Cognition 164 (2017) 46–60

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

The semantic representation of prejudice and stereotypes

Sudeep Bhatia

Department of Psychology, University of Pennsylvania, Philadelphia, PA, United States

ARTICLE INFO

Article history:
Received 22 March 2016
Revised 23 March 2017
Accepted 24 March 2017

Keywords:
Semantic representation
Latent semantic analysis
Prejudice
Stereotyping
Implicit association test

ABSTRACT

We use a theory of semantic representation to study prejudice and stereotyping. Particularly, we consider large datasets of newspaper articles published in the United States, and apply latent semantic analysis (LSA), a prominent model of human semantic memory, to these datasets to learn representations for common male and female, White, African American, and Latino names. LSA performs a singular value decomposition on word distribution statistics in order to recover word vector representations, and we find that our recovered representations display the types of biases observed in human participants using tasks such as the implicit association test. Importantly, these biases are strongest for vector representations with moderate dimensionality, and weaken or disappear for representations with very high or very low dimensionality. Moderate dimensional LSA models are also the best at learning race, ethnicity, and gender-based categories, suggesting that social category knowledge, acquired through dimensionality reduction on word distribution statistics, can facilitate prejudiced and stereotyped associations.

1. Introduction

Distributional models of semantic memory provide a powerful approach to understanding semantic representations (Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Kwantes, 2005; Landauer & Dumais, 1997; Lund & Burgess, 1996; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014). One of the main insights underlying these models is that the representations of words reflect the structure of word co-occurrence in natural language (Firth, 1957; Harris, 1954). Studying this structure, by applying these models to large-scale natural language corpora, can shed light on the representations that people have of common words, the relationships and associations between the concepts that these words represent, and the ways in which these relationships affect cognition and behavior.

Distributional models often characterize the words in their vocabulary as multi-dimensional vectors, with the proximity between the vectors of two words corresponding to the relatedness or association of the words. The dimensionality of these vectors is often smaller than that necessary to represent the data on which the model is trained, so that learning the vector representations involves performing some type of dimensionality reduction on word distribution statistics (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). Appropriate levels of vector dimensionality allow distributional models to accurately predict response probabilities and response times in a wide range of settings, including semantic priming tasks, free association tasks, recall tasks, word similarity tasks, and categorization tasks (see Bullinaria & Levy, 2007 or Jones, Willits, & Dennis, 2015 for a review).

The use of distributional models is typically limited to non-social psycholinguistic settings. We wish to use these models to better understand prejudice and stereotyping. In this paper, we recover race-based, ethnicity-based, and gender-based vector representations from the types of natural language environments individuals interact with on a day-to-day basis, and examine whether our recovered representations possess the prejudiced and stereotyped associations documented in social psychological research. Importantly, we test the effects of mechanisms like dimensionality reduction on the strength of these prejudices and stereotypes. These mechanisms are necessary for the efficient learning of word meaning and association, and play a key role in the learning of categories. Examining whether these otherwise desirable cognitive mechanisms also generate undesirable social biases, can shed light on the cognitive underpinnings of these biases, and the ways in which these biases depend on social category knowledge and category-based generalization.

1.1. Prejudiced and stereotyped associations

Prejudice and stereotyping are often studied in terms of the associations that automatically influence judgment and behavior when relevant social categories are activated (Allport, 1954; Devine, 1989; Fazio, Jackson, Dunton, & Williams, 1995; Gaertner

E-mail address: bhatiasu@sas.upenn.edu

http://dx.doi.org/10.1016/j.cognition.2017.03.016
0010-0277/

& McLaughlin, 1983; Greenwald & Banaji, 1995; Strack & Deutsch, 2004). These associations are often considered to be implicit, that is, outside of the awareness of the individual in consideration. For this reason, they are studied using experimental tasks with measures that do not rely on the individual's ability to consciously assess (and suppress) these associations. Perhaps the most common such task in use today is the implicit association test (IAT) (Cunningham, Preacher, & Banaji, 2001; Greenwald, McGhee, & Schwartz, 1998), which provides a latency-based measure of associations for social categories. With the use of the IAT and related measures (Fazio & Olson, 2003), researchers have found stronger associations between stereotypically African American names and negatively valenced words and stronger associations between stereotypically White names and positively valenced words (Greenwald et al., 1998; also Dovidio, Evans, & Tyler, 1986; Fazio et al., 1995; Gaertner & McLaughlin, 1983), illustrating associative prejudices favoring Whites over African Americans. Similar methods have also been applied to study stereotypes, which do not involve diverging associations with differently valenced words, but rather diverging associations with words in different semantic categories. For example, researchers have used the IAT to demonstrate a stronger association between female names and weakness-related words and a stronger association between male names and power-related words (Rudman, Greenwald, & McGhee, 2001; also Nosek, Banaji, & Greenwald, 2002).

Biased associations have been shown to play a role in influencing peoples' behaviors (Dovidio, Kawakami, & Gaertner, 2002; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Hamilton & Gifford, 1976; Judd & Park, 1988; McConnell & Leibold, 2001; Olson & Fazio, 2001; but also see e.g. Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013) and are considered to be one of the most important psychological determinants of prejudice and stereotyping. Given this importance, it becomes desirable to characterize what these associations are and the ways in which these associations are represented. One way to do this involves studying the distribution of names, words, and concepts in real-world natural language environments. People exposed to everyday language that presents African American names in negative contexts and White names in positive contexts, or female names in positions of weakness and male names in positions of power, will develop the prejudices and stereotypes documented in the above work. Equivalently, these prejudices and stereotypes will be reflected in the use of this language, causing African American names to be more likely to appear in negative contexts and less likely to appear in positive contexts, relative to White names, and female names to be more likely to appear in positions of weakness and less likely to appear in positions of power, relative to male names.

Studying the types of race-based or gender-based associations present in everyday language can not only shed light on the actual associations possessed by individuals, but also the ways in which these associations reflect social representations. This can then help us directly compare what we know about the cognitive basis of prejudice and stereotyping with what we know about the representation of non-social concepts. Does the representation of prejudice and stereotypes rely on same mechanisms involved in the representation of word relationships, categories, meanings, and associations in other settings? These mechanisms often facilitate efficient linguistic comprehension and word use, so could it be that prejudice and stereotyping are the harmful byproducts of an otherwise desirable system for making semantic inferences and generalizations?

1.2. Latent semantic analysis

These questions can be answered by applying theories of distributional semantics to common natural language datasets. The representations built using this method can then be tested for the types of associative biases observed in human participants, using, for example, stimuli from existing implicit association tests. The distributional model we consider in this paper is latent semantic analysis (LSA). LSA has been shown to be useful for a number of different applications in semantic memory research and computational linguistics, and is perhaps the most influential such model in this area (Landauer & Dumais, 1997; Landauer et al., 1998). Its core assumption is that decision makers represent words and concepts using a multidimensional word-vector space, built from word-distribution data. This vector space may have a high number of dimensions, but importantly, these dimensions are much less than those required for representing all of the information in the data. LSA achieves this dimensionality reduction using singular value decomposition.

Consider a setting with N different words occurring in K different contexts. These contexts could be different articles in newspapers, as in the dataset we consider below, chapters in books, conversations on the internet, or even non-textual experiences. The distribution of these words across the different contexts can be represented in an $N \times K$ matrix S. S captures word-context co-occurrence, so that the cell in row n and column k corresponds to the number of times word n occurs in context k.

LSA attempts to recover vector representations of the N words by performing a singular value decomposition on the matrix S, which describes S using some $M \ll K$ latent dimensions. The matrix recovered through this singular value decomposition can be written as $S^* = U \cdot V \cdot W$ where V is an $M \times M$ matrix with the M largest singular values from the decomposition, U is the corresponding $N \times M$ matrix of words, and W is the corresponding $M \times K$ matrix of contexts. U is of particular interest to us as it contains a representation of each of the N words as vectors on the M latent dimensions. The proximity between these vectors can be used to provide a quantitative account of word relationship and association. The metric typically used to compute vector proximity, and thus word association, is cosine similarity, so that the proximity between any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by $\text{sim}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y} / (||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||)$. This metric varies between -1 and $+1$ (with 0 capturing orthogonal vectors and $+1$ capturing vectors with identical directions) (see Landauer et al., 1998 for details).

In their classic article, Landauer and Dumais (1997) showed that the above technique could be used to model judgments of word similarity and their dependence on the rate of vocabulary acquisition, specify the comprehension and comprehensibility of pieces of text, predict word priming effects, learn the representation of numerals, and display desirable properties in a number of other settings. Related work has shown that similar approaches are also able to predict human behavior in free association tasks, recall tasks, semantic categorization tasks, and in a wide variety of other psycholinguistic experiments (see Bullinaria & Levy, 2007; Jones et al., 2015; Turney & Pantel, 2010).

1.3. Dimensionality reduction

Importantly these results rely on the appropriate choice of M, which is the total number of latent dimensions recovered by singular value decomposition. Dimensionality reduction facilitates induction and generalization, so that if M is too large or if $M = K$ (which is the special case with no dimensionality reduction) the model is unable to generalize what it learns about words in a certain context to other word and other contexts. Thus, although an LSA model may note that the words *car* and *gear* occur together and are related, and the words *car* and *brake* occur together and are related, unless *gear* and *brake* occur together, an LSA model with a large value of M would not be able to infer that *gear* and *brake* are related. A very small value of M, or too much dimension-

ality reduction, would also be detrimental to this type of inference: Although an LSA model with very small M may be able to claim that *gear* and *brake* are related, it would also make spurious inferences regarding other pairs of words (again, see Landauer & Dumais, 1997 for details).

These insights have been demonstrated in non-social psycholinguistic setting. However, they also hold for the types of inferences and generalizations at play when making inferences about individuals. In these settings dimensionality reduction can be seen as facilitating social categorization, and the use of social category knowledge to generalize across individuals.

Consider, for example, a corpus in which some male names occasionally co-occur with other male name, some female names occasionally co-occur with other female name, some male names occur alongside power-related words, and some female names occur alongside weakness-related words. In such a setting, an LSA model without dimensionality reduction would not have cohesive associations between groups of male name and groups of female names. For this reason, it may associate some male names with power and some female names with weakness, but would not extend these associations to all male names or all female names. This could change, however, if the number of dimensions in the model is reduced, so that the association between some female names and other female names, and some male names and other male names, is used to infer cohesive associations between all female names, and cohesive associations between all male names. This would allow the model to distinguish female names from male names, and, more generally, categorize names as being either female or male. This type of categorization would then facilitate gender-based generalization, so that the fact that some male names occur alongside power-related words, and some female names occur alongside weakness-related words, would be used to infer an association between all male names and power words, and all female names and weakness words. Of course, if M becomes very small, the model may also begin to associate various male names with female names, and, in turn male names with weakness words and female names with power words, causing these biases to disappear. This property also extends beyond associations with stereotypes and holds for groups of positively and negatively valenced words (see Bestgen & Vincze, 2012; Maas et al., 2011; Recchia & Louwerse, 2015 for discussions and applications of this property).

1.4. Datasets

For the above reasons it appears as if LSA could be used to study the learning of the associations at play in social settings, particularly in prejudice and stereotyping, in the same way as it is used to study the learning of associations in standard psycholinguistic tasks. Additionally, dimensionality reduction, a feature of this approach that facilitates the efficient learning of word relationships and associations could be responsible for the learning of prejudices and stereotypes, through the learning of social categories.

In this paper we are primarily concerned with the types of prejudices and stereotypes studied in existing experimental work. We utilize stimuli from a set of influential implicit association test (IAT) experiments (specifically Greenwald et al., 1998; Nosek et al., 2002; Rudman et al., 2001), and train our models on natural language data from the North American News Text Corpus (available from the Linguistic Data Consortium), a well-known dataset of news articles published in the United States. From this corpus we extract two sub-corpora. The first is a set of syndicated Los Angeles Times - Washington Post (LATWP) news articles published between May 1994 and August 1997 (approximately 52 million words), and the second is a set of syndicated New York Times (NYT) news articles published between July 1994 and December 1996 (approximately 173 million words). We train our LSA models on each of these two corpora separately, and test whether the biases observed in research using the implicit association test also emerge in our trained representations, whether these biases vary based on the number of dimensions in the corresponding LSA model, and whether dimensions that lead to the strongest biases are also the ones with the most accurate social categorization. Two sets of tests (one for LATWP and one for NYT) helps ensure that our final results are robust.

The use of news articles to train our LSA models reflects standard practices in semantic memory research and computational linguistics. These types of datasets provide a good measure of the nature and structure of everyday language, as well as common social attitudes and beliefs, which are communicated through this language. Indeed, the use of two relatively liberal newspaper datasets for our analysis provides a somewhat conservative tests of the existence of prejudice and stereotyping in everyday language. We would expect any documented biases to be more pronounced for conservative newspapers or social media outlets.

There are also practical reasons for using the LATWP and NYT articles. Firstly, the experiments whose stimuli we use were performed in the 1990s and early 2000s (in the United States), which is around the time at which these articles were published (also in the United States). Thus the LATWP and NYT corpora closely resemble the actual natural language environments of the participants in Greenwald et al. (1998) and Nosek et al. (2002), and other papers. Secondly, news articles are more likely than other text sources to mention individual names, and names that were common in the 1990s are especially likely to be mentioned in the LATWP and NYT corpora. The tests in this paper (as well as many existing IAT experiments) rely critically on names to measure associative biases, and such tests would not be possible with other types of datasets. Indeed, we first attempted our analysis using a publicly available internet version of LSA (http://lsa.colorado.edu), trained on the "General Reading Up to First Year College" corpus with 300 dimensions. This corpus had representations for only 4 out of the 25 female African American names in the Greenwald et al. (1998) IAT. In contrast our LATWP and NYT corpora have representations for 16 and 18 out of the 25 names, respectively.

1.5. Prior work

Prior work does suggest that models trained on everyday natural language data are likely to possess prejudices and stereotypes. Notably, Lynott, Kansal, Connell, and O'Brien (2012) have recently shown that simple co-occurrence frequencies in natural language predict the strength of IAT scores across different domains, so that IAT tests with stimuli that have systematic co-occurrence relationships in language are also the tests for which human participants display the strongest biases. Although Lynott et al. use the Web 1T corpus (a large dataset of webpages indexed by Google) and apply it to different IAT stimuli than that used in this paper, we would expect their insights to extend to our analysis as well. That is, word co-occurrence frequencies on our NYT and LATWP corpora should reflect the IAT biases observed in the work of Greenwald et al. (1998), Nosek et al. (2002), Rudman et al. (2001), and others.

Of course the goal of our analysis is different to that of Lynott et al. While Lynott et al. test the relationship between participants' IAT scores and linguistic association for different types of social and non-social stimuli, and use this to predict relative IAT scores for different stimuli, we are interested in studying how social categories are represented. To this end we uncover LSA-based semantic representations from word co-occurrence. Additionally, we examine the specific features of the semantic representations that lead to prejudice and stereotypes by testing LSA models with varying dimensionality, and by evaluating the degree to which learnt

social categories predict prejudice and stereotyping in these models. In doing so, our tests shed light on not only the linguistic correlates of IAT behavior, but the specific cognitive mechanisms that facilitate social categorization, and in turn, prejudiced and stereotyped representations.

In closely related work Lenton, Sedikides, and Bruder (2009) have shown that an LSA model (with 300 dimensions) trained on the TASA corpus, possesses gender stereotypes. Particularly, they find that masculine and feminine referents (e.g. *man* and *woman*) are strongly associated with stereotypically masculine and feminine traits. Again, despite the use of a different corpus, and the use masculine/feminine referents rather than male or female names, we would expect Lenton et al.'s results to emerge in our analysis. Thus, the representations possessed by our LSA models should have stronger associations between male names and stereotypically male traits, and female names and stereotypically female traits. Of course, our work goes beyond just testing for the presence of such associations. As discussed above, we wish to uncover the cognitive mechanisms that facilitate these representations. Thus, unlike Lenton et al., we examine numerous LSA models with varying dimensionalities, and test for both the presence of gender stereotypes (as well racial and ethnic prejudices), and the dependence of these stereotypies and prejudices on learnt category knowledge.

Finally, language in media has been argued to play an important role in the development of stereotyping and prejudice. For example, scholars have documented diverging descriptions of male and female political candidates in news media, and diverging descriptions of male and female characters in children's storybooks, and have used this to suggest that language use in media reflects and perpetuates gender stereotypes (Crabb & Bielawski, 1994; Kahn & Goldenberg, 1991). Likewise, considerable research has found that the language used to depict African Americans in news programs differs from the type of language used to depict Whites (Dixon & Linz, 2000; Entman, 1992). Our research builds on these insights to specify why the application of common semantic memory mechanisms to the information present in media datasets, leads to the biased associations observed in human participants in experimental work. As such, it connects three areas of relevance to the study of stereotyping and prejudice: social psychological research on human behavior, sociological research on media bias, and cognitive science research on the learning of semantic representations.

2. Overview of methods

We wish to train our LSA models on the LATWP and NYT datasets in order to learn the semantic representations of different names. Previous work has suggested that LSA models with 300 dimensions are best for semantic memory tasks, and so we will use 300 dimensional LSA models for our initial analysis. However, the key aim of our analysis is to examine the effect of dimensionality on social categorization and subsequently prejudice and stereotyping. Thus we will also consider LSA models with dimensions varying between 50 and 950, in increments of 50. In addition to this, we will test a special low dimensional model with only 2 dimensions, and a special high dimensional model without any dimensionality reduction. As we train each of our models separately on each of the two datasets, this leads to a total of 42 unique trained models.

As is standard with LSA model training, each of our datasets is modified by a term frequency-inverse document frequency (tf-idf) transformation, which reweighs the word frequency statistics used in the singular value decomposition, to control for word frequency effects (Landauer et al., 1998). Overall, the tf-idf value for a word in a context increases proportionally to the number of times a word appears in that context, but is offset by the frequency of the word in the corpus. Prior to the application of the tf-idf transformation, we tokenize each article using white space and punctuation, and lower-case and stem all of the words in the corpus using the porter-stemmer algorithm. Our models are trained with the help of the gensim toolbox (Řehůřek & Sojka, 2010).

In order to ensure consistency across our analyses as well as a close relationship with experimental findings, we will use prejudices and stereotypes studied using existing implicit association tests. The IAT is not the only way to measure association-based prejudices and stereotypes, but it is the most commonly used task of this nature. For this reason, materials and stimuli from different existing IATs are widely available, making it easier for us test our models for different types of prejudices and stereotypes.

The version of the IAT used to study racial prejudice typically involves two different sets of names, corresponding to stereotypical African American and White names, as well as sets of positively valenced and negatively valenced words (see Greenwald et al., 1998). Participants are asked to categorize the names and words based on their social groups and valence, with the ease of categorization (measured using reaction time) indicating the degree to which the names of the two groups are associated with positive vs. negative words. Thus the observation that participants are quicker to categories the names and words when asked to select between "African American names or unpleasant words" and "White names or pleasant words" than when asked to select between "African American names or pleasant words" and "White names or unpleasant words" suggests that African American names are more associated with negative words and White names are more associated with positive words. The version of this test used to study stereotypes is very similar, except that the sets of positively and negatively valenced word are replaced with words corresponding to stereotypes for the two groups, such as, for example, weakness and power-related words for gender-based stereotypes.

For any two sets of names, A and B, and two sets of words, C and D, we compute the association of each of the names and each of the words using the cosine similarity of the name vector and the word vector for the corresponding LSA model. These associations are then analyzed using standard techniques. Particularly we use linear regressions in which the strength of association between a given name and a given word is the dependent variable, and the category of the name (A or B), the category of the word (C or D), and the interaction between the two, are the independent variables. Testing for a significant interaction effect can indicate whether the strength of association is strongest between names in A or B and words in C or D.

Although such a technique can establish the presence or absence of biases for a given LSA model, it cannot be used to compare the strength of biases across LSA models with varying levels of dimensionality. Ultimately, overall distances (and thus differences in distances) in word vector space change with the number of dimensions in that space, implying that the size of the interaction effect coefficient, obtained from the above linear regression, is confounded with the dimensional features of the model in consideration. As comparing the strength of bias across LSA models with varying dimensions is a critical part of our analysis, we also use non-parametric, or ordinal, techniques. As with the regression techniques outlined above, we obtain, for each name i, a list of cosine similarity-based associations with the words in C and D. However, we then transform this list into a ranking, and subsequently calculate the sum of the rankings of the words in set C in this list. This number, $s_i(C, D)$, corresponds to the ordinal strength of association of name i with words in set C vs. D, so that higher values of $s_i(C, D)$ imply that the words in C are closer in ordinal distance to name i relative to the words D. Now, we have such num-

bers for each of the names in A and B, and we use these to calculate the average association of names in A to words in C relative to D, $S_A(C, D) = \text{AVE}_{i \text{ in } A}[s_i(C, D)]$, and the average association of names in B to words in C relative to D, $S_B(C, D) = \text{AVE}_{i \text{ in } B}[s_i(C, D)]$. These average relative associations are then compared against each other to obtain $S_A(C, D) - S_B(C, D)$. Positive values of this measure indicate that names in A are more associated with words in C relative to D, when compared to names in B. Negative values of $S_A(C, D) - S_B(C, D)$ indicate the opposite. Note that as $s_i(C, D)$ are not distributed according to the normal distribution, we also have to use nonparametric tests to examine the statistical significance of differences in $s_i(C, D)$ for names in A vs. names in B. This can be done with the Wilcoxson rank-sum test, which is a nonparametric alternative to the t-test. This test evaluates whether $s_i(C, D)$ for names in A are higher or lower than $s_i(C, D)$ for names in B, and thus allows us to evaluate whether there are significant difference between $S_A(C, D)$ and $S_B(C, D)$. The Appendix A provides additional details regarding this method.

An important part of our analysis also involves an examination of social categories: We wish to test whether the biases are most pronounced in models which closely associate names in A or B with other names in A or B. The above techniques can also be used to perform this type of test. For a given name i, this involves replacing C with A and D with B, and removing name i from its corresponding group (to ensure that the strong association between name i and itself doesn't influence the category judgment). The relative association between the names in A and B and other names in A and B can then be evaluated using $S_A(A, B) - S_B(A, B)$, and the Wilcoxson rank-sum test can be used to determine whether the corresponding differences are significantly positive or negative. Again, the Appendix A provides additional details regarding this method.

3. Prejudice

In this section we examine prejudiced associations in our trained LSA models using stimuli from three existing implicit association tests. Our first two tests pertain to prejudiced implicit associations for male African American first names relative to male White first names, and female African American first names relative to female White first names, and examine the strength of association of these names with various positively and negatively valenced words. For these tests we use the names and words from the classic race IAT of Greenwald et al. (1998). Our third test pertains to prejudiced implicit associations for Latino last names compared to White last names, and considers the names and words from the IAT used by Pérez (2010). Note that in order to ensure the comparability of our two corpora, our tests exclude the relatively few names that are not present in both of our corpora.

3.1. African American names vs. White names

One of the most prominent examples of prejudiced associations pertains to racial biases against African Americans. This application was a key feature of the seminal IAT paper of Greenwald et al. (1998), which used a list of 25 common male African American first names, 25 common male White first names, 25 common female African American first names and 25 common female White first names, as well as 25 positively valenced words and 25 negatively valenced words. Using the IAT procedure, it found that nearly all of the participants in the experiment displayed prejudiced racial associations by having quicker response times for categorizing African American names with negatively valenced words and White names with positively valenced words.

Do our LSA models reflect these biases? We begin our analysis with 300 dimensional models trained on the LATWP and NYT corpora. As discussed in the previous section, this is often considered to be the ideal size for LSA-based vector representations, and has been previously shown to lead to the best performance in a number of psychological tasks. We find that both our 300 dimensional LSA display stronger associations between African American names and negatively valenced words, and stronger associations between White names and positively valenced words, as evaluated by a positive regression interaction effect (methods described above and in the Appendix A). This bias emerges for both male names ($\beta = 0.015$, $t = 2.98$, $p < 0.01$, 95% CI = [0.005, 0.025] for LATWP and $\beta = 0.012$, $t = 2.46$, $p < 0.05$, 95% CI = [0.002, 0.021] for NYT) and female names ($\beta = 0.019$, $t = 2.82$, $p < 0.01$, 95% CI = [0.006, 0.033] for LATWP and $\beta = 0.035$, $t = 5.04$, $p < 0.01$, 95% CI = [0.021, 0.048] for NYT). More generally, 35 of our 42 LSA models display stronger negative associations for male African American names relative to male White names (with 18 of these associations reaching statistical significance at $p < 0.05$), and 41 of our 42 LSA models display stronger negative associations for female African American names relative to female White names (with 28 of these associations reaching significance at $p < 0.05$). An analysis that pools all of the dimensions and permits random intercepts for each dimension, reveals a very strong aggregate bias against African Americans for both male names ($\beta = 0.010$, $z = 6.79$, $p < 0.01$, 95% CI = [0.007, 0.013] for LATWP and $\beta = 0.012$, $z = 9.32$, $p < 0.01$, 95% CI = [0.009, 0.015] for NYT) and female names ($\beta = 0.016$, $z = 8.11$, $p < 0.01$, 95% CI = [0.012, 0.020] for LATWP and $\beta = 0.026$, $z = 16.71$, $p < 0.01$, 95% CI = [0.023, 0.029] for NYT).

These results do not only emerge with the interaction-based regressions. They are also reflected in a non-parametric analysis of name associations. Particularly, using the methods described in the prior section, we measure the average ordinal association between different groups of names, and positively and negatively valenced words. Figs. 1 and 2 illustrate these associations, separately for male names and female names, by plotting $S_A(C, D) - S_B(C, D)$ in solid lines for each of the 42 models. Here A is the set of African American names, B is the set of White names, C is the set of negatively valenced words, and D is the set of positively valenced words. $S_A(C, D)$ captures the average ordinal association of names in A to words in C relative to D, and $S_B(C, D)$ captures the average ordinal association of names in B to words in C relative to D. Here positive values of $S_A(C, D) - S_B(C, D)$ correspond to stronger associations between African American names and negatively valenced words and stronger associations between White names and positively valenced words, and thus represent prejudice against African Americans. These figures also indicate statistically significant differences between $s_i(C, D)$ for names in A vs. names in B, as evaluated with the Wilcoxson rank-sum test. Here differences with $p < 0.01$ shown using black circles and $p < 0.05$ shown using grey circles. Differences with $p > 0.05$ are shown with white circles.

What is perhaps more interesting than the fact that we observe a racial bias in our models is the fact that this bias appears to depend strongly on the dimensionality of the model. Overall, we find that the models with the greatest prejudice, that is, models with the most positive values of $S_A(C, D) - S_B(C, D)$, are models with moderate dimensionality. For example, the maximum racial bias in the LATWP corpus emerges for 200 dimensional models for male African American and White names and 150 dimensional models for female names. Likewise, the maximum racial bias in the NYT corpus emerges for 150 and 250 dimensions for male and female names respectively.

We can better understand the nature of these biases by examining the strength of association between names in one group with other names in the same group compared to names in the second group. As discussed above, this can be done by replacing C and D in our analysis with A and B (and removing the name in consideration

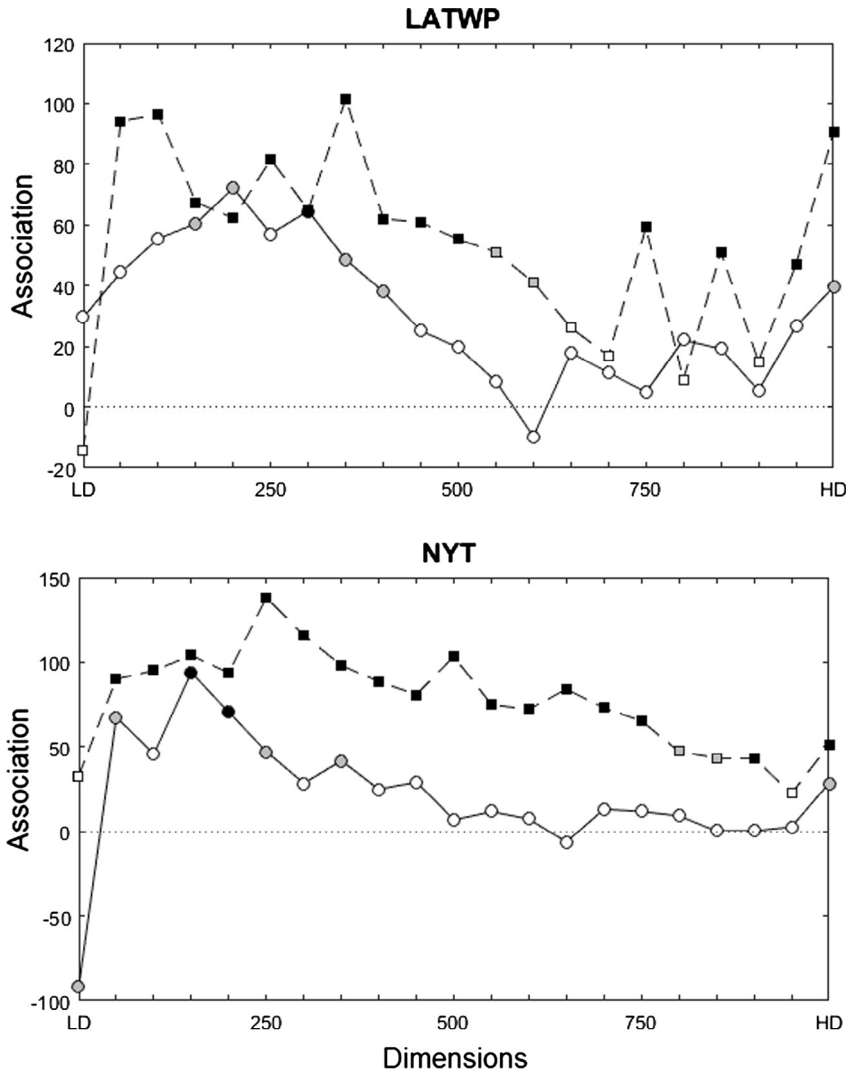

Fig. 1. Prejudiced associations (solid lines/circles) and success at racial categorization (dashed lines/squares), as a function of LSA dimensionality for male African American and White names. The two panels indicate associations for the LATWP and NYT corpora, using stimuli from Greenwald et al. (1998). Here positive values for the solid line/circles represent bias against African Americans, and positive values for the dashed line/squares represent success at race-based categorization. Black circles/squares indicate $p < 0.01$, grey circles/squares indicate $p < 0.05$, and white circles/squares indicate $p > 0.05$.

from A or B to avoid the association between that name and itself from influencing our measured associations). With this analysis we find that most of our 42 LSA models succeed at racial categorization for both male and female names, as shown in Figs. 1 and 2. These figures plot $S_A(A, B) - S_B(A, B)$ in dashed lines for each model. Positive values correspond to stronger associations between African American names and other African American names, and stronger associations between White names and other White names, and indicate an increased accuracy in categorizing names by race. Statistical significance (as determined by the Wilcoxson rank sum test) is displayed using black ($p < 0.01$), grey ($p < 0.05$), or white ($p > 0.05$) squares.

As can be seen in these figures, the effect of dimensionality observed with negatively and positively valenced words also emerges with names. Particularly, the most accurate race-based name categorization, corresponding to the most positive values of $S_A(A, B) - S_B(A, B)$, occurs for 350 dimensional models in the LATWP corpus for both male and female names, and 250 and 500 dimensional models in the NYT corpus for male and female names respectively. For this reason, we find a correlation of 0.47 ($p < 0.01$) between $S_A(C, D) - S_B(C, D)$ and $S_A(A, B) - S_B(A, B)$ for male names, and a correlation of 0.49 ($p < 0.01$) between $S_A(C, D) - S_B(C, D)$ and $S_A(A, B) - S_B(A, B)$ for our female names, across our 42 LSA models. This suggests that dimensionality reduction increases prejudice partially by facilitating race-based name categorization: Ultimately, the most prejudiced models are also the ones that are able to strongly associate African American names with other African American names and White names with other White names, and these are the models with moderate dimensionality.

A final test involves looking at the similarities between LSA models trained on our two corpora. Does dimensionality reduction have a similar effect for both the LATWP and NYT articles? To test this, we examine the correlation between the $S_A(C, D) - S_B(C, D)$ for each dimension for the LATWP models and for the NYT models, and find a correlation of 0.58 ($p < 0.01$) for male names and 0.35 ($p = 0.11$) for female names. A similar test involving $S_A(A, B) - S_B(A, B)$ finds a correlation of 0.57 ($p < 0.01$) for male names and a correlation of 0.59 ($p < 0.01$) for female names. These positive correlations indicate that there are similarities in the ways in which dimensionality reduction influences representations in our two corpora.

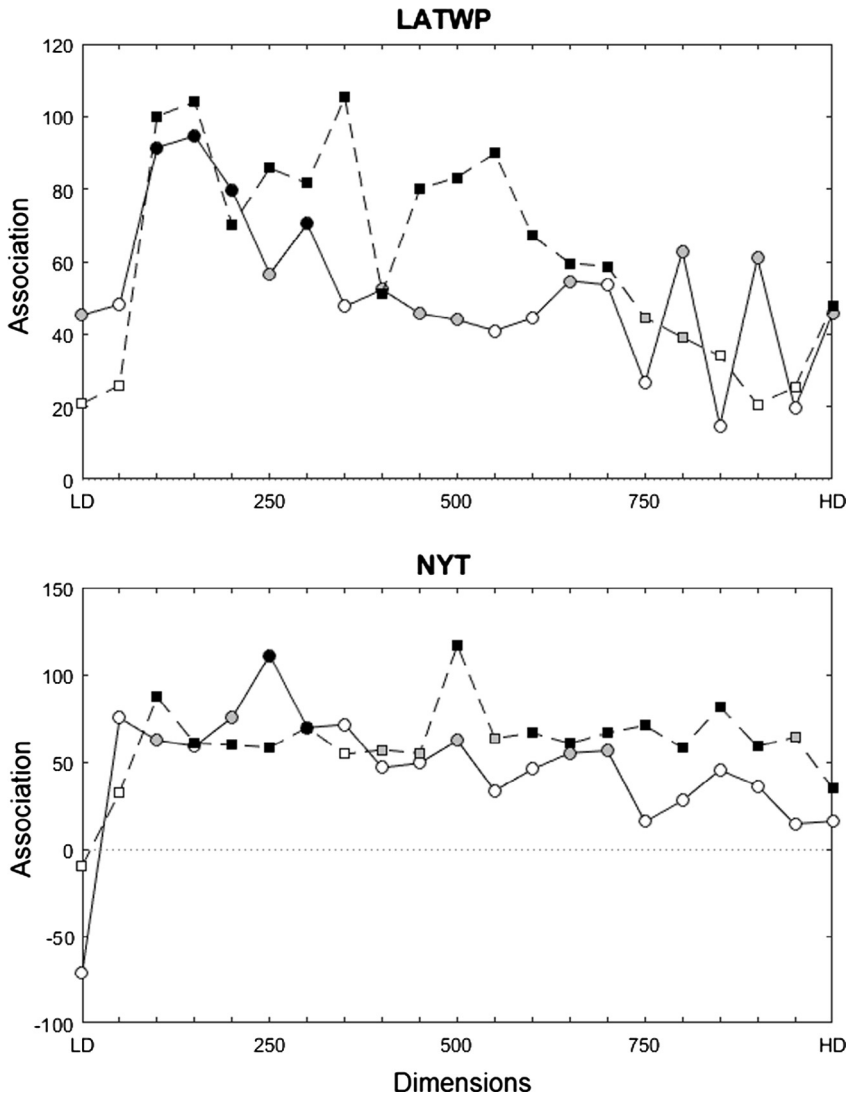

Fig. 2. Prejudice and success at racial categorization for female African American and White names, using stimuli from Greenwald et al. (1998). Again, positive values for the solid line/circles represent bias against African Americans, and positive values for the dashed line/squares represent success at race-based categorization. Black circles/squares indicate $p < 0.01$; grey circles/squares indicate $p < 0.05$; white circles/squares indicate $p > 0.05$.

3.2. Latino names vs. White names

We also tested for prejudice using stimuli from Pérez (2010). Pérez performed an IAT on a representative sample of US adults, with 10 Latino last names, 10 White last names, as well as 10 positively and 10 negatively valenced words. He found that, on aggregate, his participants displayed associative prejudices towards Latinos, and that this predicted immigration policy judgments and political ideology.

We find that our 300 dimensional LSA models do not display any significant biases across the two sets of names when evaluated using interaction effects regressions ($p = 0.08$ for LATWP and $p = 0.56$ for NYT). Results for other dimensions vary across corpora. For the LATWP corpus, there is a consistent bias associating Latino names with negatively valenced words and White names with positively valenced words, for dimensions lower than 300. Overall, 3 out of the 21 LATWP models display a significant bias ($p < 0.05$) against Latino names when evaluated with an interaction effect regression (the remaining models do not display a significant difference). Pooling all dimensions (and permitting random intercepts for dimensions), we do find an aggregate bias against Latinos ($\beta = 0.007$, $z = 2.17$, $p < 0.05$, 95% CI = [0.001, 0.013]), but this appears to be weaker than that observed in the previous section.

Fig. 3 shows these effects by plotting $S_A(C, D) - S_B(C, D)$ in solid lines, with A corresponding to the set of Latino names, B corresponding to the set of White names, C corresponding to the set of negatively valenced words, and D corresponding to the set of positively valenced words. Again $S_A(C, D) - S_B(C, D)$ captures the relative ordinal associations for the names, with positive values indicating a prejudice against Latinos. Statistical significance evaluated using the Wilcoxson rank-sum test is indicated using black, grey, or white circles. As above, these effects are stronger for LSA models with a moderate number of dimensions, with the magnitude of $S_A(C, D) - S_B(C, D)$ being largest for the LSA model with 50 dimensions.

Many of these patterns also emerge when C and D are replaced with A and B. In fact, all 22 of our LATWP models more strongly associate Latino names with other Latino names and White names with other White names, leading to positive values of $S_A(A, B) - S_B(A, B)$. As shown in Fig. 3, $S_A(A, B) - S_B(A, B)$ (indicated in dashed lines, with black, grey, or white squares for significance, based on

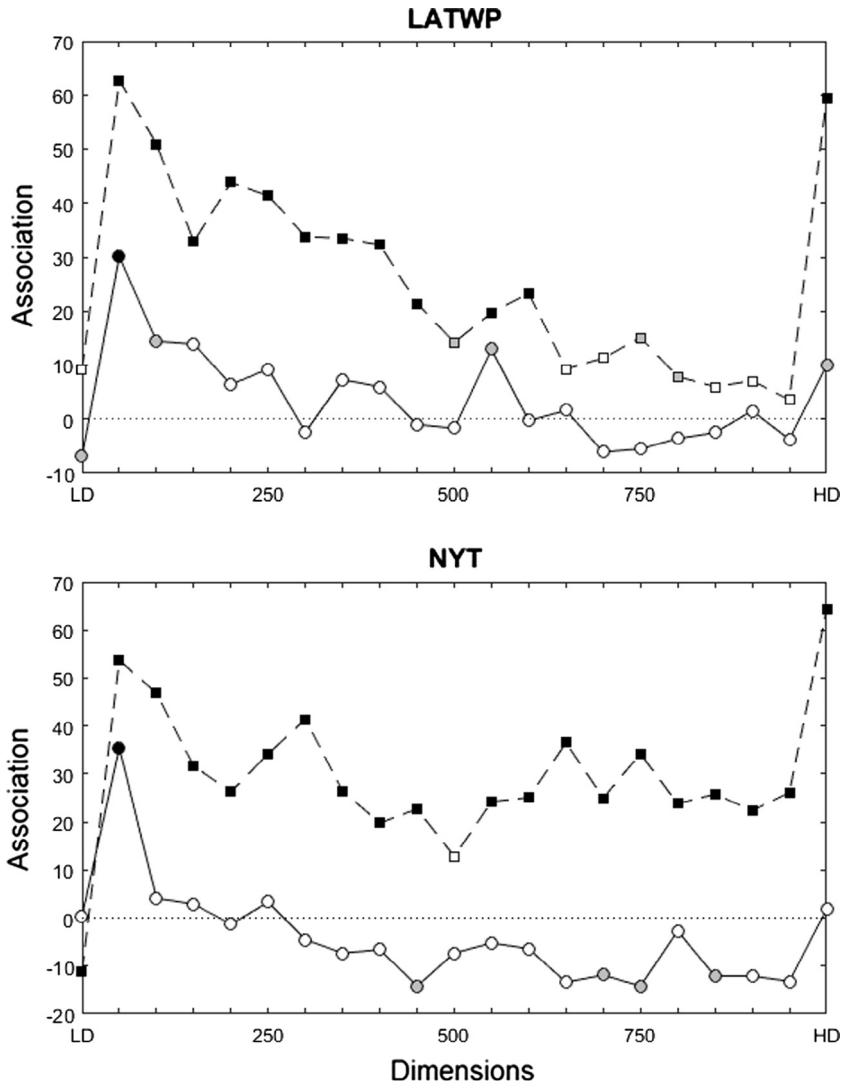

Fig. 3. Prejudice and success at ethnic categorization as a function of LSA dimensionality for Latino and White names, using stimuli from Pérez (2010). Here positive values for the solid line/circles represent bias against Latinos, and positive values for the dashed line/squares represent success at ethnicity-based name categorization. Black circles/squares indicate $p < 0.01$; grey circles/squares indicate $p < 0.05$; white circles/squares indicate $p > 0.05$.

the Wilcoxson rank-sum test) is most positive for the 50 dimensional model. Additionally, there is a correlation of 0.79 ($p < 0.01$) between $S_A(C, D) - S_B(C, D)$ and $S_A(A, B) - S_B(A, B)$, again suggesting that our LSA models are especially prejudiced when they are able to accurately categorize last names as being Latino or White.

These results are not as clear for the NYT corpus. Here one of the NYT models displays a significant bias ($p < 0.01$) against Latino names, and six of the models display a significant bias ($p < 0.05$) against White names, when evaluated using a regression interaction effect. Pooling all dimensions, and allowing for random intercepts for each dimension, we do not find an aggregate bias against either Latinos or Whites ($p = 0.77$) in an interaction effects regression.

The relative ordinal strength of associations for the different NYT models can be seen in Fig. 3, which plots $S_A(C, D) - S_B(C, D)$ in solid lines for each model (with significance, as evaluated by the Wilcoxson rank-sum test, shown in black, grey, and white circles). Here we do find that the largest magnitude of $S_A(C, D) - S_B(C, D)$ emerges for an LSA model with 50 dimensions. For this model $S_A(C, D) - S_B(C, D)$ is positive (and crosses our $p < 0.01$ threshold), again demonstrating a prejudice against Latino names. However, this bias weakens and then reverses for higher dimensions. Although the magnitude of $S_A(C, D) - S_B(C, D)$ when it is positive is much larger than the magnitude of $S_A(C, D) - S_B(C, D)$ when it is negative, negative values of $S_A(C, D) - S_B(C, D)$ do often cross our statistical significance thresholds.

When replacing C and D with A and B in order to study the success of our LSA models in associating Latino names with Latino names and White names with White names, we find that 21 of our 22 models do successfully make this categorization, by generating positive values of $S_A(A, B) - S_B(A, B)$ (shown with dashed lines in Fig. 3, with black, grey, or white squares for significance). However, the best performing model, with the most positive $S_A(A, B) - S_B(A, B)$, has no dimensionality reduction whatsoever. Indeed, unlike with the analysis of the LATWP corpus, the correlation between $S_A(C, D) - S_B(C, D)$ and $S_A(A, B) - S_B(A, B)$ on the NYT corpus is only 0.41, and does not reach statistical significance ($p = 0.06$). This suggests that some of the ambiguous results regarding prejudice observed for our LSA models on the NYT corpus could stem from their inability to accurately categorize Latino and White last names.

Lastly, how correlated are the different LSA models for our two corpora? Although LATWP and NYT do generate different results on aggregate, we find strong significant correlations across dimen-

sions for these two corpora. Particularly, there is a correlation of 0.81 ($p < 0.01$) for $S_A(C, D) - S_B(C, D)$ and a correlation of 0.68 ($p < 0.01$) for $S_A(A, B) - S_B(A, B)$ for each dimension for the LATWP models and for the NYT models. It seems thus that dimensionality reduction is having a similar effect across the two corpora: In both cases biases against Latinos are higher for moderate and low dimensions.

3.3. Discussion

We have examined whether LSA models, trained on the Los Angeles Times-Washington Post and New York Times news corpora, display prejudiced associations, the degree to which these associations depend on dimensionality, and whether dimensionality reduction generates prejudice by facilitating social categorization. Using names and words from the race IAT of Greenwald et al. (1998), we found that our LSA models do possess stronger associations between African American first names and negatively valenced words, and stronger associations between White first names and positively valenced words. These biases emerge consistently for both male and female names and emerge across a wide range of dimensions for both corpora. However, these biases are strongest for moderate dimensional LSA models. We also found that LSA models with moderate dimensionality have the strongest associations between different African American names and between different White names, and that the ability to successfully categorize names by race is strongly correlated with the degree of prejudice a model displays.

We also tested for these effects in a related IAT involving Latino and White last names. This IAT was performed by Pérez (2010), who found that a representative sample of US adults displayed associative prejudices against Latinos. Our LSA models mimicked the above findings for low and moderate dimensional models trained on the LATWP corpus. These dimensions were also the ones with the strongest associations between Latino names and other Latino names, and between White names and other White names, again demonstrating a close relationship between prejudice, dimensionality reduction, and successful social categorization.

These results did not, however, emerge as consistently for the NYT corpus. Here we did find prejudice against Latinos for some small dimensions, but higher dimensions generated a (weak) reversed effect. Additionally, dimensionality reduction did not help facilitate name categorization. One reason for these differences could be variation in the content of the LATWP and NYT articles, which may reflect diverging demographics for Los Angeles and New York City. It is also useful to note that Pérez's (2010) results were less robust than the results of Greenwald et al. (1998). The latter found that all their participants were prejudiced against African Americans, whereas Perez found that a substantial minority of his participants were prejudiced against Whites, with IAT-based prejudice being heavily correlated with preferences for immigration policy and political ideology. Note that although the results of our Latino vs. White names study do suggest that there are differences across corpora in terms of the absolute magnitude of prejudice, we found that the effect of dimensionality reduction was fairly constant across all corpora.

Ultimately, our results indicate that many of the prejudices documented by Greenwald et al. (1998) and Pérez (2010) can be found in natural language. Importantly, these prejudices rely critically on dimensionality reduction, a key component of the LSA framework. This is due to the beneficial effect of dimensionality reduction on social categorization: Moderate dimensional LSA models often have the strongest associations between names from the same racial or ethnic group. These are also models that most strongly associate White names with positive words and African American and Latino names with negative words.

4. Stereotypes

Thus far, we have considered prejudiced associations, in which names from different social categories are more or less likely to be associated with positively or negatively valenced words. In this section, we consider stereotypes, which are primarily semantic in nature, and which may not necessarily have a valence-based component. However, the properties of LSA that we propose govern the learning of stereotypes are identical to those discussed above, with regards to prejudice. Indeed, given the semantic basis of theories like LSA, and their prior successes across a variety of semantic tasks, it is possible that the biases these models display for stereotyping are even greater than the ones they display for prejudice.

In order to test the ability of our LSA models to learn and represent stereotypes, we consider two existing implicit association tests. Both utilize common male and female first names. The first test examines the associations between names and words related to power and weakness, and uses the names and words from Rudman et al. (2001). The second examines the strength of association of names with family-related words and career-related words, and uses names and words from Nosek et al. (2002).

4.1. Power and weakness

Our first stereotyping test uses stimuli from Rudman et al. (2001). Rudman et al. examine stereotypes pertaining to potency, that is, power and weakness, using an IAT task with 15 female names, 15 male names, 15 power-related words, and 15 weakness-related words. They find evidence for this stereotype in their participant pool, with participants responding more quickly when given "male or power" and "female or weakness" categories compared to "female or power" and "male or weakness" categories. Rudman et al. also note that this tendency is more pronounced for male participants, indicating gender differences in stereotyping.

We find that our 300 dimensional LSA models do display these stereotypes for the NYT dataset ($\beta = 0.029$, $t = 3.85$, $p < 0.01$, 95% CI = [0.014, 0.044]) they do not do so in a significant manner for the LATWP dataset ($p = 0.60$), when evaluated using interaction effect regressions Overall, however, 39 out of our 42 LSA models generate these associations (with 21 of these reaching statistical significance) with these regressions. An analysis that pools all of the dimensions, and permits random intercepts for the dimensionality of the model, also indicates a very strong stereotyping effect for both the LATWP corpora ($\beta = 0.011$, $z = 4.57$, $p < 0.01$, 95% CI = [0.006, 0.016]) and the NYT corpora ($\beta = 0.015$, $z = 5.49$, $p < 0.01$, 95% CI = [0.010, 0.020]).

As in the previous section, this tendency is most pronounced for LSA models with moderate dimensionality. This is shown in Fig. 4, which plots relative ordinal associations, $S_A(C, D) - S_B(C, D)$, in solid lines (with significance – as determined by the Wilcoxson rank-sum test – shown with black, grey, or white circles). Here A and B correspond to female and male names, and C and D correspond to weakness and power-related words, implying that positive values of $S_A(C, D) - S_B(C, D)$ correspond to female-weakness male-power stereotypes. Overall, the model with the strongest stereotyping, in terms of the magnitude of $S_A(C, D) - S_B(C, D)$, is the 450 dimensional model on the LATWP corpus, and the 400 dimensional model on the NYT corpus. The strength of stereotyping greatly weakens for the very low and very high dimensional models on the LATWP corpus, and actually reverses for these models on the NYT corpus.

Can our LSA models learn to represent gender as a social category? All of our 42 LSA models can, as shown in Fig. 4. This figure plots $S_A(A, B) - S_B(A, B)$ in dashed lines for each model. Positive val-

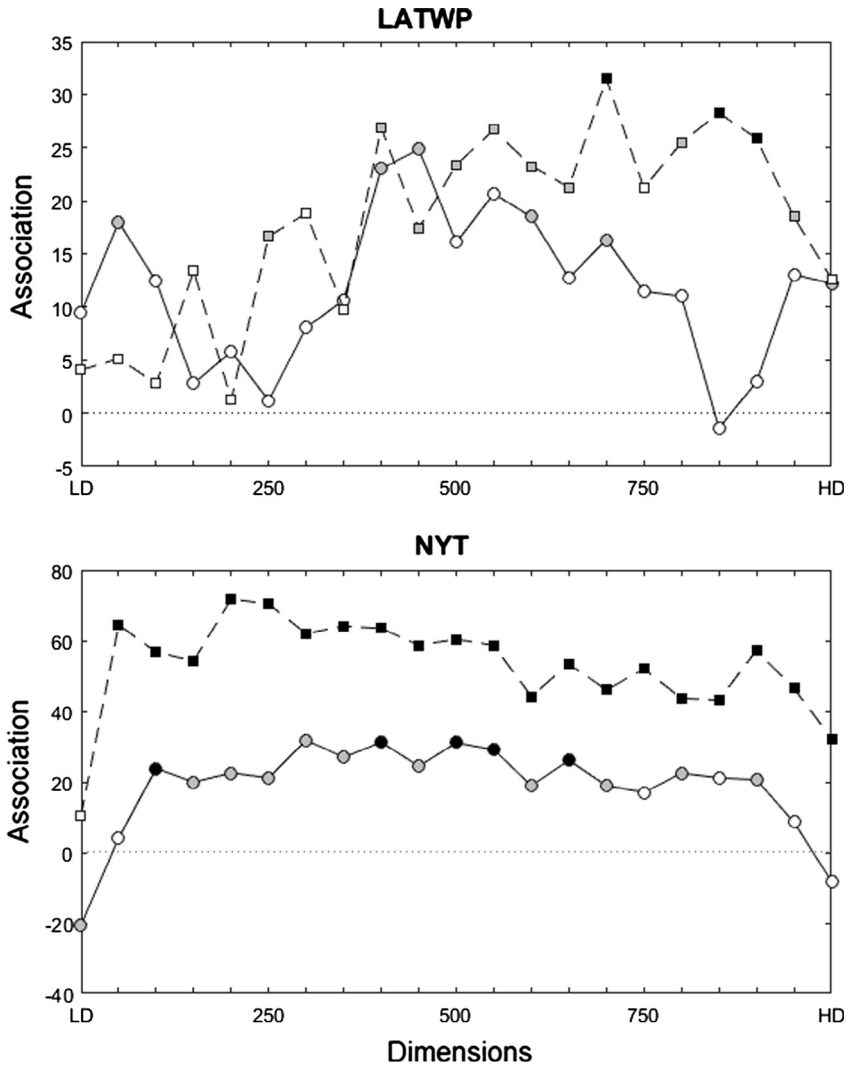

Fig. 4. Power/weakness stereotyping and success at gender categorization as a function of LSA dimensionality for female and male names, using stimuli from Rudman et al. (2001). Here, positive values for the solid line/circles represent weakness stereotypes for women, and positive values for the dashed line/squares represent success at gender-based categorization. Black circles/squares indicate $p < 0.01$; grey circles/squares indicate $p < 0.05$; white circles/squares indicate $p > 0.05$.

ues correspond to stronger associations between female names and other female names, and stronger associations between male names and other male names, and indicate an increased accuracy in categorizing names by gender. Statistical significance of the corresponding effect – as determined by the Wilcoxson rank-sum test – is displayed using black, grey, or white squares.

Once again, dimensionality reduction plays a role in facilitating social categorization. The best performing model, with the largest value of $S_A(A, B) - S_B(A, B)$, has 700 dimensions on the LATWP corpus and 200 dimensions on the NYT corpus. Indeed, there is a correlation of 0.57 ($p < 0.01$) between $S_A(C, D) - S_B(C, D)$ and $S_A(A, B) - S_B(A, B)$, suggesting that our LSA models are especially stereotyped when they are able to accurately categorize names as being female or male, and that this is facilitated by dimensionality reduction.

Unlike our previous tests we do not observe similar effects for dimensionality reduction across these datasets. LATWP and NYT corpora display correlations of only 0.10 ($p = 0.67$) for $S_A(C, D) - S_B(C, D)$ for each dimension considered, and correlations of −0.01 ($p = 0.94$) for $S_A(A, B) - S_B(A, B)$ for each dimension considered. Although both corpora display stronger effects for moderate dimensions compared to very large or very small dimensions, the actual dimensions with the strongest effects are slightly larger for LATWP compared to NYT.

4.2. Career and family

Our second test of stereotyping uses stimuli from Nosek et al. (2002), which examines stereotypes pertaining to career vs. family associations for women and men. Nosek et al. use an IAT task with 8 female names, 8 male names, 8 career-related words, and 8 family-related words, and find a robust association between female names and family-related words and male names and career-related words. Using their stimuli to test our 300 dimensional LSA models, we also obtain these associations ($\beta = 0.054$, $t = 2.56$, $p < 0.05$, 95% CI = [0.012, 0.097] for LATWP and $\beta = 0.118$, $t = 4.85$, $p < 0.01$, 95% CI = [0.069, 0.166] for NYT), when tested with a regression interaction effect. Overall, 41 out of our 42 models display this stereotype (with 25 models reaching statistical significance). A similar analysis on an aggregate level (controlling for dimensionality using random intercepts) also indicates a very strong gender-career stereotype for both the LATWP and the NYT corpora ($\beta = 0.043$, $z = 7.99$, $p < 0.01$, 95% CI = [0.032, 0.054] for LATWP and $\beta = 0.089$, $z = 13.48$, $p < 0.01$, 95% CI = [0.076, 0.102] for NYT).

Once again, dimensionality plays a crucial role, with the strongest stereotypes emerging for the 150 dimensional model on both the LATWP and the NYT corpora. Very high and very low dimen-

sional models do display some stereotyping, but this is weaker than that displayed by our more moderate dimensional models. This is illustrated in Fig. 5, which plots $S_A(C, D) - S_B(C, D)$, with A and B corresponding to female and male names, and C and D corresponding to family and career-related words, for each dimension. Statistical significance, as evaluated by the Wilcoxson rank-sum test, is shown with black, grey, and white circles.

As in the previous section, our models are also successfully able to categorize names based on gender (note that a separate test of this is necessary as Nosek et al. (2002) and Rudman et al. (2001) use different sets of male and female names). In fact, almost all of our 42 LSA models are able to perform this type of categorization and generate positive values of $S_A(A, B) - S_B(A, B)$. This can be seen in Fig. 5, which plots $S_A(A, B) - S_B(A, B)$ using dashed lines with black, grey, and white squares for statistical significance (as evaluated by the Wilcoxson rank-sum test). The model with the largest value of $S_A(A, B) - S_B(A, B)$ has 250 dimensions on the LATWP corpus and 150 dimensions on the NYT corpus. There is also a correlation of 0.56 ($p < 0.01$) between $S_A(C, D) - S_B(C, D)$ and $S_A(A, B) - S_B(A, B)$. This again shows that our LSA models display the strongest stereotypes when they are able to accurately categorize names as being female or male.

Finally, we observe significant similarities across our two corpora in terms of the effect of dimensionality reduction. Particularly, LATWP and NYT corpora display correlations of 0.73 ($p < 0.01$) for $S_A(C, D) - S_B(C, D)$ for each dimension considered, and correlations of 0.72 ($p < 0.01$) for $S_A(A, B) - S_B(A, B)$ for each dimension considered.

4.3. Discussion

This section tested whether our LSA models, trained on the Los Angeles Times-Washington Post and New York Times news corpora, displayed the power-weakness and career-family gender stereotypes observed in prior experimental work (Nosek et al., 2002; Rudman et al., 2001). Using stimuli from these experiments, we found that the LSA models consistently associated female names with weakness words and family words, and male names with power words and career words. Again, these biases were strongest for LSA models with moderate dimensionality. Moderate dimensional LSA models were also the best at categorizing names based on gender, and were able to strongly associate female names with other female names and male names with other male names. We also found strong correlations between stereotyping and suc-

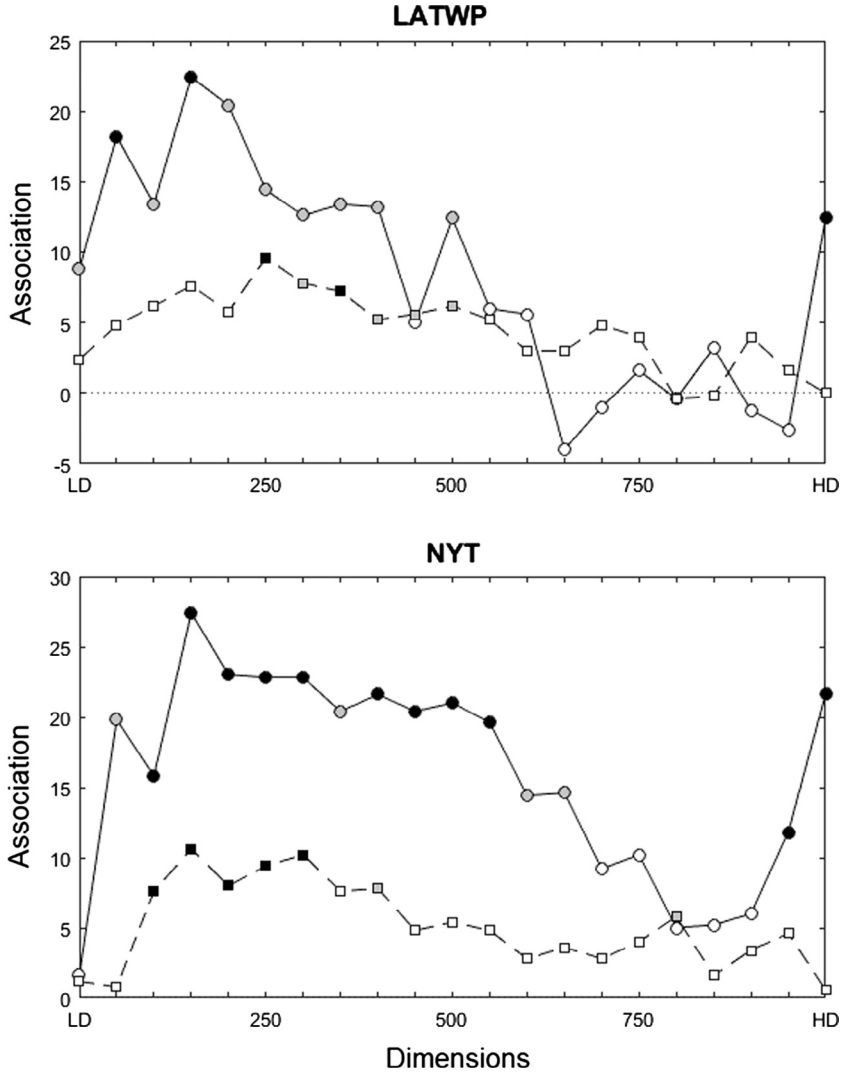

Fig. 5. Career/family stereotyping and success at gender categorization as a function of LSA dimensionality for female and male names, using stimuli from Nosek et al. (2002). Here, positive values for the solid line/circles represent family stereotypes for women, and positive values for the dashed line/squares represent success at gender-based categorization. Black circles/squares indicate $p < 0.01$; grey circles/squares indicate $p < 0.05$; white circles/squares indicate $p > 0.05$.

cessful name categorization, suggesting that dimensionality reduction generates stereotyping in part by facilitating better name categorization. Once again, these results suggest that the types of gender stereotypes documented in experimental work (Nosek et al., 2002; Rudman et al., 2001) are also present in natural language. As with prejudice, stereotyping relies critically on dimensionality reduction, due to the beneficial effect of dimensionality reduction on gender-based categorization.

5. General discussion

Racial, ethnic, and gender-based discrimination is a very important topic of inquiry in the social and behavioral sciences. Psychologists interested in these issues have, over the past decades, attempted to characterize the cognitive underpinnings of this phenomenon, and have discovered the importance of associations in generating prejudice and stereotypes for members of different social categories (Allport, 1954; Devine, 1989; Fazio & Olson, 2003; Greenwald & Banaji, 1995). For example, experimental tasks such the implicit association test have shown that many individuals display stronger associations between African American names and negatively valenced words, and stronger associations between White names and positively valenced words (Greenwald et al., 1998), as well as stronger associations between female names and weakness-related words, and stronger associations between male names and power-related words (Rudman et al., 2001). Similar prejudices and stereotypes have also been documented for a variety of other social categories, including immigrants, Muslims, the aged, and the disabled (e.g. Nosek et al., 2009).

Although there have been many insightful attempts to study how the above associations are learnt and represented, there is a gap between research in this area and cognitive science research the representation of more general concepts and categories. In this paper we attempted to bridge this through the use of distributional theories of semantic memory (Griffiths et al., 2007; Jones & Mewhort, 2007; Kwantes, 2005; Landauer & Dumais, 1997; Lund & Burgess, 1996; Mikolov et al., 2013; Pennington et al., 2014). Particularly we applied latent semantic analysis (Landauer & Dumais, 1997; Landauer et al., 1998) to a large dataset of Los Angeles Times-Washington Post and New York Times news articles. We then gave our trained LSA models five existing implicit association tests, pertaining to prejudices for male African Americans vs. Whites, female African Americans vs. Whites, and Latinos vs. Whites, and power-weakness and career-family stereotypes for women vs. men. As with participants in prior experiments, we found that our trained LSA models often display race, ethnicity, and gender-based prejudices and stereotypes. This suggests that the biased associations that characterize these prejudices and stereotypes are likely to emerge when semantic memory models are applied to the information present in everyday language, such as language used in news media (see also Lenton et al. (2009) and Lynott, Kansal, Connell, and O'Brien (2012) for similar results with different stimuli and different language datasets).

5.1. Social categorization and the role of dimensionality

One important result obtained from our analysis involves the cognitive mechanisms that generate these biases. We find that the magnitude of prejudice and stereotyping in our LSA models is at its maximum for LSA models with moderate dimensionality, and often weakens or disappears for models without dimensionality reduction. For example, although our corpora may present some African American names in negative contexts, and some White names in positive contexts, these co-occurrences are not sufficient by themselves to generate a strong prejudice against African-Americans. It is the fact that LSA applies dimensionality reduction to word-context co-occurrence data that leads to negative associations for some African Americans being generalized to most African-Americans and positive associations for some Whites being generalized to most Whites.

Why does dimensionality reduction facilitate prejudice? We tested this by examining the degree to which African American and White names were associated with other names in their own social category. Overall, we found that the ability of our models to associate African American names with other African American names and White names with other White names, was at its maximum for models with moderate dimensionality. Additionally, the success of an LSA model of a given dimensionality in categorizing names based on race was strongly correlated with the degree of prejudice that it displaced.

These results suggest that moderate dimensionality strengthens prejudice by facilitating social categorization. LSA models with dimensionality reduction possess stronger associations between names within a group, and thus apply knowledge (in this case, prejudiced associations) learnt for some of the members of a group to the remaining members of the group. High dimensional models do not have these cohesive associations, and thus are unable to generalize prejudice. Too much dimensionality reduction also reduces prejudice: Very low dimensional LSA models fail to distinguish social categories, and subsequently generalize too much. Indeed, our trained LSA models with only two underlying dimensions are often not prejudiced at all.

This mechanism is also at play for the learning of stereotypes. Moderate levels of dimensionality (unlike very high or very low dimensionality) are ideal for learning gender based social categories, and for associating female names with other female names, and male names with other male names. It is ultimately due to these categories that the co-occurrence of some female names with family or weakness-related words, and the co-occurrence of some male name with career or power-related words, gets generalized to most female and male names.

Dimensionality reduction is crucial features of LSA, and is often considered necessary for the efficient learning of word meaning and association. Indeed, prior work has found that LSA models do best in word-similarity and semantic categorization tasks when they have a moderate number of dimensions (Landauer & Dumais, 1997; Landauer et al., 1998). Likewise, the value of approaches like LSA for sentiment analysis and valence-based word categorization stems primarily from their ability to reduce the dimensionality of the word-context co-occurrence space, and subsequently learn second, third, and higher degrees of word-valence associations (Bestgen & Vincze, 2012; Maas et al., 2011; Recchia & Louwerse, 2015). Ultimately, the generalization necessary to adequately infer word meaning and valence from natural language appears to be the same type of generalization at play in learning social categories, and in turn, learning prejudices and stereotypes. This relationship does not mean that prejudice and stereotyping is desirable in any manner. Social categorization and generalization have profoundly negative consequences for how we organize our societies, and the types of associations possessed by our LSA models undoubtedly represent the darker side of human cognition. That said, it does suggest that the associations involved in prejudice and stereotyping are an outgrowth of the otherwise beneficial cognitive mechanisms that people possess for representing knowledge in everyday environments.

5.2. Bias

Everyday language environments, such as the news corpora studied in this paper, play a crucial role in the learning of prejudiced and stereotyped representations. Children and adults

exposed to this language will not only learn to distinguish names based on their race and gender, but will also develop strong associations between names of a given race or gender and various positively or negative valenced words and various stereotype-related words. This is one reason why we observed similarities between the behavior observed in prior IAT experiments and the behavior generated our LSA models.

Of course we cannot rigorously justify this type of causal claim. Everyday language it itself determined by social attitudes and beliefs. Thus the correlations we observe may merely be a product of peoples' existing associations (of course these correlations could also be due to some other third factor, and not necessary reflect any bias on the part of the writers and editors of the newspapers we consider). Regardless of causality, our analysis indicates that the structure of word co-occurrence in environments like LATWP and NYT reflects prejudice and stereotyping. These language environments present different races, ethnicities, and genders in systematically different contexts, so that semantic learning mechanisms possessed by humans develop biased associations when exposed to these language environments.

One related issue involves whether or not observed behavior with the IAT reflects actual prejudices and stereotypes, and whether this type of prejudice and stereotyping is reasonable. There is debate on this (see e.g. Arkes & Tetlock, 2004; Fiedler, Messner, & Bluemke, 2006) and addressing this debate is outside the scope of this paper. However, this paper does show that the types of associations documented by the IAT in humans, emerge naturally with common semantic memory models. Thus, in some sense, prejudiced and stereotyped associations do not need any type of specialized mechanisms in order to be learnt. Although it is certainly possible that individuals process information about different races, ethnicities or genders differently, the type of behavior observed with the IAT can be explained solely by an unbiased semantic memory mechanism applied to everyday language environments.

5.3. Theoretical extensions

The tests in this paper have relied on latent semantic analysis, which is one of the simplest and most influential distributional models of semantic memory. However, it is not the only such model. There are many other related approaches for studying the learning of semantic representations (Griffiths et al., 2007; Jones & Mewhort, 2007; Kwantes, 2005; Lund & Burgess, 1996; Mikolov et al., 2013; Pennington et al., 2014), most of which also represent words as vectors in multidimensional spaces, and perform some type of dimensionality reduction in order to generalize word meanings and associations. Due to these similarities, the insights of this paper should apply broadly and we should recover equally strong racial, ethnic, and gender prejudices and stereotypes were we to use some of these alternate techniques for building semantic representations.

It may also be useful to move beyond simple word vectors and examine representations that can accommodate context. The context in which a word or name appears provides valuable information about the meaning of the word or the name. However, LSA relies on a bag-of-word format which represents each article only in terms of the frequency of its components. Word order as well as sentence-level and paragraph-level compositional and contextual details are ignored. This also the case for more sophisticated approaches (e.g. Jones & Mewhort, 2007; Mikolov et al., 2013) which are sensitive to word order but nonetheless recover representations only for individual words.

Recent work suggests a number of techniques for accommodating context and building semantic representations for sentences and paragraphs. Many of these techniques involve sophisticated operations for combining vector representations for individual words, whereas others provided broader theoretical frameworks for studying the emergence of meaning in large pieces of text (see e.g. Kintsch, 1988, 2001; Mitchell & Lapata, 2010). These approaches have primarily been applied to the study of non-social representations, but can easily be extended to examine social categorization and its relationship with prejudice and stereotyping. This is a valuable topic for future work.

Another extension to our approach involves using theories of distributional semantics to explain well-known qualitative findings regarding the learning of prejudiced and stereotyped group-based associations. For example, an important result in social judgment research involves illusory correlations, according to which individuals believe relationships between categories that are actually independent due to differences in the number of observations for these categories or due to prior expectations for these categories (Hamilton & Gifford, 1976; Hamilton & Rose, 1980). Other work finds that beliefs about social category membership affect how people evaluate differences on features that determine these categories, how people evaluate stereotype-consistent and inconsistent information about these categories, and how people evaluate the degree of variation in the features of these categories (Hewstone, 1994; Judd & Park, 1988; Tajfel, 1969). These findings, and others, have been successful explained by cognitive and statistical models involving exemplar memory, recurrent neural networks, and tensor products (Fiedler, 1996; Van Rooy, Van Overwalle, Vanhoomissen, Labiouse, & French, 2003), however none of these models possess the types of natural representations that are involved in prejudice and stereotyping in the real world. The approach presented in this paper illustrates one way of uncovering these representations, and when combined with the above types of cognitive models, can be used to build theories of social judgment that are not only able to describe the qualitative patterns of associations observed in judgments, but also quantitatively predict the magnitude of these associations, and the ways in which these associations relate to real-world information environments.

Relatedly, models of semantic representation, such as those studied in this paper, can be used to provide a cognitive account of the learning of various social perception biases, such as the better-than-average effect (e.g. Chambers & Windschitl, 2004). Such biases are often explained using motivational reasoning, but others have argued that they are the natural consequences of biased social samples (Denrell, 2005; Galesic, Olsson, & Rieskamp, 2012). Again, the approach outlined in this paper can be used to formally specify how biased social samples lead to biases is self and other knowledge. Such an application would not only describe the qualitative patterns underlying social perception, but also quantitatively predict these patterns. Examine this relatively novel way of modelling social judgment appears to be a promising avenue for future work.

Acknowledgement

Funding was received from the National Science Foundation grant SES-1626825.

Appendix A

In this appendix we provide additional details regarding the statistical methods used to evaluate differences in associations across groups of names. As an illustration we consider two female names (*Lisa* and *Elaine*), two male names (*Daniel* and *Peter*), two power-related words (*command* and *triumph*), and two weakens related words (*fragile* and *timid*), taken from Rudman et al.'s (2001) power-weakness stereotypes test.

Table A1
Associations between various male and female names and various power-related and weakness-related words.

Word	Power word	Name	Male name	Association
fragile	0	*Lisa*	0	−0.0191
timid	0	*Lisa*	0	0.0617
command	1	*Lisa*	0	−0.0261
triumph	1	*Lisa*	0	−0.0157
fragile	0	*Elaine*	0	0.0468
timid	0	*Elaine*	0	0.0033
command	1	*Elaine*	0	−0.0936
triumph	1	*Elaine*	0	0.0556
fragile	0	*Daniel*	1	0.0025
timid	0	*Daniel*	1	−0.0904
command	1	*Daniel*	1	0.0169
triumph	1	*Daniel*	1	0.0352
fragile	0	*Peter*	1	−0.0351
timid	0	*Peter*	1	−0.0062
command	1	*Peter*	1	−0.0422
triumph	1	*Peter*	1	0.0849

The analysis in this paper uses cosine similarity on the relevant corpus to measure the associations between different names and words. For the example here, we consider the 300 dimensional LSA model trained on the NYT corpus. The model gives associations, shown in Table A1. Our first test examines applies standard parametric methods to evaluate whether male or female names are closer to power or weakness words. This involves regressing the *association* variable in Table A1, on the *power_word* and *male_name* variables, as well as their interaction (*power_word*∗*male_name*). A positive interaction effect would indicate that associations are higher when the name in consideration is male and the word in consideration is a power-related. For the simple setting considered here we do find a positive interaction effect ($\beta = 0.099$, $t = 2.02$, $p = 0.07$, 95% CI = [−0.008, 0.077]), though of course this test is not sufficiently powered to evaluate significance. In the main text of the paper we apply this test to pairwise associations between all names and all words used in a given IAT stimuli set.

Our second test uses non parametric methods to measure the strength of this effect. Such methods are necessary as the magnitude of the interaction effect coefficient (e.g. $\beta = 0.099$ above) reflects not only relative associations, but also distances in the corresponding LSA space. It thus cannot be used to compare the strength of the biases across LSA models with different dimensionalities, which is the primary focus of this paper.

The nonparametric methods we use involve examining the ordinal distances between a given name and the various sets of words. This method first transforms the list of associations between the name and the set of words into rankings, before calculating the total ranking of a one of the set of words. To maintain consistency with the notation used in the main text, we refer to the set of weakness-related words in our example as C and the power-related words in our example as D, and the average ranking of words in C in terms of distances to name i, as $s_i(C, D)$. Thus, for example, the associations between *Lisa* and *command*, *triumph*, *fragile*, and *timid* are −0.026, −0.015, −0.019, and 0.062. When transformed into ranks, these are 1, 3, 2, and 4, respectively. Subsequently we obtain the total rank of words in C, $s_{Lisa}(C, D) = 2 + 4 = 6$. Performing this exercise for the remaining names gives us $s_{Elaine}(C, D) = 5$, $s_{Daniel}(C, D) = 3$, and $s_{Peter}(C, D) = 5$.

After this, our method averages the values of $s_i(C, D)$ for the names in the two groups. Again, to maintain consistency with the text, we refer to the set of female names as A and the set of male names a B. Subsequently the average association of female names with weakness words relative to power words, $S_A(C, D) = \text{AVE}_{i \text{ in } A}[s_i(C, D)] = 5.5$, and the average association of male names with weakness words relative to power words, $S_B(C, D) = \text{AVE}_{i \text{ in } B}[s_i(C, D)] = 4$.

The final step in our method involves comparing these average associations against each other to obtain $S_A(C, D) - S_B(C, D)$. Positive values of this measure indicate that names in A are more associated with words in C relative to D, when compared to names in B, whereas negative values of this measure indicate the opposite. Here we obtain $S_A(C, D) - S_B(C, D) = 1.5$, indicating that female names are more associated with weakness word and male names are more associated with power words.

Now we also wish to perform statistical tests to examine whether $s_i(C, D)$ varies for i in A vs. i in B, that is, whether the values of $S_A(C, D)$ and $S_B(C, D)$ are significantly different. As $s_i(C, D)$ are not distributed according to the normal distribution, we also have to use nonparametric tests to examine the statistical significance of differences in $s_i(C, D)$ for names in A vs. names in B. We do this with the Wilcoxson rank-sum test, which is a non-parametric alternative to the t-test. Applying this test to the values of $s_i(C, D)$ in this example involves comparing the list of numbers $[s_{Lisa}(C, D), s_{Elaine}(C, D)] = [6, 5]$ with the list $[s_{Daniel}(C, D), s_{Peter}(C, D)] = [3, 5]$, to see if the former list is larger.

Finally, an important supplementary analysis involves testing the strength of association between names in a group with other names in the same group. This analysis calculates $s_i(A, B)$ for a name i, with that name removed from the set A or B (to ensure that the strong association between name i and itself doesn't influence the measured association). Again, referring to the set of female names as A and the set of male names as B, $s_{Lisa}(A, B)$, $s_{Elaine}(A, B)$, $s_{Daniel}(A, B)$, and $s_{Peter}(A, B)$ would measure whether each of the names is closer to other female names or other male names (for example, $s_{Lisa}(A, B)$ would measure the relative rank of the association between *Lisa* and *Elaine*, compared to the associations between *Lisa* and *Daniel* and *Lisa* and *Peter*). Subsequently, $S_A(A, B) - S_B(A, B) = \text{AVE}_{i \text{ in } A}[s_i(A, B)] - \text{AVE}_{i \text{ in } B}[s_i(A, B)]$ would measure whether female names are closer to female names relative to male names, with positive values of $S_A(A, B) - S_B(A, B)$ corresponding to successful gender-based categorization. As above, a Wilcoxson rank-sum test can be applied to compare the list $[s_{Lisa}(A, B), s_{Elaine}(A, B)]$ with the list $[s_{Daniel}(A, B), s_{Peter}(A, B)]$ to evaluate statistical significance.

References

Allport, G. W. (1954). *The nature of prejudice*. Basic Books.
Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "would Jesse Jackson 'fail' the Implicit Association Test?". *Psychological Inquiry, 15*(4), 257–278.
Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods, 44*(4), 998–1006.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39* (3), 510–526.

Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: The role of nonmotivated factors in above- average and comparative-optimism effects. *Psychological Bulletin, 130*, 813–838.

Crabb, P. B., & Bielawski, D. (1994). The social representation of material culture and gender in children's books. *Sex Roles, 30*(1–2), 69–79.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12*(2), 163–170.

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review, 112*(4), 951–978.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18.

Dixon, T. L., & Linz, D. (2000). Overrepresentation and underrepresentation of African Americans and Latinos as lawbreakers on television news. *Journal of Communication, 50*(2), 131–154.

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology, 22*(1), 22–37.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*(1), 62–68.

Entman, R. M. (1992). Blacks in the news: Television, modern racism and cultural change. *Journalism & Mass Communication Quarterly, 69*(2), 341–361.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*(6), 1013–1027.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*(1), 297–327.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. *Psychological Review, 103*(1), 193–214.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology, 17*(1), 74–147.

Firth, J. R. (1957). *Papers in linguistics*. London, England: Oxford University Press.

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, 23–30.

Galesic, M., Olsson, H., & Rieskamp, J. (2012). Social sampling explains apparent biases in judgments of social environments. *Psychological Science, 23*(12), 1515–1523.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12*(4), 392–407.

Hamilton, D. L., & Rose, T. L. (1980). Illusory correlation and the maintenance of stereotypic beliefs. *Journal of Personality and Social Psychology, 39*(5), 832–845.

Harris, Z. S. (1954). Distributional structure. *Word, 2*, 146–162.

Hewstone, M. (1994). Revision and change of stereotypic beliefs: In search of the elusive subtyping model. *European Review of Social Psychology, 5*(1), 69–109.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1–37.

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer & J. T. Townsend (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254).

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology, 54* (5), 778.

Kahn, K. F., & Goldenberg, E. N. (1991). Women candidates in the news: An examination of gender differences in US Senate campaign coverage. *Public Opinion Quarterly, 55*(2), 180–199.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163–182.

Kintsch, W. (2001). Predication. *Cognitive Science, 25*(2), 173–202.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review, 12*, 703–710.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284.

Lenton, A. P., Sedikides, C., & Bruder, M. (2009). A latent semantic analysis of gender stereotype-consistency and narrowness in American English. *Sex Roles, 60*(3–4), 269–278.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28* (2), 203–208.

Lynott, D., Kansal, H., Connell, L., & O'Brien, K. S. (2012). Modelling the IAT: Implicit Association Test reflects shallow linguistic environment and not deep personal attitudes. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 1948–1953). Austin, TX: Cognitive Science Society.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (Vol. 1, pp. 142–150). Association for Computational Linguistics.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37*(5), 435–442.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, 34*, 1388–1429.

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6*(1), 101–115.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., ... Kesebir, S. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences, 106*(26), 10593–10597.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12*(5), 413–417.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105*(2), 171.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: global vectors for word representation. *EMNLP* (Vol. 14, pp. 1532–1543). . October.

Pérez, E. O. (2010). Explicit evidence on the import of implicit attitudes: The IAT and immigration policy judgments. *Political Behavior, 32*(4), 517–545.

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology, 68*(8), 1584–1598.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*. .

Rudman, L. A., Greenwald, A. G., & McGhee, D. E. (2001). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. *Personality and Social Psychology Bulletin, 27*(9), 1164–1178.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*(3), 220–247.

Tajfel, H. (1969). Cognitive aspects of prejudice. *Journal of Biosocial Science, 1*(S1), 173–191.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial intelligence Research, 37*(1), 141–188.

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., & French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review, 110*(3), 536–563.