Gender stereotypes, language and performance

Della Giusta, M. ORCID: https://orcid.org/0000-0003-3959-4451, Jaworska, S. ORCID: https://orcid.org/0000-0001-7465-2245, Sevilla, A. and Razzu, G. ORCID: https://orcid.org/0000-0003-2640-8314 (2026) Gender stereotypes, language and performance. European Economic Review, 184. 105271. ISSN 1873-572X doi: 10.1016/j.euroecorev.2026.105271 Available at https://centaur.reading.ac.uk/127916/

European Economic Review 184 (2026) 105271

Contents lists available at ScienceDirect

European Economic Review

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

Gender stereotypes, language and performance*

Marina Della Giusta [a,*], Sylvia Jaworska [b], Almudena Sevilla [c], Giovanni Razzu [d]

[a] Università di Torino, Department of Economics and Statistics and Collegio Carlo Alberto
[b] Department of English Language and Applied Linguistics, University of Reading
[c] Department of Social Policy, the London School of Economics and Political Science
[d] Department of Economics, University of Reading

ARTICLE INFO

JEL Codes:
JEL D91 - Role and Effects of Psychological, Emotional, Social, and Cognitive Factors on Decision Making J16 - Economics of Gender Non-labor Discrimination

Keywords:
Gender,language
Stereotypes
Performance outcomes

ABSTRACT

This paper examines how gender stereotypes about competence and language in perfor- mance evaluations influences the assessment of man and women employees. Using a unique dataset of reviews annotated by expert linguists, we identify instances of communal and agentive language used by women and men managers and find that agentive language bene- fits women more than men employees, with stronger effects for women who outperform. We find that women evaluators need to use more agentive language to effectively assess women compared to male evaluators confirming the gender and competence stereotype. Addition- ally, we also find suggestive evidence that agentive language used by women evaluators for men employees negatively affects their ratings while men evaluators can use gender incongruent language. These findings highlight the spillover effects of gendered language and suggest that women leaders may face backlash when using traditionally male-typed language for men employees.

Introduction

This paper examines the impact of the evaluator's gender and the language they use in performance reviews on the performance assessment of men and women employees. This is an important question to address, as performance evaluations are closely tied to rewards, pay, promotions, and career progression. Moreover, evaluations are a key responsibility of managers and play a crucial role in their effectiveness, directly influencing their workplace and career prospects.

Women in leadership roles experience negative feedback and backlash (Chakraborty and Serra, 2024; Born et al., 2022; Grossman et al., 2019; Blau and Kahn, 2017), which the literature has related to gender stereotypes, that is beliefs that associate greater status worthiness and competence with men than with women (Ridgeway, 2001; Della Giusta and Bosworth, 2020). These stereotypes are also reflected in actual behaviors: information on performance is considered less credible if it comes from women than from men (Reuben et al., 2014) and internalised by women themselves who self-evaluate negatively in male-typed tasks giving rise to a gender gap in self evaluation (Exley and Kessler, 2022).

More specifically, gender stereotypes are beliefs about what women and men are typically expected to do (descriptive), or should do (prescriptive) (Fiske, 1993; Koenig, 2018; Cialdini and Trost, 1998; Prentice and Carranza, 2002). The concept, which originates from psychology, is now well understood in economics (Bertrand and Hallock, 2001). Women are expected to display communal behavior, that is being warm, sensitive and cooperative; men, instead, are expected to display agentive behavior, such as being

* Corresponding author.
E-mail addresses: marina.dellagiusta@unito.it (M. Della Giusta), s.jaworska@reading.ac.uk (S. Jaworska), a.sevilla@lse.ac.uk (A. Sevilla), g.razzu@reading.ac.uk (G. Razzu).

https://doi.org/10.1016/j.euroecorev.2026.105271
Received 14 December 2024; Received in revised form 20 January 2026; Accepted 23 January 2026
Available online 29 January 2026

assertive, competitive and independent. Accordingly, it is well understood in linguistics that women's language is more personal, compliant, polite, cooperative and process-orientated; instead, men's language is more matter of fact, assertive, aggressive, humorous, ego-enhancing and goal-orientated (Baxter, 2009; Schnurr et al., 2020). Therefore, problems may arise when women occupy positions of leadership that are normally associated with men: men are often considered natural leaders when they exhibit traits like aggression, assertiveness, abrasiveness or competitiveness (Chakraborty and Serra, 2024; Babcock et al., 2017; Rudman and Phelan, 2008; Babcock et al., 2003).

An emerging literature has analyzed written evaluations to address the extent to which there are systematic language differences in the reference letters, feedbacks and reports re- ceived by women and men. The findings so far suggest that women are significantly less likely to receive excellent recommendation letters than their male counterparts in geoscience (Dutt et al., 2016; Trix and Psenka, 2003); that women are more likely to receive critical subjec- tive feedback than men in performance reviews and teaching evaluations (Mengel et al., 2019; Boring, 2017; Cecchi-Dimeglio, 2017); that there are more mentions of achievements, agentive stance (assertive, dominant) and more superlative adjectives in letters written for men (Madera et al., 2009; Schmader et al., 2007) and, in peer descriptions in the Naval Academy as well as in EconJobRumors, more negative adjectives used to describe women (Wu, 2018; Smith et al., 2019). Assertive cheap talk is also used differently by women and men in order to increase leadership credibility (Manian and Sheth, 2021)

Findings specific to the use of agentive (assertive, dominant, competitive) and communal (cooperative, nurturing, collaborative) language, considered to be key dimensions of the gender and leadership stereotype, are less conclusive. For instance, Correll et al. (2020) distinguishes between viewing (the way employees are described) and valuing (the way they are assessed) and find that men and women were equally likely to be described with agentive language, to be seen as having managerial skills, and to be seen as effectively leading teams. Conversely, Baltrunaite et al. (2022) find that, in reference letters for candidates in economics and finance positions, men are described relatively more often as brilliant, and women as hardworking and diligent, and these descriptions relate with various subsequent career outcomes. They also find that reference letters for women candidates tend to stress assertiveness and agentive skills slightly more than those written for men candidates, although this difference becomes statistically insignificant once the candidate's observable characteristics are taken into account. Findings from sociological literature have long documented the way in which managers affect the process of employee's performance assessment (Castilla, 2011). In the linguistics literature, Baxter (2009) documents that, when women use agentive language in the workplace, they also appear 'unfeminine' and experience backlash.

Existing literature has separate analysis of gender language on outcomes, such as perfor- mance, and of gender of evaluators on outcomes. This cannot shed light on the mechanisms and the way stereotypes operate. Indeed, for that we need to assess the interaction of gender stereotypes coming from the use of gender language from evaluators and the gender language used to describe employees. Therefore, existing evidence does not address how stereotypes op- erate but only partial channels. In fact, linguistic literature indicates that gender stereotypes operate at two levels: i) what is appropriate language to describe men and women; ii) what is the appropriate language for women and men to use when they express themselves. In this paper, we contribute a linguistically informed analysis of gender stereotypes in language that affects the gender gap in leadership, which is possible because of the unique dataset, which includes information on use of language by both women and men evaluators across the range of evaluations they perform as well as the complete set of evaluations received by women and men employees. Moreover, we do so by proposing a theoretical framework in which two separate channels are be considered, given that gender stereotypes in language refer to both the gen- der of the person evaluating and that of the person being evaluated (Baxter, 2009). The two channels operate through: 1) the extent to which the use of agentive (**A**) and communal (**C**) language is describing the performance of staff in congruent terms (Communal language being congruent for women, Agentive language for men); 2) how this interacts with the gender of the person (which we refer to as manager) who is describing the performance of employees. The interaction between the gender of both the manager and employees and the use of congruent agentive and communal language, creates eight possible scenarios, as represented in Table 1.

While the use of agentive language by a man manager for a man employee, and that of communal language by a woman manager for a woman employee, are gender congruent, the other six combinations potentially have ambiguous effects that need to be investigated in order to establish the full effect of language in performance evaluation.

The literature suggests that male managers are considered more capable of evaluating per- formance than women (Eagly and Karau, 2002; Cuadrado et al., 2015; del Carmen Triana et al., 2024), so we can hypothesize that when incongruent associations are present in their evaluations of employees (communal language being deployed to describe women by a man or to describe men by a man) these will not be as serious than when incongruent associations are present in female managers' evaluations of employees. Therefore, while the linguistic approach to gender stereotypes characterises as appropriate for male use and male descriptions agentive-typed language, and conversely appropriate for female use and description communal-typed language, the context of performance evaluations

Table 1
Framework: gender stereotypes in language.

Employee Man	Woman		
Man			
Woman		A congr.	
C weak incongr.		A weak incongr. C weak congr.	
		A weak congr. C weak incongr.	A weak incongr.
C congr.			

also has to take into account further overarching stereo- types concerning the ability of women and men to carry out evaluations per se. The gender competence stereotype has been found in the literature in a variety of setting including specif- ically that of evaluating performance. For this reason, we generate the hypothesis that gender incongruent use of language by men will have less serious implications than gender incongruent language used by women. Our analysis fully assesses how gender stereotypes in language affect the performance evaluations of men and women, via a unique dataset comprising the whole universe of evaluations written by men and women managers for men and women employees of an international services organization, in which both agentive and communal behavior are considered positive contributors to performance.

Our analysis reaches three key conclusions: i) we find evidence that agentive descriptions are beneficial to women, there being increasing returns to the use of agentive language that are larger for women (who in our sample are also more likely to be out-performing) than for men; ii) we also find suggestive evidence that women evaluators need to deploy a larger amount of agentive language when describing women than is the case for men evaluators, who are more effective, overall, in obtaining high scores for employees; iii) suggestive evidence is also present that the use of agentive language by women evaluators in the evaluations of men employees has negative effects on the performance scores received by these men.

We interpret our findings in the light of both the literature on gender stereotypes in lan- guage and the experimental literature on backlash, to suggest that women leaders' competence, with respect to describing men's performance with male language, is not believed (Baxter, 2009; Reuben et al., 2014): indeed, it appears to potentially have negative spillovers on the men that work with them. This is a novel and interesting result, as it highlights a spillover of backlash against men who work under women in leadership positions.

Performance Evaluation Process

Our data are from an international company in the service industry. This organization has developed, over the past years, a manifested commitment to equality, diversity and inclusion. For instance, it has undertaken work to assess the extent to which per-formance guidance doc- uments effectively incorporate the consideration of diversity and inclusion or, inadvertently, risk developing or preserving gender biases. The guidance documents contain examples of be- haviors expected for each of the areas of activity, grades and performance levels. The areas of activity cover technical excellence, exceptional client service, people engagement, diversity and inclusiveness, outcome based metrics, innovation and transformational change and driving internal initiatives. There are 5 hierar-chical grades from highest to lowest: Director, Assistant Directors, Senior Executive, Executive; 3rd year analyst. Our data refer to the performance of Assistant Directors, as evaluated by Directors and Partners.

Our analysis focuses on the performance evaluation outcome obtained by Assistant Direc- tors (whom we refer to as employees) at the end of the organisation's performance evaluation cycle, which is reached through a set of stages (Fig. 1). All those at Assistant Director level are included in the performance evaluation cycle; they are assessed three times a year, at the end of quarterly assessment periods. During these periods, the process requires them to ask for written evaluations from colleagues at Director or Partner level (whom we refer to as managers) whom they have worked with, as part of the same team or just for a single individual project of varying degree of collaboration and time involvement. There is no specific prior matching between the employees who request and the managers who provide the written evaluations. Employees and managers belong to the same team, and we have data from all five teams of the organisation, which provide advice on corporate finance, but also on economic evaluation, taxation, restructuring and strategy and diligence. Teams do not compete as they work on separate projects, and within teams roles are assigned and collaboration is often essential to carry out the projects for external customers.

Managers are asked to provide written feedback on one or all of the following three dimen- sions: i) quality, risk management and technical excellence; ii) impact; iii) individual collabo- rative strengths which matter to the core business of a services organization. This means that one employee might receive written evaluations from more than one manager and the same manager might provide evaluations to more than one employee, if they have collaborated on one or more projects during the year. In our dataset, one employee receives evaluations by an average of three managers who might provide feedback on one of the three dimensions, so that they can have up to nine pieces from up to three different managers. As managers are asked by employees to provide written eval-uations on their performance, strategic behavior could take place, both from employees who could seek reviewers mostly from "benevolent" managers, but also from managers who might have an incentive to write in relatively more positive terms on the per-formance of those employees whom they would like to help achieving a higher score.

However, the overall performance evaluation process, including the final meeting [and the use of a forced distribution of scores], will mediate for this.[1] Fig. 2 shows two examples of anonymized written evaluations.

The final stage of the process consists of a moderating meeting for each Team, attended by all managers but not by the employees being evaluated, during which they consider and discuss the evidence gathered for each candidate and agree a final performance assessment score: Not Performing, Performing, Outperforming. Performing is generally associated with an employee performing well and progressing in line with the peer group's performance and expectations for the grade. An outperformer employee, instead, makes a difference and stands out from others by having had, for example, strategic impact on the business or project in which the employee has worked during the year. The organization uses variations in performance evaluations to establish employees' entitlement to performance awards and bonuses. The promotion process is not directly linked to performance evaluation and is rather based on many

[1] We do not have information on whether managers, who have been asked to write an evaluation, can refuse or have refused to do so.

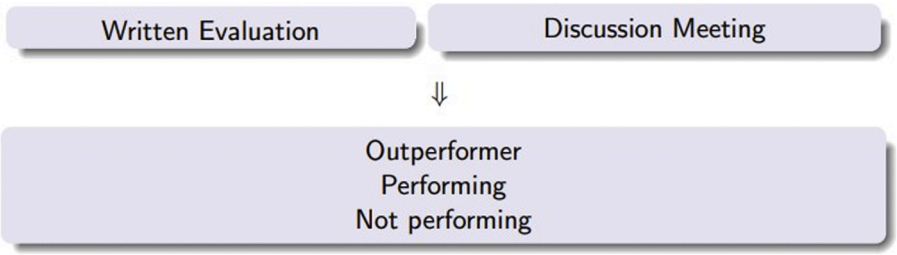


Fig. 1. Evaluation Process.

Following on from <PERSON_NAME> strong performance and progression in Cycle 2 I have made the following observations of his performance while working with him on Project <PROJECT_NAME> as well as internal initiatives such as knowledge management: From the get go on Project <PROJECT_NAME> <PERSON_NAME> displayed a strong intellectual curiosity which helped in understanding the <TEAM_NAME> landscape and red flags for our client. Furthermore he also encouraged junior members of the team to engage in similar discussions with various stakeholders. For example given the target state <ORG> solution was a core consideration in determining the deal thesis he encouraged me to leverage my past experience on managing <ORG> products to run management meetings in order to further learn about the <ORG> structure on Project <PROJECT_NAME> which is reflective of his transformative leadership style <PERSON_NAME> effectively navigates complex relationships and contexts with a strong showcase of emotional intelligence. There were numerous stakeholders on Project <PROJECT_NAME> including Private Equity Managing Director level executives C suite leadership from prospective co bidders and internal EY stakeholders such as the <ORG> team and <ORG> <PROJECT_NAME> leadership whom <PERSON_NAME> developed sustainable and trusted relationships with which were leveraged over the course of the engagement to achieve positive outcomes for the client <PERSON_NAME> ability to manage expectations of <TEAM_NAME> leadership on engagements and strong engagement delivery has made him one of the most trusted members of the pre deal team. He manages resources on engagements effectively and creates open communication channels with team members which ensures that a continuous circle of feedback exists within the team. This in turn leads to the delivery of strong diligence reports that have been reviewed iteratively on multiple occasions <PERSON_NAME> continued to showcase his inclusive leadership style on Project <PROJECT_NAME>. For example there were multiple occasions where he encouraged me to speak up take more responsibility in terms of ownership of deliverables such as the ERP landscape considerations document and capex estimates model and also walk through the work done with <TEAM_NAME> leadership. When the client provided positive feedback on these deliverables he ably recognised this as a team achievement. He also constantly provided feedback and coaching to junior members of the team so they had the opportunity to learn and grow professionally <PERSON_NAME> also played an active role in promoting initiatives that addressed priorities of the pre deal team internally. For example he prioritised the knowledge management initiative since new joiners were expected to join the team in <EVENT>. He also ensured that the compilation of best practice documents and deliverables on <TEAM_NAME> pre deal was managed effectively and went live for team members to utilise on future engagements Leveraging the depth of his <TEAM_NAME> pre deal knowledge <PERSON_NAME> also supports the delivery of proposals in the pursuit strategy for engagements. This in turn provides scope for additional revenue recognition opportunities for the <TEAM_NAME> pre deal team Whilst pricing for <TEAM_NAME> resources on engagements <PERSON_NAME> takes the lead in forecasting the budget and recommends appropriate responses to ensure that engagements are not undersold. This also ensures that the true value of the delivery performed by team members is seen in terms of revenue recognised during engagements
<PERSON_NAME> delivers consistent high quality work with strong commercial outcomes whilst maintaining a focus on self wellbeing as well as the well being of others. She manages both client and team professionally and exceptionally effectively. This year she has stepped up to undertake challenging and non standard projects with higher degrees of complexity / risk e.g. the Schemes of arrangement litigation with <TEAM_NAME> for Project <PROJECT_NAME> all with various stakeholders to manage (internal and external. She has consulted appropriately and reached out to risk on a number of occasions to manage EY and team exposure. She has managed multi discipline teams and brought to bear several specialisms collectively to meet the client requirements. e.g. on <PERSON> she was a conduit for the <TEAM_NAME> team with translating points/arguments from the technical <ORG> areas <PERSON_NAME> has built internal networks and routes to market through <TEAM_NAME> and the Restructuring team She has provided invaluable support in the team recruitment drive which this year has been exceptionally involved and has required a a significant investment of time. <PERSON_NAME> continues to make strides in both her personal development and her market presence and should be seriously considering her timeline to progression to director.

Fig. 2. Example of anonymized written feedback.

rounds of performance assessment so it is possible that doing very well in one assessment cycle does not translate in a promotion.

Before the meeting, two members of staff collect, assemble and check all the information for each of the employees in their team, so to ensure that the meeting is run effectively and with relevant and appropriate data and information. During the meeting, chairs ask managers to present the employees' case. Most often, the case is based on scores that have already been subjected to some moderation during the pre-meeting checks. A discussion follows the pre- sentation of the case, including inputs from other managers who have also worked with the employees. Rarely, further checks may be needed after the meeting, which the two staff mem-bers will carry out before finalizing the score. The final score results from allocating points to all criteria considered and applying equal weighting to all criteria.[2]

[2] We do not have detailed information on the exact points systems within each team and criteria, as we have been given only the final performance score.

Data

Our dataset consists of 1054 evaluations on the performance of assistant directors written by directors and partners, which comprises the universe of all employees being evaluated at this grade level in this performance evaluation round. The dataset also contains information about the gender of the person being evaluated, the gender of the managers who has written the reviews, the performance assessment score received by each employee at the end of the performance cycle (which we described in the previous section) and also the team the employee has been working with during the performance assessment cycle. Our sample is described in Table 2. The 1054 written evaluations, written for 164 employees by 335 managers, comprise the universe of all employees being evaluated at this grade in the organisation: each evaluation is one observation in our sample. Men represent 72.5% of managers and almost 68% of employees, a clear indication of a substantial pipeline issue already present at this middle-career level in the organisation. Therefore, of the full sample we use in analysis, men who write an evaluation for men represent 51% of all evaluations, men who write for women 24%; women who write for men 17% and women who write for women 8%.

Table 2 also shows that women are more likely to be assessed as outperforming than men. Of the 53 women employees, 46% are assessed as Performing and 54% as Outperforming. Of the 111 men employees, instead, 49.5% are assessed as Performing and 50.5% as Outperforming. Although not performing is also one of the three possible performance scores, no one in our sample receive it. Moreover, although we do not have any evidence of a formally written policy on the forced distribution of performance scores, we believe that the nature of the performance evaluation process cannot result in all employees receiving an outperforming score. The number of employees and of criteria considered in the evaluation, as well as the need to assess relative performance, alongside the nature of the process that includes three evaluation points in the cycle, meetings to discuss each employee's performance relative to others, all ensure some variation at least among performing employees.

From the written evaluations, such as the one shown in Fig. 1, we construct the indicator of Agentive and Communal language, through a process of double written annotation of the entire dataset by linguistic experts, who identify agentive language in context, something that automatic semantic classifiers do not allow, and with respect to the corpus of the language of the organization, as opposed to the use of set dictionaries, such as the corpus of the English language in general. Therefore, the meanings assigned to the verbs and adjectives were based on the ways in which the words were used in context reflecting organisational activities and practices. This approach also ensured the effective consideration of text polarity, the possible negative, positive or neutral meaning and underlying sentiment in the use of the term in the evaluation. We have also considered the use of automated analysis, such as one that employs the LWIC software, to assess the positive or negative sentiment of the evaluations but we found it to be inaccurate. Indeed, when we tested it on our data it has produced incorrect results. For example, 117 (out of 1054) evaluations were recognised as having some negative tone and when we scrutinised the top 'most negative' evaluations, these were not negative at all.[3]

The resulting indicator is based on the frequency of agentive or communal words (both adjectives and verbs) in each of the 1054 performance evaluations, relative to the length of the written evaluation, as per equation 1:

$$X_i = \sum_i W_{Li} \tag{1}$$

Where $\sum_i W_{Li}$ is total number of words in each evaluation, L $\in$ (*Agentive, Communal*).

In our sample, on average, one employee receives evaluations from three managers. 74% of the evaluations received by men employees are written by men managers - the remaining 26% by women managers. Similarly for the evaluations received by women, 75% of which being written by men managers and the rest by women managers, thus suggesting that the data we have does not have a disproportionately balanced allocation of evaluations providers by gender and, consequently, no selection by gender. However, on average, women managers write longer evaluations than men managers, the average number of words per evaluation being 109 and 97 respectively. Women managers write longer evaluations for women than for men (19 more words on average), while men managers write longer evaluations for men than for women (9 more words on average). Although there does not appear to be sorting by gender, there could be sorting by perceived effectiveness of the manager in getting employees ranked highly. We cannot check for this in our sample, as we do not have data on whether employees perceive some managers to be more effective than others, but we control for managers' heterogeneity in the specification.

The extent of agentive terms used in evaluations is shown in the bottom part of Table 2, first in absolute and then in relative terms, the latter controlling for the length of the written evaluation. In absolute terms, although women receive evaluations containing, on average, the same number of agentive terms to that received by men (12), it is women managers that employ slightly more agentive terms in evaluations for women than they do for men and men managers that employ more agentive terms in evaluations for men than they do for women: on average, for every agentive terms used in evaluations of women, men managers use 1.17 agentive terms in men's evaluations; for every agentive term used in evaluations of men, women managers use 1.27 agentive terms for women. However, when taking into account the length of the evaluations, men managers employ the same relative number of agentive terms whether they write an evaluation for a woman or a man, while women managers deploy slightly less agentive terms for women employees than for men employees.

Some observations emerge from our sample, in relation to the evidence from the literature. One is related to the context we study, which is of an organisation with expressed commitment to equality, diversity and inclusion. Indeed, our sample shows that women are

[3] Annex A details the process employed to construct the language indicator. Annex B discusses the results from using the LWIC to assess the sentiment of the written evaluations.

Table 2
Descriptive statistics.

	Male	Female
Managers (335)	243(72.5%)	92(27.5%)
Employees (164)	111(67.7%)	53(32.3%)
	Sample = 1054 written evaluations	
%Performing	49.51	45.97
%Outperforming	50.49	54.03
Average evaluation length		
received by/written by	Male manager	Female manager
Male employee	100	103
Female employee	91	124
Number of agentive terms - absolute*		
received by/ written by	Male manager	Female manager
Male employee	12.2	11.2
Female employee	10.4	13.4
Number of agentive terms - relative**		
received by/ written by	Male manager	Female manager
Male employee	0.12	0.11
Female employee	0.12	0.10

* Absolute: total number of agentive terms present in written evaluations. ** Relative: total number of agentive terms divided by the length (total number of words) of written evaluations.

more likely to be assessed as outperforming than man. In this context, we would perhaps expect gender differences to be less likely not just in performance assessment that offers limited options, such as not performing, performing and outperforming, but also in the language used to evaluate the performance of employees: ratings are easily compared, making gender differences in ratings unlikely, especially in a company committed to being transparent about their progress on gender diversity. Moreover, it is worth excluding from the outset the often advanced point that men's evaluations contain more agentive terms than those of women, because men are better at leadership and dispel leadership skills more pronouncedly than women do. However, this is not the case in our sample, which has no substantive difference in the average use of agentive terms by gender.

Model

We have described how employees can collect variable numbers of evaluations from variable numbers of managers and how there is a range of factors that combine into the final score. In order to address our research question and analyse the role of managers, the key is to assess the relevant unit of observations at which it is possible to consider simultaneously the gender of the manager that writes the evaluation, the gender of employee being evaluated and language used in each evaluation. This unit of observation is the single report written by a man/woman manager for a man/woman employee. This is also the unit of analysis necessary to investigate gender stereotypes in use of language, which refer to both the gender of person writing (what language is appropriate for a woman/man to use) and the gender of the person being described (what language is appropriate for describing a woman or a man), as shown in the language stereotype framework represented in Table 1.[4]

Therefore, we estimate a logistic regression,[5] as per equation 2:

$$PSi = \alpha Gi + \beta Xi + \gamma XiGi + \delta Ti + \lambda GMi + \sigma Li + \epsilon i \quad (2)$$

The dependent variable is the performance evaluation score PS_i, associated with each writ- ten evaluation, our unit of observation i. The independent variables consist of the gender of the employee, G_i; the language indicator X_i, in addition to the interaction between the language indicator and the gender of the employee. We also control for the team the employee belongs to, T_i, a reflection of the fact that teams maybe characterized by distinct underlying work- place climate and practices and specific team culture may negatively affect the performance assessment of employees in ways that are unrelated to the interpretation of language in the evaluations. For example, team members might perceive men with women supervisors as weak, interpreting their choice of supervisor as a negative signal. In the general baseline regression in equation 2, we also control for the gender of the manager providing the written evaluation, GM_i and for the length of each of the written evaluation, L_i. To test our hypotheses regarding the use of gender congruent language, we then run the regression by splitting the sample by the gender of the manager.

In addition, we integrate the specification in equation 2 in order to: i) consider the possible difference in managers' perceived ability to be effective in writing evaluations, therefore the possible within-manager variation; ii) take account of the possible sentiment of the evaluation computed with an automated tool, specifically the Linguistic Inquiry and Word Count (LIWC); iii) consider that agentive and communal terms may be used strategically by the manager in some form of complementary or compensatory way; iv)

[4] If the unit of analysis became the employee, with the sum of all evaluations received we would lose the information on the gender of the manager who wrote them.

[5] We prefer the logit regression as the linear probability model (LPM) does not force the predicted values to be between 0 and 1, which is of course incorrect. In any case, estimates based on the LPM do produce the same findings.

consider the role that different shares of female managers may have on the performance scores of employees to see whether their evaluations have the same importance as male managers.[6]

Results

Table 3 reports the odd ratios from the logistic regression in equation 2. The key coefficient of interest is that of the interaction between the gender of the employee receiving the evaluation and the language indicator, i.e. $\gamma i X_i G_i$. Given that in a logistic regression the value of the interaction effect changes depending upon the value of the continuous predictor variable (in our case, the language indicator), we focus on the predicted probabilities at different levels of the language indicator. These are particularly helpful to assess whether the use of, for instance, the same agentive language in men and women's evaluation is associated with significant dif- ferences in performance scores. Indeed, each predicted probability is computed at a specific level of agentive terms, therefore directly showing whether this probability differs by gender when the evaluation - whether provided to a woman or a man - contains the same amount of agentive or communal terms. Moreover, by splitting the sample according to the gender of the manager, we are able to test the extent to which the chances of receiving financial rewards and of career progression in the organization are related to the way men and women managers employ language and how this is perceived depending on the combination of the gender of those who evaluate and those who are being evaluated. These results are shown in Fig.s 4 to 6.

Fig. 3 and Fig. 4 report the marginal effects computed from the logistic regression in equation 2, namely the probabilities of being assessed as outperforming at each level of agen- tive terms present in the evaluation and controlling for the Team the employee belongs to, the gender of the manager and the length of the written evaluation. We identify four main findings.

First, women get higher returns then men from use of agentive terms in evaluation, regard- less of whether the evaluation is written by a woman or by a man. In fact, as shown in the left panel of Fig. 3, for each level of agentive terms, and across the whole range, women are more likely than men to be assessed as outperforming: the gender gap in the probability of being assessed as outperforming (in favor of women) is significantly positive and becomes larger as the number of agentive terms used in evaluations increases. For example, when evaluation for both men and women contains 10 agentive terms, women are around 1.5 percent more likely to be assessed as outperforming than men. When the evaluation contains 20 agentive terms, women are 7.5 percent more likely than men to be assessed as outperforming. The gender gap in this probability increases to almost 36 percent when the numbers of agentive terms contained in evaluations approaches 90, the maximum in our sample. The right-hand panel of Fig. 3 reveals that the increasing gender gap results from increasing returns (measured in terms of likelihood to be assessed as outperforming) associated with the use of agentive terms for women but not for men. In fact, the probability of being assessed as outperforming for women grows much more markedly than that of men as the number of agentive terms used in evaluations increases. Across the whole range of agentive terms in our data, the predicted probability for men is essentially constant at around 0.5, while that of women almost doubles from 0.5 to 0.9.[7] Second, Fig. 4 shows that when men managers write evaluations this is associated with all employees having positive and increasing probabilities to be assessed as outperforming, and this is particularly the case for women than it is for men employees.

Third, women evaluators need to use relatively more agentive terms in order to achieve the same results as men evaluators. More specifically, looking across the two panels of Fig. 4, for women employee to have, for instance, a 0.6 probability of receiving an outperforming score, women managers have to use around 30 agentive terms while around 10 would suffice in eval- uations written by man managers.

Fourth, when it is women writing the feedback, the use of agentive terms in evaluation of men employees appears to backfire. The right panel of Fig. 4 shows that the probability of being assessed as outperforming for men actually decreases as women managers write evalua- tions that contain increasing numbers of agentive terms.

We then consider the combination of gender of the manager and communal language on employees' performance assessment. Fig. .5 shows that use of communal language also helps to achieve high scores, but again this appears to be the case particularly when the manager is a man, while the returns from use of communal language are much less pronounced when the manager is a woman.

As the figures clearly suggest, there are indeed different returns to the use of agentive and communal language by men and women managers for the employees they write evaluations for, and these, as predicted, vary also by gender of the employees. Table 4 summarizes these differences and relates them to the hypotheses on gender stereotypes in language discussed in the introduction.

Our first result illustrates that, indeed, when an organization values communal behaviors alongside agentive ones, these are reflected in performance evaluations. It is however clear that the gender of managers is the driving force of whether agentive and communal language produce returns for the employees they evaluate. Man managers can effectively deploy both types of language, generating returns that vary between ten and forty percent to the employees they evaluate. Conversely, women managers can only be very effective in evaluating women employees but not men employees. The data further suggests that women managers appear to understand the role that the use of agentive terms may have in addressing gender gaps in performance evaluations and perhaps aim to compensate for the disadvantage faced by women. Results show that women evaluators need to use relatively more agentive terms to be effective, as this is incongruent at two levels: that of women evaluators not being credible (Reuben et al., 2014) and the fact they are assessing leadership skills of men using male language (Baxter, 2009). The most striking feature of our results is the markedly negative returns to men employees being evaluated in agentive terms by women managers. We interpret this to suggest for the presence of possible spillovers of backlash against women in leadership positions, which is, in this case, experienced by men evaluated by them.

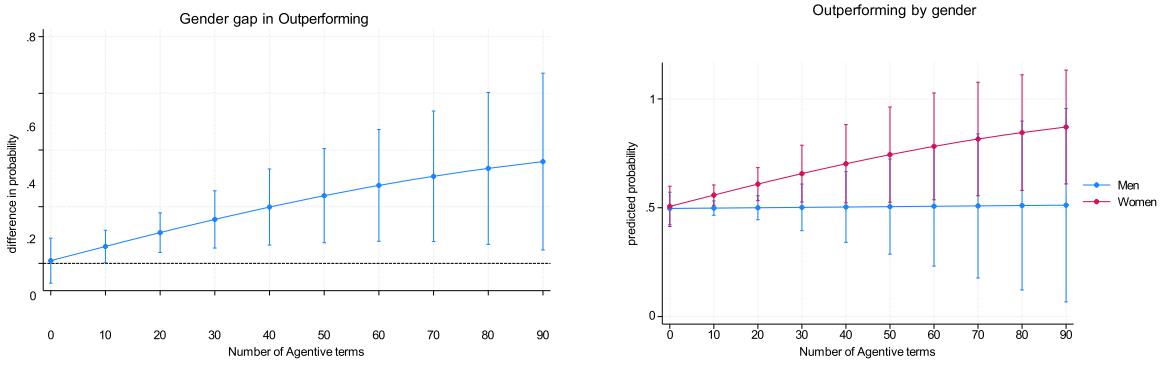
[6] We thank anonymous referees for all these suggestions.

[7] These results do not change when we employ a linear probability model, which forces the probabilities to be linear.

Table 3
Logistic Regression - Model in equation 2: Odds Ratios.

	(1) Agentive Language	(2) Communal Language
Language A	1.001	
	(0.0143)	
Language C		1.053**
		(0.0230)
Gender of employee	1.041	1.227
Gender × Language A Gender × Language C	(0.208) 1.021* (0.0127)	(0.236)
		1.013
		(0.0237)
Manager Gender	0.872	0.846
	(0.129)	(0.126)
Team=2	0.564* (0.167)	0.514** (0.153)
Team=4	1.283	1.335
	(0.256)	(0.268)
Team=5	2.011*** (0.345)	1.944*** (0.333)
Team=6	0.947	0.917
	(0.190)	(0.184)
Length of text	1.000	0.998*
	(0.00177)	(0.00128)
Observations	1048	1048

Exponentiated coefficients; Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



(a) Gender gap in outperforming **(b)** Outperforming by gender

Fig. 3. Probability of Outperforming by gender of employee and number of agentive terms in performance evaluation.

This could of course result in very serious dynamic consequences within the workplace, which are beyond the scope of this paper and not possible to address with the available data.

Our results on the sole effects of stereotypes in language and competence hold but become suggestive rather than significant when controlling for differences in managers fixed effects, sentiment indicators and proportion of women managers in an employees' overall feedback.[8] We employ two possible indicators of sentiment from the LWIC tool, one on the extent of positive tone and the other on overall emotion present in each evaluation. We fully discuss the limitations of these measures in Annex B, which also shows, that despite these limitations, the inclusion of these sentiment measures does not changes the gender gaps in the probability of being assessed as outperforming. We also consider the proportion of women managers that provide an evaluation for each employees. Note that, in this, the unit of assessment becomes the employee and the sum of respective evaluations they received, which means that we lose information on the gender of the managers writing each evaluation. This necessarily obscures the effect of gender stereotypes that the language literature argues emerge from the perceived appropriateness of the language and competence of a manager (men or woman) in assessing an employee (man or woman), as well as the type of language being used.

Concluding Remarks

Organizations that adopt active diversity policies benefit women, but there remains an area of concern regarding those who are managed by women. In fact, as more women get into leadership positions, they are more likely to assess men and be ineffective at doing

[8] See Table 5 in Annex C.

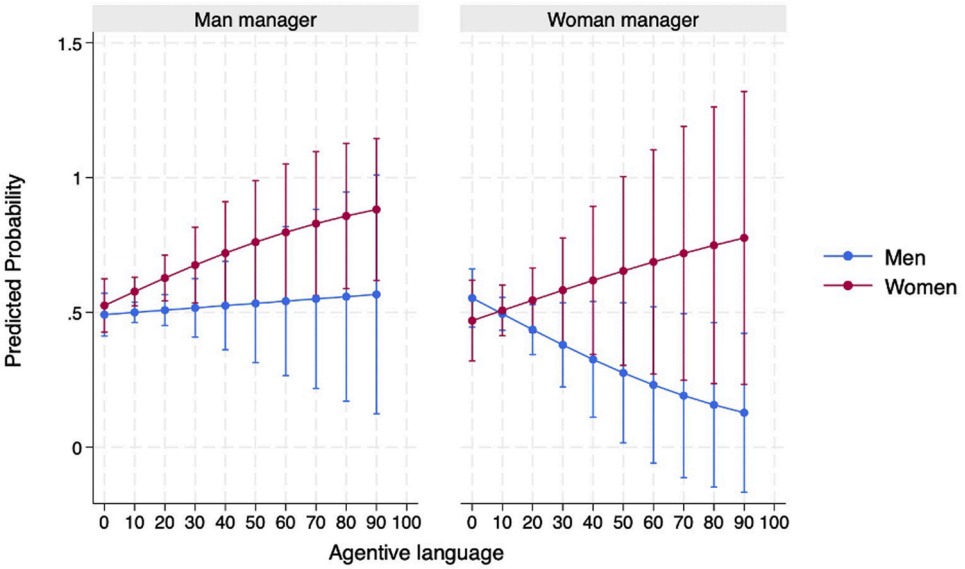


Fig. 4. Probability of Outperforming by gender of manager and employee and number of agentive terms used in performance evaluation.

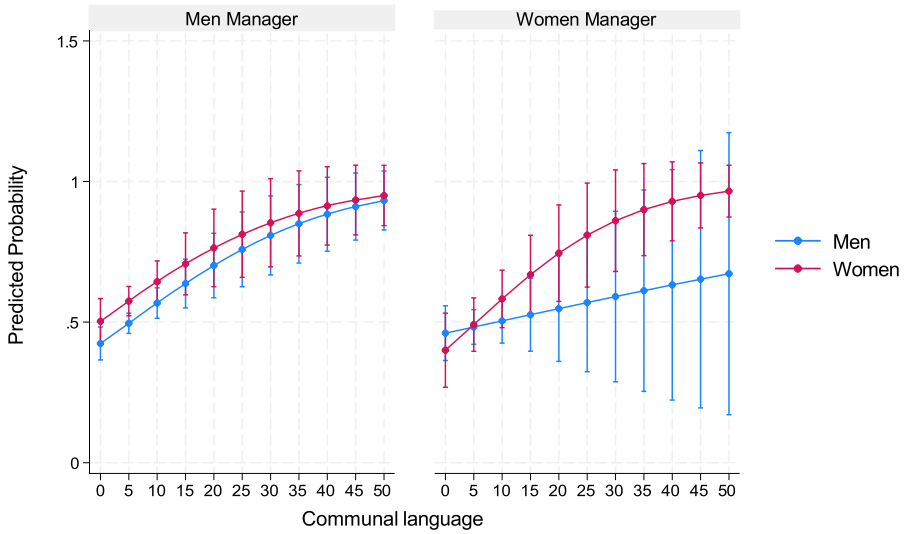


Fig. 5. Probability of Outperforming by gender of manager and employee and number of communal terms used in performance evaluation.

Table 4
Results: gender stereotypes in language.

Employee		
	Man	Woman
Man	A+C+	A+C+
Woman	A-C+	A+C+

so, thereby potentially creating backlash from men against working with them.

Our study reveals that there may be problems even in organizations that aim to redress imbalance in the pipeline to leadership positions (three-quarters of leaders are men and the proportion of women amongst assistant directors is 33 percent in our sample) committing to evaluation criteria that are clearly specified and applied equally to women and men. Since pay gap reporting has been a requirement, more women have been promoted in positions of leadership.

We do find that women are assessed as outperforming more than men, and that, unlike in other studies, agentive language benefits

them (Madera et al., 2009; Rudman et al., 2012). However, in line with the more recent findings of Baltrunaite et al. (2022) and of Correll et al. (2020) we find increasing returns from use of agentive terms for women more than men. In addition, we find that the gender gap in performance assessment in favor of women increases with the use of agentive language, as a consequence of the fact that, for men employees, the returns are constant while those for women increase. This suggests that agentive descriptions are not expected to be used when describing a woman; furthermore, to be credible, they need to be over-stressed, indicating that gender stereotypes persist.

This links to our main novel result: women managers not only need to use relatively more agentive terms to be effective in describing the leadership skills of women but, importantly, when they do so in supporting men' evaluations, this appears to backfire. This suggests that a note of caution is in order when recommending women managers to employ more agentive language (Manian and Sheth, 2021). The result that the returns for men decrease when women managers employ an increasing number of agentive terms, while the same does not occur when it is men managers to write the evaluation, is a strong indication that the use of agentive language by a woman to evaluate a man is a violation of the stereotype that a woman can assess competently the agentive and leadership skills of a man. This offers a novel insight into the debate about the operation of backlash, which occurs when individuals are perceived as violating status expectations about how their group should or should not behave. It is women's perceived status violations, for instance when acting and behaving aggressively, that produce backlash. Various studies have discussed and evidenced backlash against agentic women (Chakraborty and Serra, 2024; Exley et al., 2020; Bowles et al., 2007; Amanatullah and Morris, 2010; Amanatullah and Tinsley, 2013; Rudman and Phelan, 2008). In our case, the backlash is experienced by the men who are managed by women leaders: it is, therefore, a concerning finding given that it might lead to men not wanting to work for women leaders.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2026.105271.

Appendix A: Language Indicator

Because agentive and communal stances have been associated with gender bias in professional contexts, the decision was made to explore the extent to which agentive vs communal stance was expressed in language choice and how they differed in descriptions of female vs male counselees. Two language indicators were constructed, one including language choices pertaining to agentive stance and the second to communal stance. A written text is normally composed of a selection of content words, that is, words that carry meanings such as verbs, adjectives, adverbs and nouns, and function words, that is, words that do not have clearly defined meanings and are used as grammatical building blocks to construct sentences (e.g., articles, auxiliary verbs, prepositions etc). Because we were interested in agentive vs communal meanings, we considered content words only. From this group, we selected adjectives and verbs. As descriptive words, adjective describe and evaluate states, processes and relations, while verbs, as 'doing' verbs, express physical, mental, communicative and relational activities as well as causation and occurrence. Both parts of speech are central to the genre of performance review and thus, it was felt that analysing both types of content words is sufficient to capture the gist of the texts and identify agentive vs communal stance. We excluded nouns from our analysis because they expressed mostly roles, responsibilities, project names and other business entities that seemed less relevant to the agentive and communal stance. The performance reviews were divided into two text corpora, the first containing descriptions of male employees and the second of female employees. Both were uploaded onto the linguistic software programme Sketch Engine, which includes the Penn Treebank parser that automatically identifies parts of speech (PoS) in any text of English. The accuracy of the correct detection of POS by the parser can be up to 97% for written texts, which is high. Yet, there is always a scope for error also due to the fact that some words might belong to the different classes of PoS. We used Sketch Engine and the Penn Treebank parser to retrieve automatically all adjectives and verbs from our two data sets. A manual check was conducted to ensure that all retrieved adjectives and verbs were correctly identified and words classified inaccurately were removed. Subsequently, two researchers independently classified all adjectives and verbs into two main semantic, that is, agentive and communal. The meanings assigned to the verbs and adjectives were based on the ways in which the words were used in context reflecting organisational activities and practices. This can be seen in Fig. A1, which shows the main meanings associated with each category and examples of adjectives and verbs in each category. Using set dictionaries or automatic semantic classifiers would not have identified the diversity of context-specific meanings. After two researchers finished coding all retrieved adjectives and verbs, the coding agreement was calculated using Cohen's kappa coefficient, a statistic commonly employed to measure inter-rater reliability for qualitative (categorical) items. The score stood at 0.86 indicating a substantial agreement. Instances of disagreements were resolved by double checking the meanings of words in the texts. Subsequently, the frequency of agentive and communal words (adjectives and verbs) in each performance review and the proportion of those in each (as of the total number of words per review) was automatically calculated in Excel. Thus, we obtained a ratio agentive and communal stance as expressed in each feedback using agentive and communal verbs.

PoS: Verbs	Meanings of actions	Examples
Agentive	Beginning, initiating, creating, shaping and reshaping, adapting, adjusting, managing, achieving, completing, solving, expanding, aspiring	*start, introduce, commence change, refine, rework adapt, align, adjust organise, maintain, ensure deliver, achieve, complete lead, drive, accelerate*
Communal	Communicating, presenting collaborating, networking, mentoring being 'warm'	*ask, discuss, communicate join, collaborate, connect guide, coach, mentor respect, care, trust*
PoS: Adjectives	**Meanings/ descriptors of …**	**Examples**
Agentive	achievement, ambition, strong personality, dedication, management, agility/energy	*accomplished, experienced, strategic ambitious, motivated, driven confident, assertive, impactful dedicated, enthusiastic, passionate executive, managerial, organisational quick, fast, alert, dynamic, energetic*
Communal	collaboration, communication, mentorship, warmth	*collaborative, cooperative, interactive clear, articulate, persuasive consultative, parental, pastoral friendly, honest, pleasant, warm, nice*

Fig. A1. Examples of contextual verbs ad adjectives within the organisation's corpus.

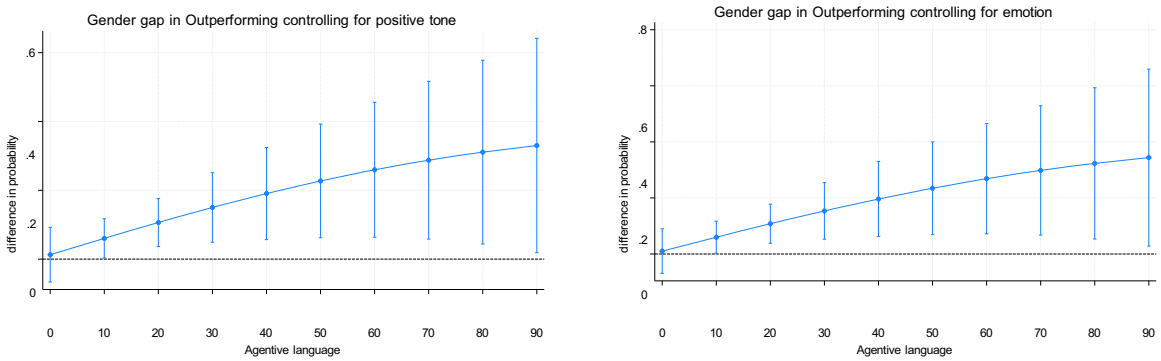
Appendix B: Sentiment Analysis

The use of computational linguistic automated tools to analyse language has become widespread. One of these tools is the Linguistic Inquiry and Word Count (LWIC), which analyses several dimensions of a text, including text polarity and sentiment, that is the negative, positive or neutral stance of a text. We employ LWIC to evaluate the positive and negative sentiment of each evaluation of our sample and compare the results with those obtained through manual annotations by linguistic experts.

With LWIC, 117 (out of 1054) evaluations were recognised as having some negative tone with scores from 1 to 10 (out of 100) which indicates low or negligible levels of negativity. However, when we scrutinise the top 'most negative' evaluations, we find that these do not have a negative connotation. For example, of the top 5:

- Ensure he does not take on too much. Leverage PERSON NAME and I to help push the client back where required. (LIWC negative tone score 10)
- PERSON NAME continues protecting the integrity of our analysis on profit warningsand beyond (LIWC negative tone score 7.69)
- Brilliant to work with, clued up on everything and all round smashingly brilliant (LIWC negative tone score 7.69)
- Calmness under pressure and also being very thorough, it is rare that there is a mistake in PERSON NAM work (LIWC negative tone score 5.26)
- Critical insights; building connections with team members bring the team together when facing difficulties bring diverse ideas and perspective (LIWC negative tone score 5.26)

This mis-attribution occurs because LIWC is based on generic dictionaries and analyses words independently of the context of their use in sentences. While the evaluations above include words that can have negative meanings as standalone words (push back, warnings, pressure, critical), in the context and in combination with other words they contribute to the positive tone and/or qualities of the evaluates ('critical insights', 'calmness under pressure' etc). A recent study by Hunter and Grant (2025) assessed the accuracy and reliability of LIWC and confirmed that the tool cannot be relied upon as the accuracy of identifying semantic content is below 50 percent. Given the very low accuracy, we find LIWC to be not a reliable and usable tool for the identification of semantic meanings in texts/evaluations. Human raters are in this respect reliable and more accurate reflecting also better the conditions of how such evaluations are read and interpreted (by decision-making humans). That being said, we consider two possible indicators of sentiment obtained from LWIC, one the positive tone of each written evaluation and the other the emotion. As discussed above, the former reflects the sentiment and general feeling of the text, while the emotion indicator reports the frequency of explicit emotion words in the text, focusing on the direct expression of feelings rather than overall sentiment. The consideration of these two indicators does not alter our results, as shown in Fig. B1: the gender gap in the probability of being assessed as outperforming remains the same as that shown in the left-panel of Fig. 3.



(a) Positive tone in LWIC **(b)** Emotion in LWIC

Fig. B1. Gender gaps in probability of Outperforming by number of agentive terms and sentiment indicators in written evaluations.

Appendix C: Integration to Benchmark model in equation 2

Table 5
Logistic Regression: Odds Ratios.

	(1) Original Model	(2) Manager FE	(3) With Sentiment	(4) Communal Terms	(5) Prop. Women manager
agentive lang.	1.001	1.014	1.003	1.000	1.000
	(0.0143)	(0.0252)	(0.0144)	(0.0151)	(0.00134)
employee gender	1.041	0.926	1.240	1.087	1.866
	(0.208)	(0.345)	(0.333)	(0.220)	(1.150)
employee gender × agentive lang.	1.021*	1.031	1.020	1.042**	0.999
	(0.0127)	(0.0210)	(0.0128)	(0.0216)	(0.00079)
manager gender	0.872	1	0.854	0.849	
	(0.129)	(.)	(0.127)	(0.127)	
Team=2	0.564*	0.179*	0.573*	0.538**	0.410
	(0.167)	(0.172)	(0.170)	(0.160)	(0.333)
Team=4	1.283	0.574	1.333	1.338	1.149
	(0.256)	(.)	(0.268)	(0.269)	(0.576)
Team=5	2.011***	.	2.079***	1.936***	1.258
	(0.345)	(.)	(0.359)	(0.334)	(0.589)
Team=6	0.947	0	0.946	0.927	0.919
	(0.190)	(.)	(0.190)	(0.186)	(0.484)
Text length	1.000	0.997	1.000	0.997	1.000
	(0.00177)	(0.00320)	(0.00178)	(0.00219)	(0.00017)
sentiment score			1.053**		
			(0.0224)		
employee gender × sentiment score			0.969		
			(0.0304)		
communal lang.				1.071***	
				(0.0253)	
employee gender × communal lang.				0.953	
				(0.0368)	
prop. women managers					1.138
					(0.874)
agentive lang. × prop. women manag.					0.998
					(0.00168)
employee gender × prop. women manag.					0.964
					(1.255)
employee gender × agentive lang. × prop. women manag.					1.002
					(0.00249)
Observations	1048	424	1048	1048	163

Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Data availability

The authors do not have permission to share data.

References

Amanatullah, E.T., Morris, M.W., 2010. Negotiating gender roles: Gender differences in assertive negotiating are mediated by women's fear of backlash and attenuated when negotiating on behalf of others. J. personal. soc. psychol. 98 (2), 2565.

Amanatullah, E.T., Tinsley, C.H., 2013. Punishing female negotiators for asserting too much or not enough: Exploring why advocacy moderates backlash against assertive female negotiators. Organ. Behav. Hum. Decis. Process. 120 (1), 110–122, 5.

Babcock, L., Laschever, S., Gelfand, M., Small, D., 2003. Nice girls don't ask. Harv. bus. rev. 81 (10), 14, 141.

Babcock, L., Recalde, M.P., Vesterlund, L., Weingart, L., 2017. Gender differences in accepting and receiving requests for tasks with low promotability. Am. Econ. Review 107 (3), 714–747, 1.

Baltrunaite, A., Casarico, A., Rizzica, L., 2022. Women in economics: the role of gendered references at entry in the profession. Cent. Econ. Policy Res. 1, 5.

Baxter, J., 2009. The language of female leadership, 1. Springer, p. 4.

Bertrand, M., Hallock, K.F., 2001. The gender gap in top corporate jobs. ILR Re- view 55 (1), 3–21, 1.

Blau, F.D., Kahn, L.M., 2017. The gender wage gap: Extent, trends, and explanations. J. econ. lit. 55 (3), 789–865, 1.

Boring, A., 2017. Gender biases in student evaluations of teaching. J. Public Eco- nomics 145 (C), 27–41, 1.

Born, A., Ranehill, E., Sandberg, A., 2022. Gender and willingness to lead: Does the gender composition of teams matter? Rev. Econ. Stat. 104 (2), 259–275, 1.

Bowles, H.R., Babcock, L., Lai, L., 2007. Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask. Organ. Behav. hum. decis. Process. 103 (1), 84–103, 5.

Castilla, E.J., 2011. Bringing managers back in: Managerial influences on workplace inequal- ity. Am. Sociol. Rev. 76 (5), 667–694, 1.

Cecchi-Dimeglio, P., 2017. How gender bias corrupts performance reviews, and what to do about it. Harv. Bus. Rev. 12 (04), 2017, 1.

Chakraborty, P., Serra, D., 2024. Gender and leadership in organisations: The threat of backlash. Econ. J. 134 (660), 1401–1430, 1, 5.

Cialdini, R. B. and M. R. Trost (1998). Social influence: Social norms, conformity and com- pliance. 1.

Correll, S.J., Weisshaar, K.R., Wynn, A.T., Wehner, J.D., 2020. Inside the black box of organizational life: The gendered language of performance assessment. Am. Sociol. Rev. 85 (6), 1022–1050, 1, 5.

Cuadrado, I., García-Ael, C., Molero, F., 2015. Gender-typing of leadership: Evaluations of real and ideal managers. Scand. J. Psychol. 56 (2), 236–244, 1.

del Carmen Triana, M., Song, R., Um, C.T., Huang, L., 2024. Stereotypical perception in management: A review and expansion of role congruity theory. J. Management 50 (1), 188–215, 1.

Della Giusta, M., Bosworth, S., 2020. Bias and discrimination: what do we know? Oxf. Rev. Econ. Policy 36 (4), 925–943, 1.

Dutt, K., Pfaff, D.L., Bernstein, A.F., Dillard, J.S., Block, C.J., 2016, November. Gen- der differences in recommendation letters for postdoctoral fellowships in geoscience. Nat. Geosci. 9 (11), 805–808, 1.

Eagly, A.H., Karau, S.J., 2002. Role congruity theory of prejudice toward female leaders. Psychol. rev. 109 (3), 573, 1.

Exley, C.L., Kessler, J.B., 2022. The gender gap in self-promotion. Q. J. Econ. 137 (3), 1345–1381, 1.

Exley, C.L., Niederle, M., Vesterlund, L., 2020. Knowing when to ask: The cost of leaning in. J. Polit. Econ. 128 (3), 816–854, 5.

Fiske, A.P., 1993. Social errors in four cultures: Evidence about universal forms of social relations. J. Cross-Cult. Psychol. 24 (4), 463–494, 1.

Grossman, P.J., Eckel, C., Komai, M., Zhan, W., 2019. It pays to be a man: Rewards for leaders in a coordination game. J. Econ. Behav. Organ. 161, 197–215, 1.

Hunter, M., Grant, T., 2025. Is liwc reliable, efficient, and effective for the analysis of large online datasets in forensic and security contexts? Appl. Corpus Linguist. 5 (1), 100118, 5.

Koenig, A.M., 2018, June. Comparing Prescriptive and Descriptive Gender Stereotypes About Children, Adults, and the Elderly. Front. Psychol. 9, 1086, 1.

Madera, J.M., Hebl, M.R., Martin, R.C., 2009. Gender and letters of recommendation for academia: agentic and communal differences. J. Appl. Psychol. 94 (6), 1591, 1, 5.

Manian, S. and K. Sheth (2021). Follow my lead: Assertive cheap talk and the gender gap *Management Science* 67 (11), 6880–6896. 1, 5.

Mengel, F., Sauermann, J., Zölitz, U., 2019. Gender bias in teaching evaluations. J. Eur. econ. assoc. 17 (2), 535–566, 1.

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. Psychol. women q. 26 (4), 269–281, 1.

Reuben, E., Sapienza, P., Zingales, L., 2014. How stereotypes impair women's careers in science. Proc. Natl. Acad. Sci. 111 (12), 4403–4408, 1, 1, 4.

Ridgeway, C.L., 2001. Gender, status, and leadership. J. Soc. issues 57 (4), 637–655, 1.

Rudman, L.A., Moss-Racusin, C.A., Phelan, J.E., Nauts, S., 2012. Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. J. exp. soc. psychol. 48 (1), 165–179, 5.

Rudman, L.A., Phelan, J.E., 2008. Backlash effects for disconfirming gender stereotypes in organizations. Res. organ. behav. 28, 61–79, 1, 5.

Schmader, T., Whitehead, J., Wysocki, V.H., 2007. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. Sex roles 57 (7), 509–514, 1.

Schnurr, S., Zayts, O., Schroeder, A., Le Coyte-Hopkins, C., 2020. it's not acceptable for the husband to stay at home': Taking a discourse analytical approach to capture the gendering of work. Work Organ. 27 (3), 414–434, 1.

Smith, D.G., Rosenstein, J.E., Nikolov, M.C., Chaney, D.A., 2019. The power of language: Gender, status, and agency in performance evaluations. Sex Roles 80 (3), 159–171, 1.

Trix, F., Psenka, C., 2003. Exploring the color of glass: Letters of recommendation for female and male medical faculty. Discourse Soc. 14 (2), 191–220, 1.

Wu, A.H., 2018, May. Gendered Language on the Economics Job Market Rumors Forum. AEA Pap. Proc. 108, 175–179, 1.