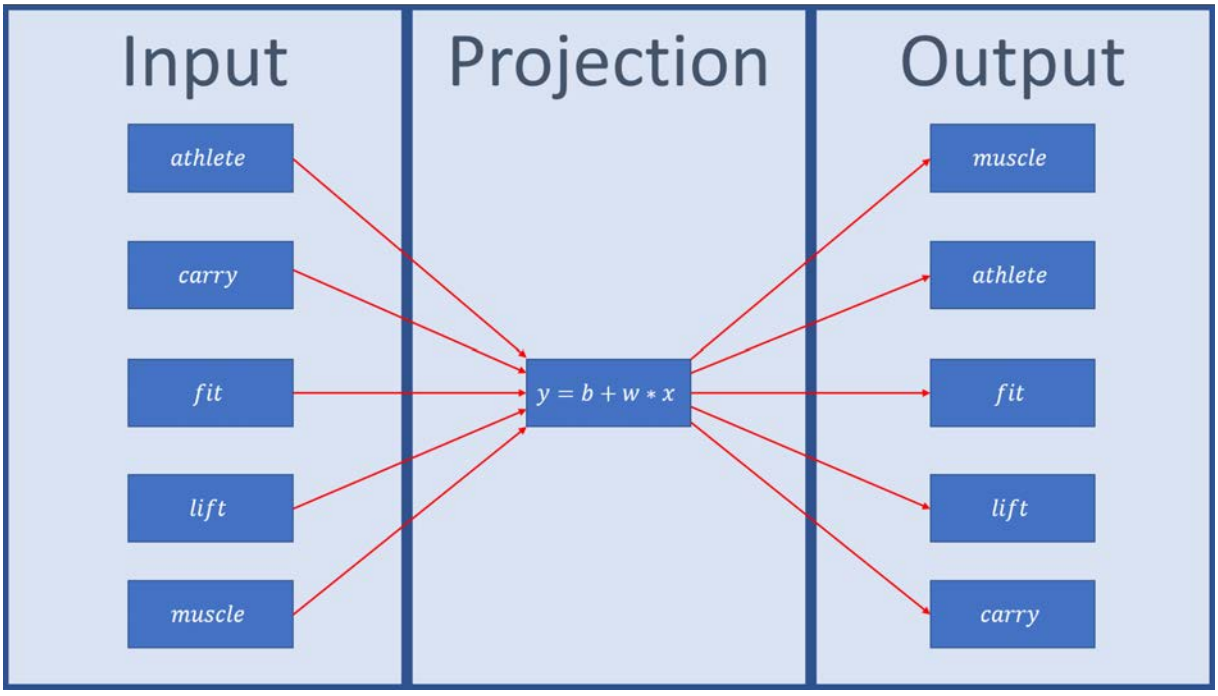
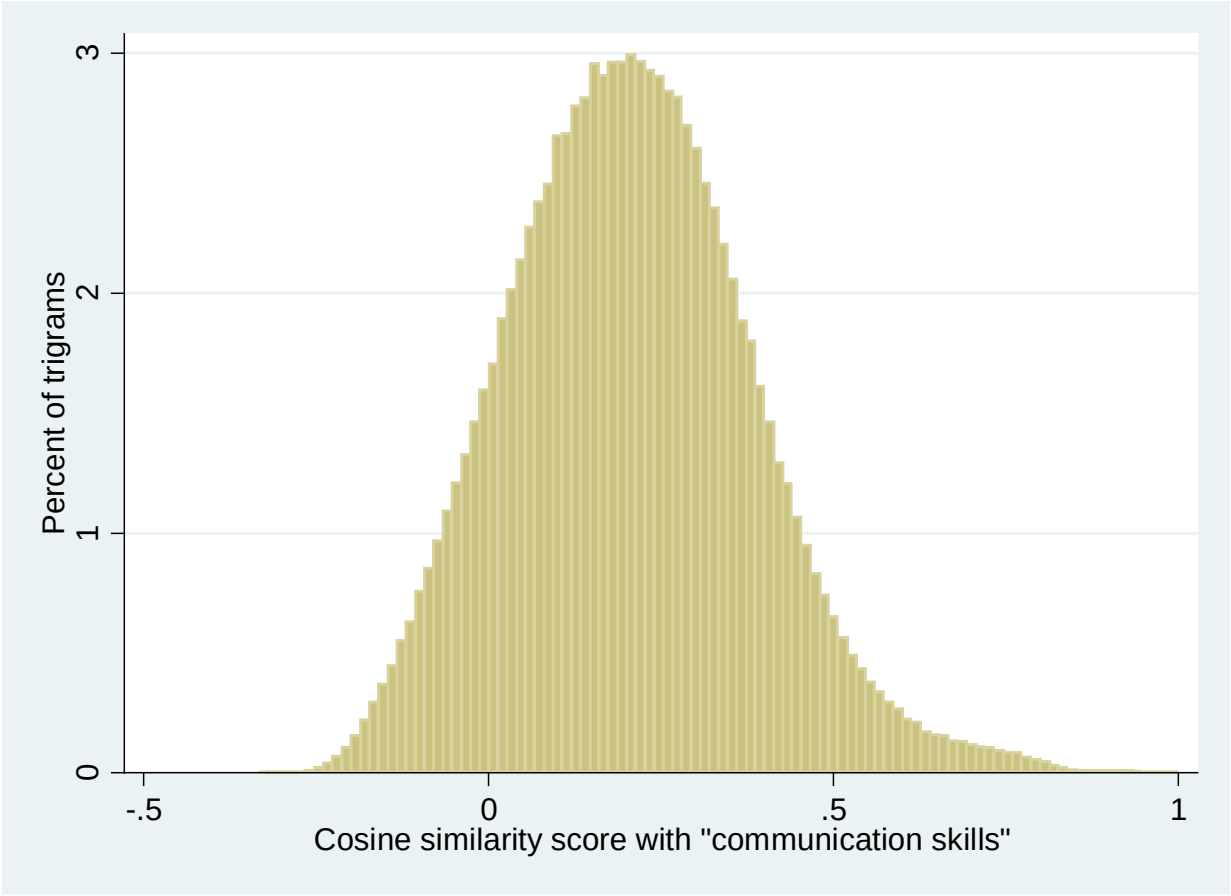
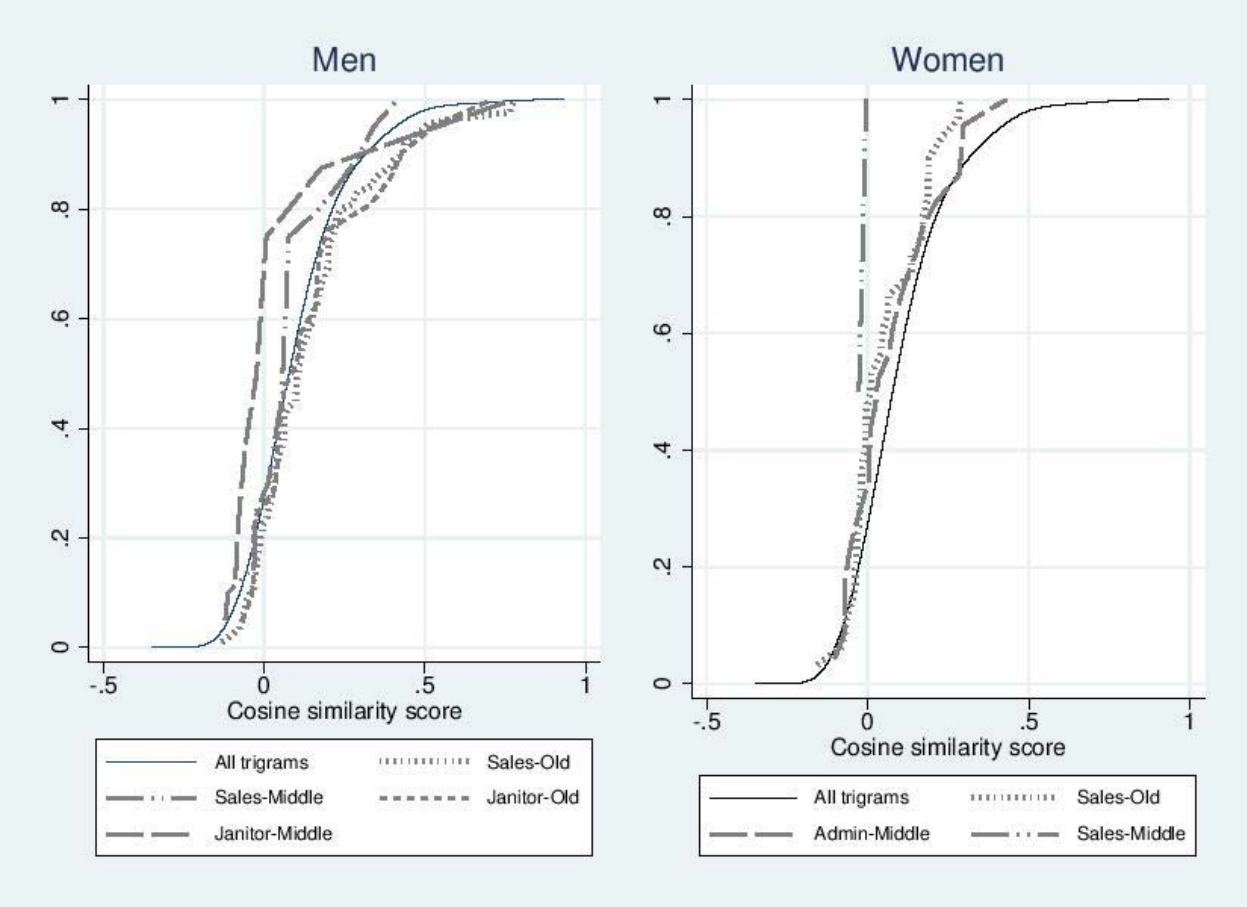


Figure 4: Example of the Distribution of Cosine Similarity (CS) Scores



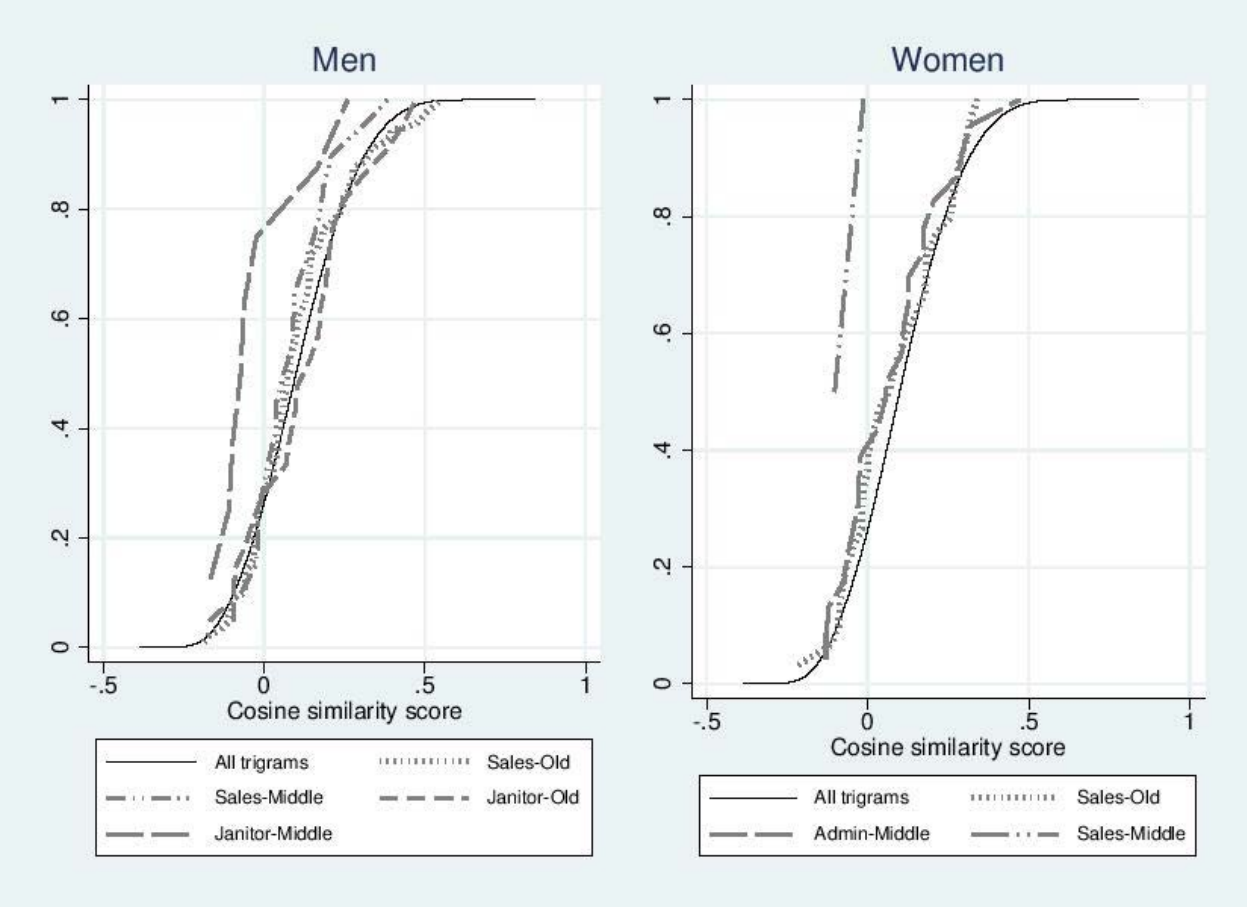
Note: Figure reports the distribution of cosine similarity scores for all trigrams from the job ads. The higher the cosine similarity score, the more related the trigram is to “communication skills.”

Figure 5a: Distribution of Cosine Similarity Scores of Elastic-Net Selected Trigrams Relative to All Trigrams – Less Physically Able



Note: The continuous line represents the CDF of semantic similarity scores among the approximately 1 million trigrams. The remaining lines are the CDFs of the trigrams selected by the elastic net algorithm. If the CDF of an elastic net lies to the right of the CDF of all trigrams, it means that the trigrams that the elastic net determined predict discrimination are more related to the stereotype than the average trigram.

Figure 5b: Distribution of Cosine Similarity Scores of Elastic-Net Selected Trigrams Relative to All Trigrams – More Careful



Note: See Figure 5a for description.

Table 1: Stereotypes about Older Workers' Health

Aggregate Stereotype	Phrasing	Source
Less Attractive	"wrinkled," "unattractive," "not neat"	Kite et al. (1991)
	"less attractive"	Levin (1988)
	"worse-looking when older"	Zepelin, Sills, and Heath (1987)
Hard of Hearing	"hard of hearing"	Kite et al (1991)
	"worse hearing," "think people speak too softly,"	Ryan et al. (1992)
	"frustrated when not hearing," "think other people speak too fast," "often ask others to repeat"	
	"worse hearing"	Hummert, Gartska, and Shaner (1995)
Worse Memory	"Worse memory"	Hendrick et al. (1988)
	"Worse memory"	Ryan (1992)
	"Worse memory"	Ryan and Kwong See (1993)
	"Worse memory"	Hummert, Gartska, and Shaner (1995)
Less Physically Able	"lower physical capacity"	Kroon et al. (2016) (p. 16)
	"[worse] physical capability and health"	van Dalen, Henkens, and Schippers (2009) (p. 21)
	"sedentary," "physically handicapped," "slow moving," "sick," "shaky hands," "fragile," "poor posture"	Schmidt and Boland (1986)
	"less qualified for a physically demanding job"	Finkelstein, Burke, and Raju (1995)
	"tired," "scared of becoming sick or incompetent"	Hummert et al. (1994)
	"[lower] activity," "[less] energy," "[worse] health," "[less] speed"	Levin (1988) (p. 142)
	"less physically active," "unhealthy," "moves slowly"	Kite et al. (1991)
	"worse psychomotor speed"	Hendrick et al. (1988)

Table 2: Stereotypes about Older Workers' Personality

Aggregate Stereotype	Phrasing	Source
Less Adaptable	"[less] flexible in doing different tasks," "[less likely to] try new approaches"	AARP (2000) (p. 6)
	"occupationally flexible"	Karpinska et al. (2013)
	"[more] flexibility"	Levin (1988) (p. 142)
	"[less likely to] adapt to change," "[less likely to] grasp new ideas"	Lyon and Pollard (1997) (p. 252)
	"older workers are less flexible than younger workers."	McCann and Keaton (2013)
	"resistant to change"	McGregor and Gray (2002)
	"find difficult to change," "old-fashioned"	Schmidt and Boland (1986)
	"adapt less well to change," "are less able to grasp new ideas"	Warr and Pennington (1993) (p. 89)
	"resistant to change"	Weiss and Maurer (2004)
	"talks of past," "focuses away from future toward past"	Kite et al. (1991)
	"less flexible," "more old-fashioned"	Stewart and Ryan (1982)
Careful	"think before they act"	Lyon and Pollard (1997) (p. 251)
	"older workers are more cautious than younger workers."	McCann and Keaton (2013)
	"cautiousness," "self-discipline"	Truxillo et al. (2012) (p. 2623)
	"think before they act"	Warr and Pennington (1993) (p. 89)
	"better practical judgment," "better common sense"	Hendrick et al. (1988)
Less Creative	"[lower] creativity"	Levin (1988) (p. 142)
	"[lower] creativity"	van Dalen, Henkens, and Schippers (2009) (p. 21)
Dependable	"loyal"	AARP (2000) (p. 6)
	"[more] stability"	Crew (1984) (p.433)
	"more reliable," "committed to the organization"	van Dalen, Henkens, and Schippers (2009) (p. 21)
	"stable"	Finkelstein, Burke, and Raju (1995)
	"trustworthy," "reliability," "commitment"	Kroon et al. (2016) (p. 16)
	"are loyal to the organization"	Lyon and Pollard (1997) (p. 251)
	"reliability," "loyalty," "job commitment"	McGregor and Gray (2002)
	"loyal to the company," "are reliable"	Pitt-Catsouphes et al. (2007) (p. 8)
	"more loyal to the organization" "more reliable"	Warr and Pennington (1993) (p. 89)
	"more stable"	Singer (1986)
	"more trustworthy"	Stewart and Ryan (1982)
Negative Personality	"dejected," "poor," "hopeless," "unhappy," "lonely," "insecure," "complains a lot," "grouchy," "critical," "miserly"	Kite et al. (1991)
	"[less] pleasantness"	Levin (1988) (p. 143)
	"ill-tempered," "bitter," "demanding," "complaining," "annoying," "humorless," "selfish," "prejudiced," "suspicious of strangers," "easily upset," "miserly," "snobbish"	Schmidt and Boland (1986)
	"[less] friendliness," "[less] cheerfulness"	Truxillo et al. (2012) (p. 2623)
Warm Personality	"warm," "good-natured," "benevolent," "amicable"	Krings, Sczesney, and Kluge (2010)
	"Warm personality"	Kroon et al. (2016) (p. 16)
	"more conscientious"	Warr and Pennington (1993) (p. 89)
	"warm"	Fiske et al. (2002)

Table 3: Stereotypes about Older Workers' Skills

Aggregate Stereotype	Phrasing	Source
Lower Ability to Learn	"will [not] participate in training programs"	AARP (2000) (p. 6)
	"learn new techniques" "personal development"	Armstrong-Stassen and Schlosser (2008)
	"[less] potential for development"	Crew (1984) (p.433)
	"lack willingness to be trained"	van Dalen, Henkens, and Schippers (2009) (p. 21)
	"training more appropriate for younger workers"	Dedrick and Dobbins (1991) (p. 373)
	"[less] ability and willingness to learn"	Kroon et al. (2016) (p. 16)
	"[less likely to] want to be trained"	Lyon and Pollard (1997) (p. 252)
	"Less interest in learning."	Maurer at al. (2008)
	"learn less quickly," "are less interested in being trained"	Warr and Pennington (1993) (p. 89)
	"less potential for development"	Finkelstein, Burke, and Raju (1995)
	"lower potential for development"	Singer (1986)
Better Communication Skills	"[better] interpersonal skills"	Crew (1984) (p.433)
	"better social skills"	van Dalen, Henkens, and Schippers (2009) (p. 21)
	"more interpersonally skilled"	Kroon et al. (2016) (p. 16)
	"sincere when talking," "tells more enjoyable stories"	Ryan et al. (1992)
Worse Communication Skills	"less interpersonally skilled"	Finkelstein and Burke (1998) (p. 331)
	"unable to communicate"	Schmidt and Boland (1986)
	"worse interpersonal skills"	Singer (1986)
	"talks slowly," "less sociable," "has few friends"	Kite, Deaux, and Meile (1991)
	"worse conversational skills," "hard to understand when noisy," "lose track of who said what," "lose track of topic," "lose track of what talked about," "hard to speak if pressed for time," "use fewer difficult words," "recognize meanings of fewer words"	Ryan et al. (1992)
	"less outgoing," "quieter voice," "more hoarse"	Stewart and Ryan (1982)
More Experienced	"solid experience"	AARP (2000) (p. 6)
	"[more] experience"	Finkelstein, Higgins, and Clancy (2000)
	"[more] experience"	Finkelstein, Ryan, and King (2013)
	"have useful experience"	Lyon and Pollard (1997) (p. 251)
	"having more experience which is useful in the job"	Warr and Pennington (1993) (p. 89)
More Productive	"strong work ethic"	Pitt-Catsouphes et al. (2007) (p. 8)
	"working harder"	Warr and Pennington (1993) (p. 89)
Less Productive	"[lower] performance capacity"	Crew (1984) (p.433)
	"attributed low performance more to the stable factor of lack of ability when the subordinate was old"	Dedrick and Dobbins (1991) (p. 368)
	"less economically beneficial"	Finkelstein and Burke (1998) (p. 331)
	"high performance rating is positively related with youth"	Lawrence (1988) (p. 328)
	"[less] competence"	Levin (1988) (p. 142)
	"younger workers are seen as having higher performance capacity"	Singer (1986) (p. 691)
Worse with Technology	"[less likely to] understand new technologies" "[less likely to] learn new technologies," "[less] comfortable with new technologies"	AARP (2000) (p. 6)
	"lack capacity to deal with new technologies"	van Dalen, Henkens, and Schippers (2009) (p. 21)
	"[less] technological competence" "[less] technological adaptability"	Kroon et al. (2016) (p. 16)
	"[less likely to] accept new technology"	Lyon and Pollard (1997) (p. 252)
	"Older workers adapt to new technology slower than younger workers." "Younger workers are less fearful of technology than older workers.	McCann and Keaton (2013)
	"problems with technology"	McGregor and Gray (2002)
	"less readily accept the introduction of new technology"	Warr and Pennington (1993) (p. 89)

Table 4: Evaluating the Elastic Net Algorithms

Age	Gender	Occupation	Observations	Selected trigrams	λ	α
Old	Male	Janitor	329	21	0.014	0.5
		Sales	1,680	89	0.004	0.5
		Security	932	1	0.022	0.5
	Female	Admin	7,330	1	0.004	0.5
		Sales	1,861	30	0.005	0.5
Middle	Male	Janitor	331	8	0.018	0.5
		Sales	1,612	20	0.006	0.5
		Security	956	0	0.035	0.5
	Female	Admin	6,827	23	0.002	0.5
		Sales	987	2	0.013	0.5

Notes: λ is the penalization parameter, which is estimated. α sets the weights on ridge vs. LASSO regularization and is set to 0.5. Elastic net models include the same controls for resume characteristics used in NBB.

Table 5: Distributions of Selected Job-Ad Phrases in Relation to Age Stereotypes

Stereotype	Old-Male-Janitor	Old-Male-Sales	Old-Female-Sales	Middle-Male-Janitor	Middle-Male-Sales	Middle-Female-Admin
Health						
Less Attractive						
Hard of Hearing						
Worse Memory						Top 10%: 21.7%* (p=0.073)
Less Physically Able	Top 10%: 23.8%* (p=0.052) Top 5%: 14.3%* (p=0.086)	Top 10%: 16.9%** (p=0.049) Top 5%: 11.2%** (p=0.013) Top 1%: 4.5%** (p=0.012)		Top 1%: 12.5%* (p=0.077)		
Personality						
Less Adaptable	Top 10%: 23.8%* (p=0.052)					
Careful	Top 5%: 14.3%* (p=0.084)	Top 1%: 4.5%** (p=0.012)				
Less Creative						
Dependable	Top 10%: 23.8%* (p=0.052) Top 5%: 19.1%** (p=0.018) Top 1%: 9.5%** (p=0.019)	Top 10% 16.7%** (p=0.048) Top 5%: 10.1%** (p=0.044)				Top 10%: 21.74%* (p=0.089) Top 5%: 17.39%** (p=0.025)
Negative/Warm Personality		Top 1%: 3.37%* (p=0.60)				
Skills						
Lower Ability to Learn						
Worse/Better Communication Skills	Top 5%: 14.3%* (p=0.085) Top 1%: 9.5%** (p=0.019)		Top 1%: 6.7%** (p=0.036)			Top 10%: 21.7%* (p=0.073) Top 1%: 8,7%** (p=0.022)
More Experienced					Top 10%: 35.0%*** (p=0.002) Top 5%: 20.0%** (p=0.016)	
Less/More Productive	Top 5%: 14.3%* (p=0.084)					
Worse with Technology					Top 10%: 30.0%*** (p=0.011)	Top 10%: 21.7%* (p=0.073)

Note: For results significant at the 10% level or higher, we report the cutoff (top 10%, 5%, or 1% of semantic similarity scores for all trigrams) for the age stereotypes indicated in the rows, the share of trigrams above this threshold selected by the elastic net, and the p-value of the binomial test of probability testing whether there is a significant difference between the share selected and what would be expected if we randomly drew trigrams. Blank cells indicate no significant difference between the elastic net trigrams and a random draw from the distribution of all trigrams. Also not shown are the results for age-gender-occupation triplets for which only

two or fewer trigrams were selected (Old-Male-Security, Old-Female-Admin, Middle-Male-Security, and Middle-Female-Sales), for which the tests of over-representation of selected trigrams are uninformative.
* $p<0.1$, ** $p<0.05$, *** $p<0.01$

Table 6a: Significance of the Difference Between Elastic-Net Selected Trigrams and All Trigrams – Less Physically Able

	Old-Male-Janitor	Old-Male-Sales	Old-Female-Sales	Middle-Male-Janitor	Middle-Male-Sales	Middle-Female-Admin
Total elastic net selected trigrams	21	89	30	8	20	23
Proportion of elastic net selected trigrams in the 91st centile or above	23.81%	16.85%	0.00%	12.50%	10.00%	4.35%
Proportion of elastic net selected trigrams in the 96th centile or above	14.29%	11.24%	0.00%	12.50%	5.00%	4.35%
Proportion of elastic net selected trigrams in the 100th centile or above	4.76%	4.49%	0.00%	12.50%	0.00%	0.00%
Total not selected trigrams	210,651	210,583	210,642	210,664	210,652	210,649
Proportion of not selected trigrams in the 91st centile or above (top 10%)	10.00%	10.00%	10.00%	10.00%	10.00%	10.00%
Proportion of not selected trigrams in the 96th centile or above (top 5%)	5.00%	5.00%	5.00%	5.00%	5.00%	5.00%
Proportion of not selected trigrams in the 100th centile or above (top 1%)	1.00%	1.00%	1.00%	1.00%	1.00%	1.00%
Hypothesis testing: p-value						
H_0: proportion of selected trigrams in the 91st centile or above=10% H_A: proportion of selected trigrams in the 91st centile or above ≠ 10%	0.052	0.049	1.000	0.569	1.000	1.000
H_0: proportion of selected trigrams in the 96th centile or above=5% H_A: proportion of selected trigrams in the 96th centile or above ≠ 5%	0.086	0.013	1.000	0.336	1.000	1.000
H_0: proportion of selected trigrams in the 100th percentile=1% H_A: proportion of selected trigrams above the 100th percentile ≠ 1%	0.190	0.012	1.000	0.077	1.000	1.000

Note: We report the binomial tests of probability to determine if the elastic net is selecting trigrams related to the age stereotypes at a higher rate than we would expect from a random draw of trigrams. We test how over-represented trigrams are in the top 10%, the top 5%, and the top 1%.

Table 6b: Significance of the Difference Between Elastic-Net Selected Trigrams and All Trigrams – Careful

	Old-Male-Janitor	**Old-Male-Sales**	**Old-Female-Sales**	**Middle-Male-Janitor**	**Middle-Male-Sales**	**Middle-Female-Admin**
Total elastic net selected trigrams	21	89	30	8	20	23
Proportion of elastic net selected trigrams in the 91st centile or above	14.29%	12.36%	6.67%	0.00%	5.00%	8.70%
Proportion of elastic net selected trigrams in the 96th centile or above	14.29%	7.87%	0.00%	0.00%	5.00%	4.35%
Proportion of elastic net selected trigrams in the 100th centile or above	0.00%	4.49%	0.00%	0.00%	0.00%	0.00%
Total not selected trigrams	210,651	210,583	210,642	210,664	210,652	210,649
Proportion of not selected trigrams in the 91st centile or above (top 10%)	10.00%	10.00%	10.00%	10.00%	10.00%	10.00%
Proportion of not selected trigrams in the 96th centile or above (top 5%)	5.00%	5.00%	5.00%	5.00%	5.00%	5.00%
Proportion of not selected trigrams in the 100th centile or above (top 1%)	1.00%	1.00%	1.00%	1.00%	1.00%	1.00%
Hypothesis testing: p-value						
H_0: proportion of selected trigrams in the 91st centile or above=10% H_A: proportion of selected trigrams in the 91st centile or above ≠ 10%	0.461	0.477	0.763	1.000	0.714	1.000
H_0: proportion of selected trigrams in the 96th centile or above=5% H_A: proportion of selected trigrams in the 96th centile or above ≠ 5%	0.084	0.216	1.000	1.000	1.000	1.00
H_0: proportion of selected trigrams in the 100th percentile=1% H_A: proportion of selected trigrams above the 100th percentile ≠ 1%	1.000	0.012	1.000	1.000	1.000	1.000

Note: See notes to Table 6a.

Table 7: Coefficients from Elastic Net: Old-Male-Janitor

Trigram	Elastic Net Coefficient	Semantic Similarity Score (Centile of Score)					
		Less Physically Able	Less Adaptable	Careful	Dependable	Worse/Better Communication Skills	Less/More Productive
ability communicate effectively	0.410	0.531 (99)	0.344 (96)	0.301 (89)	0.169 (75)	0.624 (99)	0.212 (85)
criminal background drug	0.144	0.066 (47)	-0.038 (14)	0.193 (72)	0.025 (26)	0.234 (58)	-0.013 (25)
able lift pounds	0.113	0.707 (100)	0.177 (70)	0.240 (81)	0.184 (79)	0.118 (32)	0.103 (59)
pass background check	0.106	0.105 (58)	0.035 (30)	0.212 (76)	0.068 (40)	0.177 (45)	-0.094 (9)
must reliable transportation	0.105	0.430 (96)	0.338 (96)	0.472 (99)	0.418 (100)	0.351 (81)	0.319 (97)
monday friday pm	0.102	-0.034 (19)	-0.143 (2)	-0.095 (10)	-0.105 (2)	-0.049 (7)	-0.229 (1)
good communication skills	0.096	0.333 (92)	0.320 (94)	0.431 (98)	0.404 (99)	0.903 (100)	0.344 (98)
mirrors glass partitions	0.087	0.041 (40)	0.039 (32)	0.079 (46)	-0.097 (3)	0.009 (14)	-0.088 (10)
administrative assistant janitor	0.075	0.030 (37)	0.035 (31)	-0.031 (21)	0.097 (51)	0.113 (31)	0.080 (51)
heritage community initiatives	0.075	-0.070 (12)	0.118 (54)	0.002 (27)	-0.004 (17)	0.277 (67)	0.142 (70)
maintain high level	0.074	0.158 (71)	0.287 (91)	0.141 (61)	0.211 (85)	0.326 (76)	0.204 (84)
cleaning supplies equipment	0.050	0.165 (72)	0.165 (66)	0.169 (67)	0.257 (92)	0.275 (67)	0.103 (59)
pm monday friday	0.044	-0.034 (19)	-0.143 (2)	-0.095 (10)	-0.105 (2)	-0.049 (7)	-0.229 (1)
part time day	0.041	-0.037 (19)	-0.166 (1)	-0.091 (11)	-0.090 (3)	-0.082 (5)	-0.076 (11)
hard working honest	0.041	0.392 (95)	0.304 (93)	0.376 (96)	0.434 (100)	0.303 (72)	0.349 (99)
background check prior	0.035	0.110 (59)	-0.015 (19)	0.201 (73)	0.041 (31)	0.122 (33)	-0.126 (5)
please send resume	-0.001	0.052 (43)	0.056 (36)	0.102 (51)	0.003 (19)	0.032 (17)	-0.074 (12)
hours per week	-0.004	0.042 (40)	-0.096 (5)	-0.056 (16)	-0.019 (14)	0.008 (14)	-0.033 (20)
customer service skills	-0.007	0.171 (74)	0.142 (60)	0.176 (68)	0.309 (96)	0.689 (100)	0.154 (73)
duties include limited	-0.034	-0.030 (20)	0.179 (70)	0.068 (43)	0.063 (39)	0.213 (53)	0.137 (68)
valid drivers license	-0.035	0.196 (78)	0.146 (61)	0.096 (50)	0.073 (42)	0.131 (35)	0.032 (37)

Note: See the notes to Table 4. The first column lists the trigram. The elastic net coefficient is reported next. This is interpreted as the percentage point increase in the discrimination rate for old men in janitor positions. The remaining columns list the stereotypes that had at least one instance of a statistically significant result. In each cell, we report the raw semantic similarity score. Numbers closer to 1 indicate a stronger similarity between the stereotype and the trigram. In parentheses, we report the centile of the semantic similarity. Trigrams in the top 10% of the distribution are shaded. The table reports results for the stereotypes for which at least one selected trigram is statistically significant overrepresented in the top 10%, 5%, or 1% of the distribution of similarity scores between the trigrams and that stereotype (as reported in Table 5).

Table 8: Coefficients from Elastic Net: Old-Male-Sales

		Semantic Similarity Score (Centile of Score)			
Trigram	**Elastic net coefficient**	**Less Physically Able**	**Careful**	**Dependable**	**Negative/ Warm Personality**
term disability insurance	0.237	0.038 (39)	-0.034 (20)	-0.003 (18)	0.095 (68)
shopping experience sales	0.231	0.031 (37)	-0.015 (24)	0.127 (62)	0.189 (85)
must team player	0.223	0.362 (93)	0.131 (58)	0.113 (57)	0.122 (74)
individual must able	0.206	0.779 (100)	0.505 (100)	0.151 (70)	0.070 (62)
diploma ged must	0.205	0.335 (92)	0.316 (91)	0.041 (31)	0.009 (45)
graduate equivalent prior	0.167	0.034 (38)	-0.030 (21)	-0.033 (11)	-0.068 (22)
point sale systems	0.128	0.043 (40)	0.069 (43)	0.057 (36)	-0.094 (16)
submit resume review	0.091	-0.131 (4)	0.140 (60)	-0.029 (12)	-0.016 (37)
greet customers enter	0.089	0.117 (61)	-0.035 (20)	0.030 (27)	-0.132 (9)
interpersonal skills ability	0.079	0.489 (98)	0.370 (95)	0.240 (90)	0.441 (100)
high school diploma	0.076	-0.059 (14)	-0.031 (21)	0.000 (18)	-0.038 (30)
pet care leader	0.071	0.101 (57)	0.027 (33)	0.171 (76)	0.201 (87)
aids healthcare foundation	0.056	-0.04 (18)	-0.025 (22)	0.047 (33)	0.017 (47)
please submit resume	0.055	0.009 (30)	0.183 (70)	-0.035 (10)	-0.088 (17)
must able work	0.052	0.767 (100)	0.547 (100)	0.221 (87)	0.080 (64)
must reliable transportation	0.052	0.430 (96)	0.472 (99)	0.418 (100)	-0.078 (20)
hiring front desk	0.047	0.052 (43)	0.039 (36)	0.110 (56)	-0.009 (39)
school diploma ged	0.047	-0.049 (16)	-0.058 (16)	-0.054 (7)	0.002 (42)
sales associates join	0.035	-0.100 (7)	-0.187 (2)	0.034 (29)	-0.024 (35)
goorin bros right	0.031	0.213 (81)	0.070 (43)	-0.013 (15)	0.060 (59)
responsibilities include limited	0.027	-0.018 (23)	0.099 (50)	0.043 (31)	-0.009 (39)
days per week	0.026	-0.009 (26)	-0.102 (9)	-0.048 (8)	-0.134 (9)
fast paced environment	0.026	0.291 (89)	0.278 (86)	0.342 (98)	0.179 (84)
people achieve health	0.025	0.119 (62)	0.063 (42)	0.044 (32)	0.117 (73)
able lift lbs	0.024	0.773 (100)	0.260 (84)	0.179 (78)	-0.033 (32)
including evenings weekends	0.023	-0.059 (14)	-0.049 (17)	0.013 (22)	-0.044 (29)
cell phone repair	0.022	0.198 (78)	0.104 (52)	0.054 (35)	0.044 (55)
resume salary history	0.018	-0.078 (10)	0.053 (39)	0.080 (45)	-0.004 (41)
must willing work	0.015	0.602 (100)	0.530 (100)	0.233 (89)	0.102 (70)
qualifications bachelor degree	0.014	-0.024 (22)	0.09 (48)	-0.030 (12)	0.133 (76)
different shifts available	0.012	0.281 (88)	0.275 (86)	0.062 (38)	0.018 (47)
great opportunity part	0.012	0.172 (74)	0.141 (61)	0.081 (45)	-0.01 (39)
generous employee discount	0.011	0.076 (50)	0.143 (61)	0.296 (95)	0.107 (71)

		Semantic Similarity Score (Centile of Score)			
Trigram	**Elastic net coefficient**	**Less Physically Able**	**Careful**	**Dependable**	**Negative/ Warm Personality**
seeking part time	0.011	0.199 (79)	0.043 (37)	0.003 (19)	-0.098 (15)
team oriented environment	0.009	0.194 (78)	0.095 (50)	0.174 (76)	0.156 (80)
candidate possess following	0.004	0.060 (45)	-0.060 (16)	0.061 (38)	0.082 (65)
possess following attributes	0.004	0.151 (70)	0.059 (41)	0.057 (36)	0.247 (92)
experience high school	0.002	0.023 (34)	0.046 (37)	0.110 (56)	0.132 (76)
perform essential functions	0.001	0.227 (82)	0.208 (75)	0.120 (59)	0.109 (71)
promoted full time	0.001	0.203 (79)	0.059 (41)	0.070 (41)	-0.054 (26)
reliable transportation outgoing	0.001	0.166 (73)	0.265 (84)	0.569 (100)	-0.020 (36)
priced low per	0.001	0.105 (58)	0.044 (37)	0.211 (85)	-0.110 (13)
requirements must high	0.001	0.400 (95)	0.395 (97)	0.157 (72)	-0.048 (28)
positions strong career	0.001	-0.013 (24)	0.023 (32)	0.263 (93)	0.223 (89)
salaried positions strong	0.001	0.137 (66)	0.146 (62)	0.304 (96)	0.084 (65)
plans priced low	0.001	0.234 (83)	0.145 (61)	0.172 (76)	-0.116 (12)
smartphone android iphone	0.001	0.136 (66)	-0.057 (16)	0.036 (29)	0.127 (75)
phones well wireless	0.000	0.102 (57)	0.046 (37)	0.242 (90)	0.042 (54)
sprint dealer plans	0.000	0.111 (60)	-0.028 (21)	0.008 (21)	0.039 (54)
personality marketsource offers	0.000	0.114 (75)	0.116 (58)	0.207 (81)	0.428 (100)
strong career path	0.000	-0.031 (20)	0.080 (46)	0.273 (94)	0.268 (93)
text data requirements	0.000	0.154 (70)	0.277 (86)	0.061 (38)	0.054 (58)
per month unlimited	0.000	-0.017 (24)	-0.134 (6)	0.019 (24)	-0.141 (8)
time salaried positions	0.000	0.060 (45)	-0.021 (23)	0.098 (52)	-0.030 (33)
opportunities promoted full	0.000	0.342 (92)	0.140 (60)	0.214 (86)	-0.032 (32)
transportation outgoing friendly	0.000	0.099 (57)	0.105 (52)	0.382 (99)	-0.005 (41)
unlimited voice text	0.000	0.043 (40)	0.116 (55)	0.080 (45)	0.323 (97)
offers opportunities promoted	0.000	0.276 (88)	0.072 (44)	0.186 (80)	-0.011 (38)
verizon sprint dealer	0.000	-0.031 (20)	-0.171 (3)	0.053 (35)	0.101 (70)
must smartphone android	0.000	0.437 (97)	0.246 (81)	0.101 (53)	0.085 (66)
voice text data	0.000	0.098 (56)	0.224 (78)	0.105 (54)	0.281 (95)
well wireless accessories	0.000	0.057 (44)	0.056 (40)	0.270 (93)	0.020 (48)
month unlimited voice	0.000	-0.024 (22)	-0.097 (10)	0.062 (38)	0.191 (85)
wireless accessories inside	0.000	0.156 (70)	-0.023 (22)	0.118 (59)	-0.027 (34)
marketsource offers opportunities	0.000	0.150 (82)	0.097 (51)	0.204 (82)	0.008 (44)
accessories inside target	0.000	0.226 (82)	0.080 (46)	0.076 (43)	-0.008 (39)
low per month	0.000	0.023 (34)	-0.015 (24)	0.095 (50)	-0.148 (7)

		Semantic Similarity Score (Centile of Score)			
Trigram	**Elastic net coefficient**	**Less Physically Able**	**Careful**	**Dependable**	**Negative/ Warm Personality**
android iphone must	0.000	0.433 (97)	0.247 (82)	0.070 (41)	0.081 (65)
iphone must reliable	0.000	0.510 (99)	0.490 (100)	0.331 (97)	0.056 (58)
benefits competitive hourly	0.000	0.200 (79)	0.184 (70)	0.246 (91)	0.065 (61)
career path opportunities	0.000	0.035 (38)	0.008 (29)	0.212 (85)	0.165 (82)
ged must smartphone	0.000	0.459 (97)	0.335 (92)	0.113 (57)	0.047 (56)
friendly personality marketsource	0.000	0.122 (77)	0.146 (67)	0.283 (95)	0.529 (100)
full time salaried	0.000	0.113 (60)	0.088 (48)	0.121 (60)	-0.060 (24)
data requirements must	0.000	0.394 (95)	0.392 (97)	0.127 (62)	-0.003 (41)
dealer plans priced	0.000	0.131 (65)	0.028 (33)	0.095 (50)	-0.04 (30)
contact job poster	-0.001	0.107 (59)	0.137 (60)	0.127 (62)	0.176 (83)
please contact job	-0.001	0.200 (79)	0.243 (81)	0.204 (84)	0.136 (77)
retail sales experience	-0.004	0.012 (31)	0.000 (27)	0.129 (63)	0.151 (79)
including nights weekends	-0.006	-0.067 (12)	-0.109 (9)	-0.024 (13)	-0.046 (28)
full time sales	-0.008	0.062 (46)	0.019 (31)	0.075 (43)	-0.03 (33)
part time sales	-0.011	-0.027 (21)	-0.103 (9)	-0.006 (17)	-0.065 (23)
level customer service	-0.020	0.063 (46)	0.057 (40)	0.208 (84)	0.001 (42)
customer service skills	-0.025	0.171 (74)	0.176 (68)	0.309 (96)	0.171 (83)
customer service sales	-0.026	-0.005 (27)	-0.013 (24)	0.195 (82)	0.001 (42)
retail sales associate	-0.030	-0.070 (12)	-0.114 (8)	0.076 (43)	0.010 (45)
strong work ethic	-0.034	0.064 (46)	0.327 (92)	0.370 (99)	0.386 (99)
part time positions	-0.035	0.019 (33)	-0.052 (17)	0.000 (18)	-0.066 (23)
outstanding customer service	-0.039	-0.006 (26)	0.080 (46)	0.340 (98)	0.069 (62)

Note: See the notes to Table 7.

Table 9: Coefficients from Elastic Net: Old-Female-Sales

		Semantic Similarity Score (Centile of Score)
Trigram	**Elastic Net Coefficient**	**Worse/Better Communication Skills**
position applying subject	0.152	0.253 (62)
dan fan city	0.133	-0.099 (4)
interpersonal communication skills	0.123	0.947 (100)
retail sales associates	0.122	0.120 (33)
time sales associates	0.070	0.099 (28)
vosges haut chocolat	0.068	-0.036 (9)
work well team	0.065	0.287 (69)
full time sales	0.059	0.086 (26)
unlimited earning potential	0.055	0.211 (53)
search inc org	0.052	0.149 (39)
fast paced environment	0.049	0.424 (90)
resume contact information	0.033	0.450 (92)
customer service skills	0.031	0.689 (100)
retail customer service	0.029	0.255 (62)
revealing beautiful skin	0.020	0.017 (15)
running cash register	0.020	-0.037 (9)
please submit resume	0.017	0.047 (19)
great customer service	0.015	0.291 (70)
customer service representative	0.012	0.236 (58)
retail sales experience	0.008	0.330 (77)
applying subject field	0.006	0.298 (78)
subject field email	0.006	0.335 (71)
write position applying	0.006	0.230 (57)
please email resume	0.004	0.167 (43)
outstanding customer service	-0.006	0.338 (79)
perform essential functions	-0.010	0.430 (90)
retail sales associate	-0.029	0.114 (32)
sales customer service	-0.030	0.254 (62)
resume cover letter	-0.034	-0.055 (7)
nights weekends holidays	-0.057	-0.017 (11)

Note: See the notes to Table 7.

Table 10: Coefficients from Elastic Net: Middle-Male-Janitor

		Semantic Similarity Score (Centile of Score)
Trigram	**Elastic Net Coefficient**	**Less Physically Able**
able lift lbs	0.217	0.773 (100)
drug free clean	0.042	0.177 (75)
monday saturday sunday	0.014	-0.078 (10)
nc title custodian	0.014	-0.090 (8)
custodian schedule monday	0.014	-0.009 (25)
schedule monday saturday	0.014	-0.024 (22)
title custodian schedule	0.014	-0.057 (14)
part time position	0.009	0.007 (30)

Note: See the notes to Table 7.

Table 11: Coefficients from Elastic Net: Middle-Male-Sales

		Semantic Similarity Score (Centile of Score)	
Trigram	**Elastic Net Coefficient**	**Experienced**	**Worse with Technology**
full paid training	0.276	0.130 (82)	0.114 (42)
requirements high school	0.081	-0.034 (30)	0.295 (84)
experience plus must	0.080	0.272 (98)	0.156 (53)
sunglass hut experience	0.075	0.196 (93)	0.111 (41)
sales associate needed	0.069	0.197 (94)	0.27 (79)
open close store	0.067	-0.077 (18)	0.059 (28)
part time merchandiser	0.060	0.04 (56)	0.012 (17)
partnership store manager	0.059	-0.083 (16)	0.155 (53)
prior experience retail	0.057	0.241 (97)	0.304 (85)
customer service experience	0.048	0.138 (84)	0.364 (92)
customer service retail	0.048	-0.046 (26)	0.370 (93)
religion national origin	0.043	-0.133 (7)	0.120 (43)
clothes general merchandise	0.033	-0.14 (6)	0.072 (31)
motor clothes general	0.033	-0.069 (20)	0.199 (64)
experience preferably retail	0.020	0.258 (98)	0.387 (94)
fast paced dynamic	0.009	0.216 (95)	0.261 (78)
retail customer service	0.003	-0.046 (26)	0.370 (93)
retail sales experience	0.001	0.222 (96)	0.363 (92)
customer service skills	-0.002	0.05 (59)	0.429 (97)
full part time	-0.020	-0.016 (36)	0.031 (21)

Note: See the notes to Table 7.

Table 12: Coefficients from Elastic Net: Middle-Female-Administrative Assistant

		Semantic Similarity Score (Centile of Score)			
Trigram	**Elastic Net Coefficient**	**Worse Memory**	**Dependable**	**Worse/Better Communication Skills**	**Worse with Technology**
administrative assistant office	0.096	-0.063 (13)	0.062 (38)	0.083 (26)	0.132 (47)
desire learn grow	0.076	0.18 (86)	0.061 (38)	0.348 (80)	0.098 (38)
knowledge microsoft office	0.058	0.263 (96)	0.101 (53)	0.450 (92)	0.509 (99)
pm mon fri	0.046	-0.057 (15)	-0.157 (1)	-0.109 (3)	-0.087 (4)
firm located downtown	0.046	-0.086 (9)	0.007 (20)	-0.097 (4)	0.152 (52)
written verbal communication	0.043	0.224 (93)	0.003 (19)	0.459 (93)	0.154 (52)
full time administrative	0.040	0.082 (60)	0.051 (34)	0.135 (36)	0.114 (42)
years administrative experience	0.020	0.144 (78)	0.048 (33)	0.303 (72)	0.156 (53)
customer service experience	0.014	0.223 (93)	0.267 (93)	0.511 (96)	0.364 (92)
hours monday friday	0.014	0.02 (38)	-0.084 (4)	0.015 (15)	-0.096 (3)
duties full time	0.013	0.095 (64)	0.101 (52)	0.191 (48)	-0.025 (11)
submit resume consideration	0.010	-0.041 (19)	-0.042 (9)	0.117 (32)	0.028 (21)
fast paced environment	0.008	0.163 (82)	0.342 (98)	0.424 (90)	0.353 (91)
must reliable transportation	0.005	0.08 (59)	0.418 (100)	0.351 (81)	0.280 (81)
mon fri pm	0.005	-0.057 (15)	-0.157 (1)	-0.109 (3)	-0.087 (4)
microsoft word excel	0.002	0.217 (92)	0.019 (24)	0.28 (68)	0.304 (85)
customer service skills	-0.006	0.19 (88)	0.309 (96)	0.689 (100)	0.429 (97)
front desk receptionist	-0.006	0.092 (63)	0.091 (49)	0.027 (16)	0.003 (15)
school diploma equivalent	-0.007	0.04 (45)	-0.076 (4)	0.214 (53)	0.221 (69)
limited answering phones	-0.007	0.147 (79)	0.073 (42)	0.26 (63)	0.273 (80)
looking part time	-0.008	0.045 (47)	0.037 (30)	0.008 (14)	-0.006 (14)
high school diploma	-0.016	0.013 (36)	0.000 (18)	0.240 (59)	0.241 (73)
excellent computer skills	-0.023	0.261 (96)	0.372 (99)	0.768 (100)	0.483 (99)

Note: See the notes to Table 7.

Table 13: Top Decile Thresholds

Stereotype	Old-Male-Janitor	Old-Male-Sales	Old-Female-Sales	Middle-Male-Janitor	Middle-Male-Sales	Middle-Female-Admin
Health						
Less Attractive						
Hard of Hearing						
Worse Memory						0.215 [0.055]
Less Physically Able	**0.330 [0.101]**	**0.330 [0.101]**		**0.330 [0.101]**		
Personality						
Less Adaptable	0.292 [0.111]					
Careful	**0.324 [0.106]**	**0.324 [0.106]**				
Less Creative						
Dependable	0.253 [0.102]	0.253 [0.102]				0.253 [0.102]
Negative/Warm Personality		0.244 [0.044]				
Skills						
Lower Ability to Learn						
Worse/Better Communication Skills	**0.442 [0.202]**		**0.442 [0.202]**			**0.442 [0.202]**
More Experienced					0.180 [0.030]	
Less/More Productive	0.252 [0.080]					
Worse with Technology					**0.356 [0.155]**	**0.356 [0.155]**

Note: For each result that was shown to be significant in Table 5, we report the threshold cosine similarity scores for the top decile of the cosine similarity score distribution for the stereotype. In brackets, we report the mean of the distribution to help scale these cutoffs. Bolded cells have higher cosine similarity thresholds for the top decile of trigrams of 0.3 or higher.

Table 14: Number of Selected N-grams When Varying Number of Words in a Phrase

Gender	Age	Occupation	1 word	2 words	3 words	4 words	5 words
Male	Old	Janitor	2	1	21	477	1
		Sales	1	2	89	55	47
		Security	0	1	1	136	402
	Middle	Janitor	1	0	8	250	167
		Sales	54	1	20	0	0
		Security	0	0	0	0	118
Female	Old	Admin	3	8	1	3	3
		Sales	12	29	30	6	3
	Middle	Admin	41	20	23	4	0
		Sales	6	13	2	0	2

Note: Each cell reports the number of N-grams selected by the elastic net when using the reported number of words in a phrase. Our main analysis uses trigrams (three words). The elastic net models are estimated using the same parameters and controls.

Table 15: Robustness of Results to Varying Number of Words in a Phrase

Age-Occupation-Gender	Stereotype	Threshold	1 word	2 words	3 words	4 words	5 words
Old Male Sales	Less Physically Able	Total selected	1	2	89	55	47
		Top 1%	0.00% (1.000)	0.00% (1.000)	4.49%** (0.012)	0.00% (1.000)	0.00% (1.000)
		Top 5%	0.00% (1.000)	0.00% (1.000)	11.24% (0.013)	7.27% (0.355)	8.51% (0.297)
		Top 10%	0.00% (1.000)	0.00% (1.000)	16.85%** (0.049)	21.82%** (0.010)	25.53%*** (0.002)
	Careful	Total selected	1	2	89	55	47
		Top 1%	0.00% (1.000)	0.00% (1.000)	4.49%** (0.012)	0.00% (1.000)	0.00% (1.000)
		Top 5%	0.00% (1.000)	0.00% (1.000)	7.87% (0.216)	5.45% (0.755)	8.51% (0.297)
		Top 10%	0.00% (1.000)	0.00% (1.000)	12.36% (0.477)	9.09% (1.000)	8.51% (1.000)
Middle Administrative Assistant Female	Worse/Better Communication Skills	Total selected	41	20	23	4	0
		Top 1%	0.00% (1.000)	10.00%** (0.0169)	8.70%** (0.022)	25.00%** (0.039)	0.00% (1.000)
		Top 5%	14.63%** (0.016)	15.00%* (0.075)	13.04% (0.105)	25.00% (0.185)	0.00% (1.000)
		Top 10%	39.02%*** (0.000)	20.00% (0.133)	21.74%* (0.073)	25.00% (0.344)	0.00% (1.000)
Old Male Janitor	Worse/Better Communication Skills	Total selected	2	1	21	477	1
		Top 1%	0.00% (1.000)	0.00% (1.000)	8.70%** (0.022)	5.24%*** (0.000)	0.00% (1.000)
		Top 5%	0.00% (1.000)	0.00% (1.000)	13.04% (0.105)	0.943%*** (0.000)	0.00% (1.000)
		Top 10%	0.00% (1.000)	0.00% (1.000)	21.74%* 0.073	15.30%*** (0.000)	0.00% (1.000)
Old Female Sales	Worse/Better Communication Skills	Total selected	12	29	30	6	3
		Top 1%	16.67%*** (0.006)	20.69%*** (0.000)	6.67%** (0.036)	0.00% (1.000)	0.00% (1.000)
		Top 5%	16.67% (0.118)	24.14%*** (0.000)	6.67% (0.661)	0.00% (1.000)	0.00% (1.000)
		Top 10%	41.67%*** (0.004)	27.59%*** (0.006)	10.00% (1.000)	0.00% (1.000)	0.00% (1.000)
Middle Male Sales	Worse with Technology	Total selected	54	1	20	0	0
		Top 1%	5.56%** (0.017)	0.00% (1.000)	0.00% (1.000)	0.00% (1.000)	0.00% (1.000)
		Top 5%	12.96%** (0.018)	0.00% (1.000)	5.00% (1.000)	0.00% (1.000)	0.00% (1.000)
		Top 10%	24.07%*** (0.002)	0.00% (1.000)	30.00%** (0.011)	0.00% (1.000)	0.00% (1.000)

Note: For each result that was bold-faced in Table 13, we estimate the elastic net models varying the number of words in a phrase. The share of selected phrases above the cutoff are reported first, followed by the p-value from the binomial test of proportions.

* p<0.1, ** p<0.05, *** p<0.01

Table 16: Correlation Between Stereotyped Phrases and Discrimination Against Younger Applicants (in Favor of Older Applicants)

Stereotype	Old-Male-Janitor	Old-Female-Sales	Old-Female-Admin	Middle-Female-Admin
Number of selected trigrams	1	1	5	10
Health				
Less Attractive				Top 10%: 10.0% (p=1.000)
Hard of Hearing			Top 10%: 20.0% (p= 0.410) Top 5%: 20.0% (p= 0.226)	
Worse Memory			Top 10%: 40.0%* (p= 0.081) Top 5%: 40.0%** (p= 0.023)	
Less Physically Able				Top 10%: 10.0% (p=1.000) Top 5%: 10.0% (p=0.401) Top 1%: 10.0%* (p=0.096)
Personality				
Less Adaptable				Top 10%: 10.0% (p=1.000)
Careful				Top 10%: 20.0% (p= 0.264) Top 5%: 10.0% (p=0.401)
Less Creative				
Dependable				Top 10%: 30.0%* (p=0.070) Top 5%: 10.0% (p=0.401) Top 1%: 10.0%* (p=0.096)
Negative/Warm Personality				Top 10%: 10.0% (p=1.000) Top 5%: 10.0% (p=0.401)
Skills				
Lower Ability to Learn				Top 10%: 10.0% (p=1.000) Top 5%: 10.0% (p=0.401)
Worse/Better Communication Skills				
More Experienced				Top 10%: 20.0% (p=0.264) Top 5%: 10.0% (p=0.401)
Less/More Productive				
Worse with Technology				

Note: Not shown are the results for age-gender-occupation triplets for which no trigrams were selected (janitor: middle-aged males, sales: middle-aged females and middle-aged and older males, and security: middle-aged and older males), for which the tests of over-representation of selected trigrams are uninformative.

* p<0.1, ** p<0.05, *** p<0.01

Table 17: Placebo Analysis

Gender	Age	Occupation	Trigrams selected using true outcomes	Trigrams selected using placebo outcomes
Male	Old	Janitor	21	0
		Sales	89	5
		Security	1	0
	Middle	Janitor	8	0
		Sales	20	0
		Security	0	0
Female	Old	Admin	1	0
		Sales	30	0
	Middle	Admin	23	8
		Sales	2	0

Note: In the placebo outcomes column, ads were randomly assigned to have discriminated against older workers. The share of ads assigned to the placebo treatment group within each age-gender-occupation cell was set to be identical to the overall share of observations in the data (11%).

Appendix Table A1: Coefficients from Elastic Net for Excluded Samples

Algorithm	Trigram	Elastic Net Coefficient
Old-Male-Security	high school diploma	0.020
Middle-Male-Security	N/A	
Old-Female-Administrative Assistant	resume cover letter	0.021
Middle-Female-Sales	sales customer service part time sales	0.045 -0.011

Note: This table presents the selected trigrams for elastic net algorithms where fewer than three trigrams were selected.