

Think Political Leader – Think He or She?

A Multi-Country Test of a Gender Fair Language Intervention for Stereotype Bias

Word Count: 8,041 (including references)

Abstract

Women are greatly underrepresented in positions of political leadership around the world. In seeking to explain this underrepresentation, some researchers have pointed to people's tendencies to stereotype leaders as more similar to men than women as these tendencies can support the belief that women are unsuited to leadership. This project aims to test whether a subtle linguistic intervention is able to ameliorate this gender bias in political leadership stereotypes across different languages and national contexts (minimum anticipated N country = 15, N sample = 3,600). Specifically, this project will examine whether gender fair language (e.g., the use of paired pronouns 'he or she') can reduce the tendency for people to stereotype political leaders as more similar to men than women. To increase the rigor with which these stereotypes are measured, this project will complement the dominant 'cheap talk' measure of stereotype content with a novel incentivised measure of this content. This project will support an understanding of the capacity for gender fair language interventions to shift gender bias in the political leadership domain in different regions of the world.

Word count: 179

Think Political Leader – Think He or She?

A Multi-Country Test of a Gender Fair Language Intervention for Stereotype Bias

Political systems are almost universally led by men. Currently, only one country in about every eight has elected a woman as Head of State or Government, and across the world women make up only one in every four parliamentarians (UN women, 2022). Some researchers have suggested that it may be possible to improve the representation of women in politics and other positions of power by changing the stereotypes that people hold about leaders (Eagly, Nater, Miller, Kaufmann & Sczesny, 2020; Eagly & Wood, 2002; Ellemers, 2018). This claim rests on an extensive body of evidence that people tend to stereotype leaders as more similar to men than women and to believe that men are therefore more suited to leadership than women (Conway, Pizzamiglio & Mount, 1996; Eagly & Karau, 2002; Heilman, 1983, 2001; Schein, 1973, 1975). If the degree of gender bias in political leader stereotypes can be reduced, the reasoning goes, women may be more likely to seek such positions and those who do may be less likely to face gender-based discrimination (e.g., Gaucher, Friesen & Kay, 2011; Heilman, 2012; Koch, D’Mello & Sackett, 2015). This reasoning highlights the importance of identifying potential interventions and testing their ability to causally affect gender bias in political leader stereotypes.

In this project, we aim to test whether a subtle linguistic intervention is able to ameliorate gender bias in political leadership stereotypes across a variety of languages and cultures. Specifically, we will examine whether exposing people to feminised gender fair language when asking them to describe political leaders (e.g., by using the paired pronouns ‘he or she’) can reduce their tendency to stereotype political leaders as more similar to men than women. In the process, we aim to build on the existing literature in three important ways. First, the existing work on the impact of gender fair language has primarily focused on

the ability of gender encoded language (e.g., the pronoun “she”) to increase the salience of the relevant gender category (e.g., women) in working memory. Little, if any, of this work has examined the impact of gender fair language on the content of people’s stereotypes. By providing a causal test of the impact of our gender fair language intervention on gendered leadership stereotypes we will inform a richer understanding of the implications of people’s everyday language for social cognition.

Second, much of the existing work on gender fair language has examined a relatively narrow set of languages and contexts (most notably English), raising questions of external validity. To assess the extent to which the impact of gender fair language on gender bias in the political leader stereotype generalizes across different languages and national contexts, we will conduct the broadest test of a gender fair language intervention to date. Third, the non-incentivised approach to measuring stereotype content that dominates the existing literature on stereotypes is vulnerable to response biases, raising questions of construct validity. To address this vulnerability, we will introduce a novel, incentivised, measure of stereotype content. By adopting these methodological improvements, this project can be expected to produce robust findings that support advancements in theory and practice.

Gendered Leadership Stereotypes

In this section, we will review the literature that supports our claim that people tend to stereotype leaders as more similar to men than women. Stereotypes are the beliefs that people hold about the attributes that the members of social categories actually have or should ideally have (Hilton & von Hippel, 1996). Reflecting the salience of gender as a social category around the world, there is evidence that people have widely shared and consensual stereotypes about the attributes of men and women (Eagly & Wood, 2012; Hentschel et al., 2019). Importantly, there are differences in the content of the stereotypes about each gender. Specifically, people tend to believe that women are more communal than men, exhibiting

more kindness, sympathy and sensitivity; they also tend to believe that men are more agentic than women, exhibiting more independence, competitiveness and determination (Bem, 1981; Eagly, Nater, Miller, Kaufmann & Sczesny, 2020; Eagly & Wood, 2002; Ellemers, 2018; Prentice & Carranza, 2002; Williams & Best, 1982).

If we turn to the stereotypes of leaders, there is evidence that leadership roles are associated with, and believed to require, the agentic attributes that are more typically granted to men than women (e.g., Heilman, 1983; Eagly & Karau, 2002; Froehlich, Olsson, Dorrrough & Martiny, 2020). For instance, Powell, Butterfield and Parent (2002; see also Powell & Butterfield, 1979, 1989) asked US undergraduate and MBA students to rate a good manager on 10 communal attributes (e.g., affectionate, warm, tender) and 10 agentic attributes (e.g., independent, forceful, dominant) and found that 55 percent of participants described managers as more agentic than communal. A mere four percent of participants showed the reverse pattern, describing managers as more communal than agentic. There is also evidence that US participants consider the attribute “masculine” to be more characteristic of managers than the attribute “feminine” (Gatton, DuBois & Faley, 1999; see also Beggs & Doolittle, 1993; Sastre, Fougereau, Igier, Salvatore & Mullet, 2000).

Consistent with the above findings, when researchers have measured the stereotypes of men, women and leaders at the same time, they have observed a greater degree of overlap between the stereotypes of leaders and those of men than between the stereotypes of leaders and those of women (Eagly & Karau, 2002; Heilman, 1983). In one of the most influential demonstrations of this phenomenon, Schein (1973, see also Schein, 1975; Brenner, Tomkiewicz & Schein, 1989) asked 300 US middle managers to rate 92 attributes according to how characteristic they were of women in general, men in general or successful middle managers. Schein observed that the ratings that participants gave for managers were more similar to the ratings that they gave to men (relative to women) on 70 percent of the

attributes. The ratings that participants gave for managers were only more similar to the ratings they gave for women (relative to men) on eight percent of attributes. On the basis of these findings, Schein concluded that when people ‘think manager’, they ‘think male’. In the intervening decades, Schein’s think manager–think male paradigm has been applied to leaders in a range of occupational settings (e.g., business, military, education), and a meta-analysis of 40 of these studies found that the stereotypes of leaders are, on average, around two to three times more similar to the stereotypes of men than the stereotypes of women (Koenig, Eagly, Mitchell & Ristikari, 2011).

Together, these findings provide a strong basis for expecting that when people think of leaders in the political domain, they are likely to think male (not female). Indeed, if anything, the association between the stereotypes of leaders and men in the political domain may be especially strong. The competitive nature of the political system and the role of the media in mediating between political leaders and their public arguably make a politician’s ability to exercise agency by seizing and exercising power an essential competency (e.g., Okimoto & Brescoll, 2010; see also Schneider & Bos, 2013). Further, the ability of a person to engage in (verbal) combat and vigorous self-promotion and, ultimately, to dominate their opponents is likely to be seen as central to the acquisition and retention of political leadership positions (Lombard, Azpeitia & Cheryan, 2021). For all these reasons, people’s stereotypes of political leaders can be expected to be more similar to those of men than those of women. This expectation provides the basis for our first hypothesis:

H1: The positive association between people’s stereotypes of men and political leaders will be stronger than the association between their stereotypes of women and political leaders.

Gender fair Language and Gendered Leadership Stereotypes

In this project, we aim to test whether the use of gender fair language (i.e., a strategy of treating men and women symmetrically in language; Sczensny, Formanowicz & Moser, 2016) may ameliorate gender bias in people's stereotypes of political leaders. Our expectation that gender fair language may reduce gender bias is based on evidence that, first, gender fair language can increase the cognitive accessibility of female exemplars and that, second, counter-stereotypical exemplars can reduce bias in stereotypes. We will review both sets of evidence in turn.

Research that has examined the impact of gender fair language has grown out of a broader programme of work that has sought to understand whether gender encoding in everyday language has consequences for gender equality in society. The starting point for this work is the observation that all languages encode for gender, whether through the lexical meaning of words or grammatical structure (Sczensny et al., 2016; Stahlberg, Braun, Irmen and Sczesny, 2011). Although all languages encode for gender, there are three general ways in which they can do so. Specifically, some languages (i.e., *neutral* gender languages, like Chinese and Hungarian) only encode gender in the lexical meaning of certain nouns (e.g., the loving *wife*). Other languages (i.e., *natural* gender languages, like English and Swedish) additionally encode gender in the lexical meaning of the corresponding personal pronouns (e.g., the loving *wife*, *she*...). Finally, *gendered* languages (like Spanish and Hebrew) encode gender grammatically in most nouns as well as in the articles, pronouns and adjectives that accompany them (e.g., *la esposa amorosa*, *ella*...).

The habitual encoding and decoding of gender in everyday communication can be expected to increase the salience of gender in working memory. When this encoding is asymmetrical, such that one gender is encoded more frequently than another, it may contribute to gender biases (Stahlberg et al., 2011; see also Sczesny et al., 2016). One such

asymmetry is the use of the masculine generic, which is when a word with masculine lexical meaning is used to describe mixed gender groups or people whose gender is unknown or irrelevant (Bodine, 1975; Hellinger & Bußmann, 2003). In English, this can be seen in the use of the noun “man” and pronoun “he” to refer to humankind (e.g., Darwin’s 1871 book, *The Decent of Man*; Hegarty & Buechel, 2006). In gendered languages, like Spanish, it can also be seen in the use of the masculine plural to refer to any group that includes at least one man. There is evidence that the masculine generic increases the salience of men and male exemplars relative to women and female exemplars (Stahlberg, Sczesny, & Braun, 2001; Moulton, Robinson & Elias, 1978; Switzer, 1990; Wilson & Ng, 1988). For instance, Bailey, Dovidio and LeFrance (2021) found that using the masculine generic title *Master* for a university leadership role (rather than the more neutral title *Head*) increased the degree to which students associated the role with men rather than women. It also increased their tendency to recall men (and forget women) who had previously occupied the role. In a similar way, Archer and Kam (2022) found that using the masculine generic title *Chairman* (rather than the more neutral title *Chair*) increased people’s assumption that the leader was a man and improved their memory for male occupants.

If an asymmetry in gender encoding is implicated in gender bias, then it should be possible to lessen this bias by increasing symmetry. There are two main ways in which language can counter asymmetrical gender encoding and therefore be gender fair: neutralisation and feminisation. The former involves replacing gender encoded words with the neutral gender version (e.g., replacing ‘Chairman’ with ‘Chair’). The latter involves retaining the male encoded word and introducing the female encoded version alongside it (e.g., replacing ‘Chairman’ with the paired nouns ‘Chairman or Chairwoman’). There is evidence that people who are exposed to neutralised or feminised gender fair language show less cognitive biases than those who are exposed to the masculine generic form. In particular,

people exposed to gender fair language are more likely to remember female exemplars or bring images of women to mind (e.g., Archer & Kam, 2022; Bailey et al., 2021; Gastil, 1990).

Although both gender fair strategies have been shown to lessen bias relative to the masculine generic, there may be differences in their effectiveness in some contexts. Indeed, the core concern of this proposal is people's tendency to associate the category of 'leader' with men rather than women, even though this category is gender neutral in many languages, including English (Koenig et al., 2011). The tendency to associate linguistically neutral social categories with men rather than women is not restricted to the leader category, or other roles where men are numerically dominant, but is also observed for a number of other common categories, like 'humans' and 'employees' (e.g., Bailey, LaFrance & Dovidio, 2019; Hegarty & Parr, 2023). In cases like these, feminisation may be more effective than neutralisation because it explicitly encourages people to cognitively activate the category of women alongside the (more accessible) category of men. Evidence for this possibility is provided by Lindqvist, Renström and Sendén (2019; see also Khosroshahi, 1989; Switzer, 1990) who described a hypothetical job applicant to English and Swedish speakers and asked them to choose the photograph from a mixed gender array that best captured their mental image of this applicant. When a feminisation strategy was used, such that the applicant was described with paired pronouns (i.e., "he or she"), participants were equally likely to select a photograph of a woman or a man. When a neutralisation strategy was used, such that the applicant was instead described with a neutral pronoun (i.e., "they"), participants were much more likely to select a photograph of a man.

If these findings hold, such that feminised gender fair language (versus neutralised language) boosts the cognitive accessibility of women when thinking about political leaders, then this may help to reduce gender bias in the associated stereotype. This expectation draws

on work that shows that stereotypes are both reasonably stable and also somewhat malleable (Allport, 1954; Blair, 2002; Katz & Braly, 1933, 1935). Specifically, there is evidence that the content of the stereotypes that are activated in people's working memory at any particular point in time can be shaped by a variety of contextual factors, including a person's social motives and aspects of their immediate social context (e.g., Blair, 2002; Haslam, Turner, Oakes, McGarty & Hayes, 1992; Garcia-Marques et al., 2006; Smith & Zárate, 1992). Of particular relevance for our current project, there is evidence that people who are exposed to, or asked to imagine, counter-stereotypical exemplars show less bias on both implicit and explicit stereotype measures than people who are not (Blair, Ma & Lenton, 2001; Dasgupta & Greenwald, 2001; Kurdi, Sanchez, Dasgupta & Banaji, 2023; for a meta-analysis see Forscher, Lai, Axt, Ebersole, Merham, Devine & Nosek, 2019).

Much of the work that has examined the impact of counter-stereotypic exemplars has relied on strong and overt manipulations (e.g., instructing participants to imagine a person who is inconsistent with a given stereotype). A feminised gender fair language intervention offers a more subtle — and more practically useful — way of increasing the cognitive accessibility of women in contexts where they may be counter-stereotypical. Our aim, in this proposal, is to test whether exposure to feminised gender fair language in the form of gender paired nouns or pronouns (versus exposure to gender neutral language) reduces people's tendency to stereotype political leaders as more similar to men than women. Based on the literature review above, we propose the following additional hypotheses:

H2a: Feminised (versus gender neutral) language will weaken the positive association between people's stereotypes of men and political leaders.

H2b: Feminised (versus gender neutral) language will strengthen the positive association between people's stereotypes of women and political leaders.

Study Overview

We will test our hypotheses by asking participants to describe the stereotypes of political leaders, men and women. Importantly, when asked to describe the political leader stereotype, participants will be randomly exposed to either feminised or neutralised language. Our primary aim is to assess the causal impact of this linguistic intervention on the association between people's stereotype of political leaders and their stereotypes of men and women. As a secondary aim, we will test the generalisability of our findings by collecting data from at least 15 different languages, each of which will be situated in a different national context. This will allow us to look beyond the (predominantly English-speaking) countries that have been the site of much of the cultural discourse around the use of gender fair language and, more recently, pronouns and the gender binary (Renström et al., 2023). In our sampling of languages, we will ensure that we collect data from at least five languages within each of the three language gender groups (neutral, natural and gendered). However, targeting these three groups will entail some flexibility in our operationalisation of a feminisation strategy because, unlike natural gender and gendered languages, gender neutral languages do not encode gender in pronouns. For this reason, in gender neutral languages, participants will be exposed to gender paired nouns rather than pronouns.

Our third aim is to increase the construct validity and the robustness of our findings by complementing the non-incentivised measure of stereotype content that is dominant in the literature with a novel, incentivised, measure. The dominant approach (as exemplified by Schein, 1973; see also Asbrock, 2010; Bye et al., 2014; Fiske et al., 2002) asks participants to rate their personal endorsement of different attributes for the social groups of interest. Because this approach does not incentivise truthful responding, it is vulnerable to social desirability concerns and may be considered to be a kind of 'cheap talk' (Crawford & Sobel, 1982). To address this vulnerability, we will introduce a new measure that is designed to elicit

stereotypes in an incentive compatible way. To create this new measure, we will adapt the incentive compatible approach that has been developed to elicit social norms for our own purposes (Krupka & Weber, 2013; see also Chang, Chen & Krupka, 2019; Krupka, Leider & Jiang, 2017). Specifically, the participant in each country whose estimate of their country-sample's modal stereotype is closest to the actual mode will receive a monetary prize. Any evidence of differences in the stereotypes that are produced by the non-incentivised and incentivised approaches is likely to advance the theory and practice of stereotype elicitation.

Method

Ethics information

This project has been selected for execution by the Psychological Science Accelerator (PSA). In line with PSA protocols, all labs that agree to participate in this study will be required to obtain and document ethical approval from their local ethics boards. We will obtain informed consent from all participants.

Participants

Participants will be a minimum of 3,600 individuals recruited from labs spanning a minimum of 15 languages (minimum language $N=240$, after exclusions) across the three gender groups that are part of the Psychological Science Accelerator (PSA) network ¹.

Participants will be adults over the age of 18 who, if they were not born in their country, have been resident in it since at least the age of 12. Participants will be excluded from the data

¹ Recent performance of the PSA network suggests that we should not have any difficulties in meeting this minimum sample size. However, if we fail to achieve our intended sample size with PSA labs, we will attempt to recruit additional labs to join the PSA network; if this fails, we will supplement our sample using paid crowd sourcing websites as appropriate for the target language (e.g., Prolific Academic or Qualtrics). For any language requiring a top-up sample, we will recruit at least 80 participants (after exclusions) with a similar demographic profile to those collected through PSA labs.

analysis if they (1) do not meet the above residence requirement, (2) fail either of the two attention checks and (3) have missing data on more than 20 percent of survey items.

Given the nature of the PSA network, we anticipate that most participants will be undergraduate university students. Baseline compensation for completing the study will be determined by local lab rules. Reflecting the makeup of the PSA network and the norms in psychology departments, participants are likely to participate in exchange for course credit although payment is possible in some labs. Participants will additionally have the opportunity of winning a monetary bonus on the basis of their performance on the incentivised measure of stereotype content. The best performing participant in each country will receive a prize of GBP100 (or Purchasing Power Parity equivalent).

We conducted a Montecarlo simulation to verify that our sample size, design, and analysis approach will be sufficient to enable reliable inferences regarding the parameters of interest. We generated 1000 simulated data sets using the proposed design and sample size and examined the extent to which the estimated parameters from models used to analyse these simulated data sets correspond to the data-generating parameter values. The results revealed a very high degree of correspondence between data-generating and estimated values (average $R > 0.99$), suggesting that sample size is sufficient to reliably test our hypotheses with our proposed design.

Design

The study uses a 2 (feminised gender fair language: present versus absent; between) X 3 (stereotype: political leader, men, women; within) mixed design. Participants will be randomly allocated to experimental condition. To control for order effects, participants will be randomly allocated to one of six possible stereotype presentation orders (this order will be retained across the two stereotype measures), where the first presentation group is either men, women or political leaders; a distractor group ‘the elderly’ will always follow the first gender

group. Participants will also be randomly allocated to one of two stereotype measure orders, such that the first stereotype measure is either non-incentivised or incentivised ².

Procedure

We will recruit participants to take part in a survey of people's perceptions of social groups. We will run the survey online and it will be accessed either in the participating lab or remotely. Surveys will be translated into the appropriate language for each country ³. This study will take no longer than 30 minutes to complete. The survey will measure stereotype content (non-incentivised and incentivised) and demographic variables. (For exploratory purposes, the survey will also administer a number of individual difference measures, including the list measure; Blair & Imai, 2012).

Stereotype Content: Non-Incentivised. The non-incentivised stereotype measure is based on Schein's (1973) original think manager – think male paradigm. Specifically, participants will be told that they will be presented with a series of descriptive terms that commonly characterise people in general and asked to use this list to indicate what they think various social groups are like. Participants will then be presented with each of the four target groups in turn: women, men, political leaders and the elderly (the latter group is a distractor that is included to minimise the risks of contrast effects or reactance due to the repeated measures design) ⁴. The order of the presentation of the groups 'women', 'men' and 'political leaders', will be randomised; the final group, 'the elderly', will always be presented after the first gender group to ensure that ratings of men and women are temporally separated. This randomisation will result in 6 presentation orders.

² Order effects will be examined in an exploratory fashion and reported as part of robustness analyses.

³ The translations will be overseen by a member of the PSA translation team. For each language, the translation will be accomplished by a team of five individuals (two individuals will conduct forward translations, another two individuals will conduct the backtranslation and a language expert will coordinate their work).

⁴ A pilot study of this approach revealed that about half of participants had some insight into the goals of the study, but there was no evidence that high versus low insight participants differed in terms of their stereotype responses.

When eliciting the stereotypes of women, men and the elderly, English speaking participants will be presented with the following instructions: *“On this page, we would like you to rate [target group]. In making your judgments, it may be helpful to imagine that you are about to meet a person for the first time and the only thing you know in advance is that the person is a [target group member]. Below, you will see 30 pairs of terms. Please select the point between the words in each pair that best describes what [target group] are like in general.”*

When eliciting the stereotype of political leaders, the instructions will vary as a function of experimental condition. English speaking participants in the control (feminised language absent) condition will be presented with the following instructions: *“On this page, we would like you to rate people who occupy political leadership positions. In making your judgments, it may be helpful to imagine that you are about to meet a person for the first time and the only thing you know in advance is that this person occupies a position of political leadership. Below, you will see 30 pairs of terms. Please select the point between the words in each pair that best describes what people who occupy political leadership positions are like in general.”* English speaking participants in the experimental (feminised language present) condition will instead be presented with text that encodes for male and female gender categories in the form of paired pronouns: *“On this page, we would like you to rate people who occupy political leadership positions. In making your judgments, it may be helpful to imagine that you are about to meet a person for the first time and the only thing you know about him or her in advance is that he or she occupies a position of political leadership. Below, you will see 30 pairs of terms. Please select the point between the words in each pair that best describes what people who occupy political leadership positions are like in*

general.” In gender neutral languages, which lack gender encoding in pronouns, the pronouns in the above text will be replaced with the paired nouns “this man or this woman”⁵.

Participants will then be presented with 30 semantic differentials. These attributes have been selected to tap into major facets of social evaluation, including 2 communal dimensions of friendliness and morality and 4 agentic dimensions of competence, assertiveness, dominance and self-sufficiency. For each dimension, we surveyed the attributes that had been used in previous stereotype research (Abele & Hauke, 2020; Abele et al., 2016; Ashton & Lee, 2001; Carli et al., 2016; Cuddy et al., 2015; Duehr & Bono, 2006; Epitropaki & Martin, 2005; Fiske, 2018; Hentschel et al., 2019; Imhoff et al., 2018; Koch et al., 2016; Louvet et al., 2019; Ma et al., 2022; Oliveira et al., 2020; Schneider & Bos, 2014) and chose additional attributes that best met the criteria of being frequently used, non-redundant, and capturing important aspects of the stereotypes of men, women and leaders. We constructed 7-point semantic differentials around these attributes (these are coded such that a score of 1 is given for total endorsement of the attribute on the left pole of the differential and 7 for total endorsement of the attribute on the right pole). We then sought feedback from native speakers of four culturally distant languages (German, Turkish, Arabic and Chinese) about ease of translation and consistency of meaning (following guidelines of the International Test Commission, 2017). This feedback led to some substitutions and amendments.

The attributes for friendliness are “caring—uncaring”, “warm—cold”, “empathetic—unempathetic”, “kind—unkind”, “good-natured—bad-natured”. The attributes for morality are “fair—unfair”, “just—unjust”, “moral—immoral”, “ethical—unethical”, “honourable—dishonourable”. The attributes for competence are “capable—incapable”,

⁵ When implementing the feminised instructions, we will follow the language’s convention for the ordering of the male- and female-encoded pronouns or nouns. We did not randomise this ordering for reasons of experimental power. However, it is possible that the unconventional ordering (e.g., she or he) would provide a greater boost to the salience of the female category, and that the conventional ordering is therefore a weaker test. The impact of using paired pronouns versus nouns will be analysed and reported in the robustness analysis.

“competent— incompetent”, “talented—untalented”, “skilled—unskilled”, “accomplished—unaccomplished”. The attributes for assertiveness are “strong—weak”, “confident—hesitant”, “decisive—indecisive”, “assertive—unassertive”, “never-giving-up—giving-up-easily”. The attributes for dominance are “dominant—submissive”, “conceited—humble”, “self-promoting—modest”, “pushy—easy-going”, “feeling-superior—feeling-inferior”. The attributes for self-sufficiency are “self-controlled—uncontrolled”, “self-reliant—other-reliant”, “mature—immature”, “independent—dependent”, “self-sufficient—unsufficient”. In the survey, the semantic differentials that relate to each dimension will be divided into sets and presented across survey pages. Within each set, approximately half of the differentials will have the positive pole on the right and the remainder will have the positive pole on the left. The dimensions will be presented in a random order.

Stereotype Content: Incentivised. This measure will be in most respects the same as the non-incentivised measure described above. Importantly, the order of presentation and the experimental condition will be retained across both measures so that participants see the experimental or control versions of the stimuli twice, in the same order. In order to incentivise truthful responding, we will make two changes to the non-incentivised measure. First, participants will be asked to estimate what *most other members of their community* (i.e., those who will also be completing the study in their country) think the four social groups are like (for a similar approach, see Kotzur et al., 2019; Kotzur et al., 2020). For example, when eliciting the stereotypes of women, English speaking participants will be presented with the following instructions: “*On this page, we would like you to rate what other members of your community think women are like. In making your judgments, it may be helpful to imagine that a member of your community is about to meet a person for the first time and the only thing they know in advance is that the person is a woman. Below, you will see 30 pairs of terms.*”

Please select the point between the words in each pair that best describes what other members of your community think women are like in general.” Second, we will adapt Krupka and Weber’s (2013) method in order to incentivise accuracy. In particular, participants will be told that we will calculate the number of times they select the response that is actually provided by most other members of their community and that the participant who attains the highest score will receive a bonus of GBP100 (or PPP equivalent).

Finally, participants will be asked to provide demographic information about their gender (man, woman, not listed here: please specify if you wish, and prefer not to say), age, country of residence (number of years resided in their country of residence, if not born there), education level, employment status and political conservatism. Participants will also be asked to respond to two attention check items: “In this survey, you were asked to rate four social groups. Was one of these groups ‘Doctors’?” and “The following question is an attention check. Please select the answer ‘very unhappy’ below.” Finally, participants will be asked to respond to an open question that asks what they thought the survey was investigating ⁶.

Sampling Plan

To achieve our intended sample size, we need to recruit a minimum of 240 participants for each of a minimum of fifteen languages (where each language is sourced from a single country). We also need to recruit a minimum of five languages from each of the three language gender groups (i.e., gender neutral, natural gender and gendered). Based on the geographical location of PSA labs and performance in recent many-lab studies, we estimate that the PSA is well placed to achieve samples in excess of 240 participants in twenty-four languages across the three gender groups. We will first solicit interest from potential labs and ask them to estimate their data collection capacity. We recognise that some

⁶ We will ask local labs to code these responses for the presence of insight into the study purpose (insight will be coded as present when participants make any reference to gender stereotypes) and include an analysis that examines whether the presence of insight conditions the finding in robustness analysis.

of the labs in less developed countries may face resource or other practical constraints that limit the size of the sample that they are able to collect. We therefore anticipate requesting that each lab collects data from a minimum of 80 participants. In some languages, it may be necessary to recruit multiple labs to achieve the minimum sample size. We will prioritise translation and data collection for languages for which we have a reasonable expectation of achieving our minimum sample. We will attempt to sample across both genders because there is some evidence that men and women differ in their endorsement of gendered leadership stereotypes (Koenig et al., 2011; Miller et al., 2014). Therefore, where gender-mixed samples are possible, we will ask labs to recruit a minimum of 32 participants of each gender (or 40 percent of the sample, if larger).

Analysis Plan

The analyses will be conducted using a fixed-effects regression approach with clustered standard errors (McNeish & Kelley, 2018; Antonakis, Bastardo & Rönkkö, 2021; Certo, Withers & Semadeni, 2017). This approach accounts for the multilevel structure of the data, where ratings are nested within participants, which themselves are nested within countries (note that each country will be represented by a single language, but some languages may be represented across more than one country). It also allows us to account for the unique effects of each semantic differential. For example, participants, on average, may be more or less likely to rate someone as ‘kind’ than to rate them as ‘never-giving-up’. The analysis will be implemented via the *fixest* (Berge, 2018) package in R (R Core Team, 2024), which efficiently estimates models such as ours with multiple fixed effects and multiway clustering. This approach overcomes the endogeneity problem (Antonakis et al., 2021; Bliese, Schepker, Essman, & Ployhart, 2020) because variation due to country, participant, or semantic differential is controlled for by treating these factors as fixed effects rather than treating this variation as a random effect.

The dependent variable in the model used to test Hypotheses 1 is the semantic differential rating associated with the non-politician. The predictors in this model are the semantic differential rating associated with the politician, an effect-coded gender variable indicating whether non-politician is a man (coded as 1) or woman (coded as -1), and the politician rating X gender interaction. The interaction term tests H1 by measuring whether the relationship between the ratings given to the man and the ratings given to the politician differs from the relationship between the ratings given to the woman and the ratings given to the politician.

In the model used to test H2a, the dependent variable is the semantic differential rating associated with the man only. The gender fair manipulation replaces the dummy-coded gender variable as the second predictor. Because the gender fair manipulation is between participants, the main effect of this manipulation cannot be identified when the fixed effect of participant is in the model. Thus, we only include the fixed effects of country and semantic differential in this model. The model that will be used to test H2b is identical to the model used to test Hypothesis 2, except that the semantic differential associated with the woman is the dependent variable.

References

- Abele, A. E., & Hauke, N. (2020). Comparing the facets of the big two in global evaluation of self versus other people. *European Journal of Social Psychology*, 50(5), 969-982.
- Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness—Communion with warmth and morality. *Frontiers in Psychology*, 7, 1810.
- Allport, G. (1954). *The nature of prejudice*. New York: Basic Books.
- Antonakis, J., Bastardo, N., & Rönkkö, M. (2021). On ignoring the random effects assumption in multilevel models: Review, critique, and recommendations. *Organizational Research Methods*, 24(2), 443-483.
<https://doi.org/10.1177/1094428119877457>
- Archer, A. M., & Kam, C. D. (2022). She is the chair (man): Gender, language, and leadership. *The Leadership Quarterly*, 33(6), 101610.
- Asbrock, F. (2010). Stereotypes of social groups in Germany in terms of warmth and competence. *Social Psychology*, 41(2), 76-81.
- Ashton, M. C., & Lee, K. (2001). A theoretical basis for the major dimensions of personality. *European Journal of Personality*, 15(5), 327-353.
- Bailey, A. H., Dovidio, J. F., & LaFrance, M. (2022). “Master” of none: Institutional language change linked to reduced gender bias. *Journal of Experimental Psychology: Applied*, 28(1), 237.
- Bailey, A. H., LaFrance, M., & Dovidio, J. F. (2019). Is man the measure of all things? A social cognitive account of androcentrism. *Personality and Social Psychology Review*, 23(4), 307-331.

- Beggs, J. M., & Doolittle, D. C. (1993). Perceptions now and then of occupational sex typing: A replication of Shinar's 1975 study. *Journal of Applied Social Psychology*, 23(17), 1435-1453.
- Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological Review*, 88(4), 354.
- Bergé, L. (2018). "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm." *CREA Discussion Papers*. Accessed from <https://CRAN.R-project.org/package=fixest>
- Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, 6(3), 242-261.
- Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, 81(5), 828.
- Bliese, P. D., Schepker, D. J., Essman, S. M., & Ployhart, R. E. (2020). Bridging methodological divides between macro- and microresearch: Endogeneity and methods for panel data. *Journal of Management*, 46(1), 70-99.
<https://doi.org/10.1177/0149206319868016>
- Bodine, A. (1975). Androcentrism in prescriptive grammar: singular 'they', sex-indefinite 'he', and 'he or she' 1. *Language in Society*, 4(2), 129-146.
- Boehm, U., Marsman, M., Matzke, D. et al. (2018). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behaviour Research*, 50, 1614–1631. <https://doi.org/10.3758/s13428-018-1054-3>
- Brenner, O. C., Tomkiewicz, J., & Schein, V. E. (1989). The relationship between sex role stereotypes and requisite management characteristics revisited. *Academy of Management Journal*, 32(3), 662-669.

- Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1). <https://doi.org/10.18637/jss.v080.i01>
- Bye, H. H., Herrebrøden, H., Hjetland, G. J., Røyset, G. Ø., & Westby, L. L. (2014). Stereotypes of Norwegian social groups. *Scandinavian Journal of Psychology*, 55(5), 469-476.
- Carli, L. L., Alawa, L., Lee, Y., Zhao, B., & Kim, E. (2016). Stereotypes about gender and science: Women ≠ scientists. *Psychology of Women Quarterly*, 40(2), 244-260.
- Certo, S. T., Withers, M. C., & Semadeni, M. (2017). A tale of two effects: Using longitudinal data to compare within- and between-firm effects. *Strategic Management Journal*, 38(7),
- Chang, D., Chen, R., & Krupka, E. (2019). Rhetoric matters: A social norms explanation for the anomaly of framing. *Games and Economic Behavior*, 116, 158-178.
- Conway, M., Pizzamiglio, M. T., & Mount, L. (1996). Status, communality, and agency: implications for stereotypes of gender and other groups. *Journal of Personality and Social Psychology*, 71(1), 25.
- Crawford, V. P. & Sobel, J. (1982). Strategic Information Transmission. *Econometrica*, 50(6), 1431–1451.
- Cuddy, A. J., Wolf, E. B., Glick, P., Crotty, S., Chong, J., & Norton, M. I. (2015). Men as cultural ideals: Cultural values moderate gender stereotype content. *Journal of Personality and Social Psychology*, 109(4), 622.
- Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, 81(5), 800.
- Duehr, E. E., & Bono, J. E. (2006). Men, women, and managers: are stereotypes finally changing? *Personnel Psychology*, 59(4), 815-846.

- Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3), 573.
- Wood, W., & Eagly, A. H. (2002). A cross-cultural analysis of the behavior of women and men: implications for the origins of sex differences. *Psychological Bulletin*, 128(5), 699.
- Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *American Psychologist*, 75(3), 301.
- Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, 69, 275-298.
- Epitropaki, O., & Martin, R. (2005). From ideal to real: a longitudinal study of the role of implicit leadership theories on leader-member exchanges and employee outcomes. *Journal of Applied Psychology*, 90(4), 659.
- Fiske, S. T. (2002). What we know now about bias and intergroup conflict, the problem of the century. *Current Directions in Psychological Science*, 11(4), 123-128.
- Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2), 67-73.
- Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, 117(3), 522.
- Froehlich, L., Olsson, M., Dorrough, A. R., & Martiny, S. (2020). Gender at work across nations: Men and women working in male-dominated and female-dominated occupations are differentially associated with agency and communion. *Journal of Social Issues*, 76(3), 484–511

- Garcia-Marques, L., Santos, A. S. C., & Mackie, D. M. (2006). Stereotypes: Static abstractions or dynamic knowledge structures? *Journal of Personality and Social Psychology*, 91(5), 814.
- Gastil, J. (1990). Generic pronouns and sexist language: The oxymoronic character of masculine generics. *Sex roles*, 23, 629-643.
- Gatton, D. S., Dubois, C. L., & Faley, R. H. (1999). The effects of organizational context on occupational gender-stereotyping. *Sex Roles*, 40(7/8), 567.
- Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, 101(1), 109.
- Haaf, J.M. & Rouder, J.N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, 26, 772–789.
<https://doi.org/10.3758/s13423-018-1522-x>
- Haslam, S. A., Turner, J. C., Oakes, P. J., McGarty, C., & Hayes, B. K. (1992). Context-dependent variation in social stereotyping 1: The effects of intergroup relations as mediated by social change and frame of reference. *European Journal of Social Psychology*, 22(1), 3-20.
- Hegarty, P., & Buechel, C. (2006). Androcentric reporting of gender differences in APA journals: 1965–2004. *Review of General Psychology*, 10(4), 377-389.
- Hegarty, P., & Parr, A. (2022). Embodied standpoints in gender difference graphs and tables: When, where, and why are men still prioritized? *Feminism & Psychology*, DOI: 09593535231181240.
- Heilman, M. E. (1983). Sex bias in work settings: The Lack of Fit model. *Research in Organizational Behavior*, 5, 269–298.

- Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior*, 32, 113-135.
- Bußmann, H., & Hellinger, M. (2003). Engendering female visibility in German. *Gender across languages: The linguistic representation of women and men*, 3, 141-174.
- Hentschel, T., Heilman, M. E., & Peus, C. V. (2019). The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. *Frontiers in Psychology*, 10, 11.
- Hilton, J. L., & Von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, 47(1), 237-271.
- Imhoff, R., Koch, A., & Flade, F. (2018). (Pre) occupations: A data-driven model of jobs and its consequences for categorization and evaluation. *Journal of Experimental Social Psychology*, 77, 76-88.
- International Test Commission, 2017. *ITC Guidelines for translating and adapting tests* (2nd Ed, Version 2.4). Accessed at https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf
- Katz, D., & Braly, K. W. (1935). Racial prejudice and racial stereotypes. *The Journal of Abnormal and Social Psychology*, 30(2), 175.
- Khosroshahi, F. (1989). Penguins don't care, but women do: A social identity analysis of a Whorfian problem. *Language in Society*, 18(4), 505-525.
- Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, 100(1), 128.
- Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110(5), 675.

- Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin*, 137(4), 616.
- Kotzur, P. F., Friehs, M. T., Asbrock, F., & van Zalk, M. H. (2019). Stereotype content of refugee subgroups in Germany. *European Journal of Social Psychology*, 49(7), 1344-1358.
- Kotzur, P. F., Veit, S., Namyslo, A., Holthausen, M. A., Wagner, U., & Yemane, R. (2020). ‘Society thinks they are cold and/or incompetent, but I do not’: Stereotype content ratings depend on instructions and the social group's location in the stereotype content space. *British Journal of Social Psychology*, 59(4), 1018-1042.
- Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3), 495-524.
- Krupka, E. L., Leider, S., & Jiang, M. (2017). A meeting of the minds: informal agreements and social norms. *Management Science*, 63(6), 1708-1729.
- Kurdi, B., Sanchez, A., Dasgupta, N., & Banaji, M. R. (2023). (when) do counterattitudinal exemplars shift implicit racial evaluations? Replications and extensions of Dasgupta and Greenwald (2001). *Journal of Personality and Social Psychology*.
- Lindqvist, A., Renström, E. A., & Gustafsson Sendén, M. (2019). Reducing a male bias in language? Establishing the efficiency of three different gender-fair language strategies. *Sex Roles*, 81(1), 109-117.
- Lombard, E. J., Azpeitia, J., & Cheryan, S. (2021). Built on Uneven Ground: How Masculine Defaults Disadvantage Women in Political Leadership. *Psychological Inquiry*, 32(2), 107-116.

- Louvet, E., Cambon, L., Milhabet, I., & Rohmer, O. (2019). The relationship between social status and the components of agency. *The Journal of Social Psychology, 159*(1), 30-45.
- Ma, A., Rosette, A. S., & Koval, C. Z. (2022). Reconciling female agentic advantage and disadvantage with the CADDIS measure of agency. *Journal of Applied Psychology, 107*(12), 2115.
- McNeish, D., & Kelley, K. (2018). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods, 24*(1), 20-35; Schunck, R. (2013).
- Moulton, J., Robinson, G., & Elias, C. (1978). Sex bias in language use: 'Neutral' pronouns that aren't. *American Psychologist, 33*(11), 1032-1036
- Okimoto, T. G., & Brescoll, V. L. (2010). The price of power: Power seeking and backlash against female politicians. *Personality and Social Psychology Bulletin, 36*(7), 923-936.
- Oliveira, M., Garcia-Marques, T., Garcia-Marques, L., & Dotsch, R. (2020). Good to Bad or Bad to Bad? What is the relationship between valence and the trait content of the Big Two? *European Journal of Social Psychology, 50*(2), 463-483.
- Powell, G. N., & Butterfield, D. A. (1979). The “good manager”: Masculine or androgynous? *Academy of Management Journal, 22*(2), 395-403.
- Powell, G. N., & Butterfield, D. A. (1989). The " Good Manager" Did Androgyny Fare Better in the 1980s? *Group & Organization Studies, 14*(2), 216-233.
- Powell, G. N., Butterfield, D. A., & Parent, J. D. (2002). Gender and managerial stereotypes: have the times changed? *Journal of Management, 28*(2), 177-193.

- Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly*, 26(4), 269-281.
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Renström, E. A., Lindqvist, A., Klysing, A., & Gustafsson Sendén, M. (2023). Personal pronouns and person perception—Do paired and nonbinary pronouns evoke a normative gender bias? *British Journal of Psychology*. DOI: 10.1111/bjop.12686
- Sastre, M., & TM, F. E., Igier, V., Salvatore, N., & Mullet, E. (2000). Perception of occupational gender typing: A replication on European samples of Shinar's (1975) and Beggs and Doolittle's (1993) studies. *Journal of Applied Social Psychology*, 30, 430-441.
- Schein, V. E. (1973). The relationship between sex role stereotypes and requisite management characteristics. *Journal of Applied Psychology*, 57(2), 95.
- Schein, V. E. (1975). Relationships between sex role stereotypes and requisite management characteristics among female managers. *Journal of Applied Psychology*, 60(3), 340.
- Schneider, M. C., & Bos, A. L. (2014). Measuring stereotypes of female politicians. *Political psychology*, 35(2), 245-266.
- Sczesny, S., Formanowicz, M., & Moser, F. (2016). Can gender-fair language reduce gender stereotyping and discrimination?. *Frontiers in Psychology*, 7, 25.
- Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, 99(1), 3.
- Stahlberg, D., Braun, F., Irmen, L., & Sczesny, S. (2011). Representation of the sexes in language. In *Social communication* (pp. 163-187). Psychology Press.

- Stahlberg, D., Sczesny, S., & Braun, F. (2001). Name your favorite musician: Effects of masculine generics and of their alternatives in German. *Journal of Language and Social Psychology*, 20(4), 464-469.
- Switzer, J. Y. (1990). The impact of generic word choices: An empirical investigation of age-and sex-related differences. *Sex Roles*, 22, 69-82.
- UN women. (2021). *Women in Politics*. Accessed from <https://www.unwomen.org/en/digital-library/publications/2021/03/women-in-politics-map-2021>
- Williams & Best, 1982
- Wilson, E., & Ng, S. H. (1988). Sex bias in visual images evoked by generics: A New Zealand study. *Sex roles*, 18, 159-168.