Government Equalities Office

Gender bias and performance feedback: an RCT

Research report

June 2021

Hannah Burd, Georgina Bremner, Andrew Schein and Monika Rudzeviciute, - Behavioural Insights Team, in partnership with Michael Yeomans (Imperial College) and Ariella Kristal (Harvard University)

Contents

List of figures

List of tables

Acknowledgements

We wish to acknowledge the roles that the following individuals had in this project:

- Ruari Johnson from YSC for supporting the technical implementation of the trial on the 360 degree feedback platform.

Executive summary

Executive Summary

If leadership performance evaluations demonstrate gender bias and reinforce harmful stereotypes, then they may hinder women's career aspirations and the likelihood of them being promoted at senior levels. Existing research finds that women are more likely to be described in terms of relationship-oriented (communal) attributes and less likely than men to be described as possessing agentic (task-focused) attributes.[1] Agentic attributes (e.g. assertiveness, competence, or persistence) are often valued more and considered more important for leaders, suggesting that men may be at an advantage when it comes to leadership progression.[2] Research also finds that women are more likely to receive vague feedback while their male counterparts receive more actionable feedback, which is more likely to support their career development.[3] In order to achieve gender equality in organisations, there needs to be a focus on reducing bias in and improving the quality of performance feedback for women to give them an equal chance to succeed.

The current study explored the language used in 360 degree feedback performance reviews carried out between 2018 and 2019, for 4,328 senior managers in a large UK public sector organisation (1,854 female, 2,149 male, and 325 other). The performance reviews included questions that cover strengths, areas for development, and overall performance as a leader. A randomised controlled trial (RCT) design was used to compare feedback provided by reviewers who saw 1) a 'business-as-usual' feedback interface versus 2) a treatment interface. The treatment interface included altered instructions for the open text response questions that encouraged reviewers to provide specific examples, advice on how to improve, and where applicable to include feedback on both relationship-oriented and task-oriented skills.

Study aims

Our research had three key aims:

1. To provide insights into how specificity of feedback may differ across genders using novel natural language processing (NLP) methods.

2. To provide insights into the use of gender stereotyped language in feedback using dictionary methods (searching for specific words or phrases thought to be associated with a particular gender).

3. To evaluate the impact of an intervention (designed to remind colleagues to provide feedback that is specific and also considers a colleague's relationship- *and* task-oriented attributes) on both the specificity of feedback given and also the prevalence of gender stereotyped language.

[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*(3-4), 159-171.
[2] Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin*, 111, 3–22; Eagly, A. H., & Karau, S. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109, 573–598; 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*(3-4), 159-171.
[3] Correll, S., & Simard, C. (2016). Vague feedback is holding women back. *Harvard Business Review.*

Methods

The trial sample included a total of 25,627 unique reviewers from the 2018-2019 performance review round who wrote a total of 46,176 unique feedback responses for 4,328 unique review subjects. We randomised reviewers into a treatment or a control group and we changed the phrasing of some of the treatment group's questions, for example asking reviewers to provide specific and actionable feedback (see Appendix A). To further counteract potential gender bias in responses, we also altered the wording of one question to encourage reviewers to provide feedback on both relationship-oriented (communal) and task-oriented (agentic) attributes.

We used a natural language processing (NLP) algorithm to measure the specificity of feedback provided. We also used dictionary methods to understand the presence of gender bias in feedback. Finally, we used a machine learning model to predict the review subject's gender based on the language in the review, analysing whether predictability of gender differed between the treatment and the control conditions. If the algorithm was less able to predict the review subject's gender based on the language used in the treatment condition, we theorised that the treatment would have successfully reduced gender bias.

Results

Specificity of feedback	
Specificity of feedback, by **gender**, under **control** conditions	Women received more specific feedback on their strengths, but less specific feedback on development. There was no difference between men and women in the specificity of feedback for the overall question.
Specificity of feedback under **treatment** conditions	The treatment increased the specificity of feedback for the overall question, did not affect the specificity of the development question, and decreased specificity in the strengths-based question.
Specificity of feedback by **gender** under **treatment** conditions	The treatment did not have a different effect on the specificity of language used for men and women.
Gendered language	
Use of communal and agentic language, by **gender**, under **control** conditions	Women received more communal words *and* agentic words than men in their strengths-based feedback, but no differences between men and women were observed on the other feedback questions.
Predicting gender from feedback under **control** conditions	Gender could be predicted from feedback due to subtle linguistic differences in reviews for women versus men, which were not picked up by the specificity algorithm or the communal / agentic dictionaries.

Predicting gender from feedback under **treatment** conditions	As under control conditions, gender could be predicted from feedback in the treatment condition. Importantly, however, the algorithm was no less able to predict gender in the treatment condition indicating that the intervention had not reduced linguistic bias.
Additional exploratory analyses	
Numeric ratings	Asking whether someone is 'accessible and approachable' appears to reflect more closely their overall job performance than asking if they are 'visible and approachable at all times'.
Benevolent sexism	Data showed that women received more positive words in their praise and 'hedges' in their development feedback than men even when controlling for performance quality.

Conclusions

In the context of this part of the public sector in the UK, our findings indicate that women receive more specific feedback about strengths and less specific feedback than men on areas to develop, which is in line with existing research.

We found that women were more likely to receive equal amounts of feedback about communal and agentic attributes, which is contrary to the existing academic thinking on stereotypes about women in leadership. However, our exploratory evidence that women were more likely than men to receive tentative feedback suggests that some gender stereotypes may persist. In particular, feedback for women seemed to be more positive and less direct, even controlling for performance quality, suggesting that development may not be prioritised for female recipients, which is consistent with the literature on benevolent sexism.

The intervention did not systematically increase the specificity of feedback, which may be due to the treatment condition instructions attempting to fulfil additional aims - notably increasing actionability and also reducing gender bias by encouraging feedback on both communal and agentic traits. The business-as-usual approach (control condition) already asked people to provide a defined number of examples, and this may also have provided a similar amount of encouragement to be specific as the instructions in the treatment condition.

Finally, we found some interesting, albeit tentative, results on asking people to rate someone on their 'visibility' versus 'accessibility' which suggest that 'accessible' correlates better with overall job performance than 'visible'. We speculate that this distinction seems particularly important for remote and flexible work, where accessibility may be more achievable than visibility. We encourage employers to avoid questions of job 'inputs' (i.e. hours worked or location of work) and instead to focus on 'outputs' (i.e. objectives met, tasks delivered, team successes).

We suggest multiple future research directions that could be fruitful. These include establishing how language used in feedback may be linked to the perceived usefulness of the feedback from the review subject's perspective, as well as how it may be linked to career progression outcomes. We also would be interested to test different interventions targeted at reducing benevolent sexism

in the feedback process to bring women's feedback more in line with the feedback that men receive. Finally, there is promise in building on the current findings to further test different ways of structuring performance feedback forms in order to get reviewers to give more actionable feedback.

Introduction

1. Introduction

Feedback is essential for performance improvement, as it provides advice for future actions recipients can take which can lead to career progression.[4] However, providing high quality performance feedback to colleagues is challenging.[5] The literature suggests that women may be at a disadvantage in performance evaluations to men for two reasons:

1. They are more likely to receive feedback about relationship-oriented attributes, which are perceived as less important for leadership than agentic attributes (which are more likely to be ascribed to men).[6]

2. They receive feedback that is less critical and more vague, both on what they do well and what they need to improve.[7] [8] For example, a study looking across three firms found that while praise for women lacked ties to specific outcomes (e.g. "you had a great year"), men were more likely to be given concrete advice on specific skills to develop (e.g. "you need to deepen your domain knowledge in the X space — once you have that understanding, you will be able to contribute to the design decisions that impact the customer."[9]

Each year senior managers and leaders in this large public sector organisation undergo a performance review. Managers nominate reviewers (e.g. their managers, peers, direct reports) they would like to receive feedback from. The feedback form includes 1-5 Likert scale questions about leadership qualities as well as open text response questions where reviewers can comment on what the review subject has done well and suggest areas for development. In keeping with the wider literature, previous text analysis of the open text feedback, covering the performance year 2017-2018, found that certain words related to communal qualities (such as 'emphatic', 'warm'

[4] Ashford, S. J., & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational behavior and human performance*, 32(3), 370-398; Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, 101(2), 127-151.

[5] Bies, R. J. (2013). The delivery of bad news in organizations: A framework for analysis. *Journal of Management*, 39(1), 136-162; Dibble, J. L., & Levine, T. R. (2010). Breaking good and bad news: Direction of the MUM effect and senders' cognitive representations of news valence. *Communication Research*, 37(5), 703-722; Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2), 254; Schaerer, M., Kern, M., Berger, G., Medvec, V., & Swaab, R. I. (2018). The illusion of transparency in performance appraisals: When and why accuracy motivation explains unintentional feedback inflation. *Organizational Behavior and Human Decision Processes,* 144, 171-186.

[6] 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*(3-4), 159-171.

[7] Jampol, L., & Zayas, V. (2020). Gendered White Lies: Women Are Given Inflated Performance Feedback Compared With Men. *Personality and Social Psychology Bulletin*, 0146167220916622; McKinsey (2016). Women in the Workplace 2016. *Lean In and McKinsey & Company*. Retrieved February, 27, 2017.

[8] Correll, S., & Simard, C. (2016). Vague feedback is holding women back. *Harvard Business Review.*

[9] Ibid.

and ‘caring’) were more likely to show up in reviews about women, whilst certain words related to agentic qualities (such as ‘analytical’, ‘methodical’ and ‘intellectual’) were more likely to appear in reviews about men.

In order to ensure a level playing field in performance reviews, this research aimed to improve the quality of feedback provided to all review subjects. Feedback to help people improve is often theorised to be more effective when it includes specific, actionable suggestions that can be followed, rather than abstract evaluations.[10] This research applied this theory, and tested the effectiveness of an intervention aiming to remind reviewers to provide more specific and actionable feedback. We also altered instructions where applicable to encourage reviewers to comment on both communal and agentic attributes of the review subject.

We hypothesised that the intervention would address potential gender bias in two ways:

1. By encouraging provision of specific examples and actionable feedback, reviewers would have to be more deliberative in their responses, which may reduce reliance on stereotypes in their feedback.

2. By encouraging reviewers to provide feedback on both communal and agentic attributes, we expected that both male and female review subjects would be more likely to receive feedback about both types of attributes.

In this way, this study sought both to increase the specificity of feedback and to reduce the extent to which language used is affected by gender biases.

[10] Reyt, J. N., Wiesenfeld, B. M., & Trope, Y. (2016). Big picture is better: The social implications of construal level for advice taking. Organizational Behavior and Human Decision Processes, 135, 22-31; Kraft, M. A., & Rogers, T. (2015). The underutilized potential of teacher-to-parent communication: Evidence from a field experiment. Economics of Education Review, 47, 49-63

Research aims and trial methodology

2. Research aims and trial methodology

This research explored the language used in performance reviews carried out between 2018 and 2019 and had three key aims:

1. To provide insights into how specificity of feedback may differ across genders using novel natural language processing (NLP) methods.

2. To provide insights into the use of potentially gender stereotyped language in feedback using dictionary methods.

3. To evaluate the impact of an intervention designed to remind colleagues to provide more specific feedback on both specificity of the language used as well as the use of gender stereotyped language.

2.1 Hypotheses

Specificity of feedback

- Feedback for women will be less specific than for men under control conditions.

- The intervention will make feedback more specific.

Gender biased language

- Women will receive more communal words in their reviews than men, under control conditions.

- Men will receive more agentic words in their reviews than women, under control conditions.

- The intervention will reduce the difference in language used (in terms of communal and agentic language) between men and women.

- The intervention will make the review subject's gender harder to 'predict' based on the language used in the review due to it being more similar across the two genders.

2.2 Sample selection

The participants in this trial consisted of all nominated reviewers invited to provide feedback in the senior management performance review cycle for 2018-2019. This included a total of 26,482 different reviewers, who provided feedback in a total of 51,223 reviews for 5,037 review subjects. However, 709 subjects were only present in self-reviews. When we removed these reviews, we were left with a sample of 25,627 unique reviewers who wrote a total of 46,176 unique feedback responses for 4,328 unique review subjects.

2.3 Outcome measures

The outcome measures for this study are summarised in Table 1.

Primary outcome measure
Specificity: The primary outcome measure of this trial was specificity of the language used in the open text feedback. Our definition of specificity overlaps with that of 'concreteness', which linguistically is defined as the degree to which a concept denoted by text refers to a perceptible entity, and is the opposite of abstract.[11] We used a function called doc2concrete to measure specificity of the language in the three open text questions in the reviews. Doc2concrete had been trained on multiple feedback datasets, where text was rated by human coders according to how concrete the language was.[12] For further details see Appendix B.

Secondary outcome measures
Gender stereotyped language: A secondary outcome measure in this trial was the use of gender stereotyped language as assessed by the use of 'agentic' and 'communal' language using Linguistic Inquiry and Word Count (LIWC) dictionaries, as specified by Gaucher, Friesen & Kay (2011)[13]. These dictionaries were developed as a measure of gender bias in job advertisements, based on the underlying theory that female dominated roles are described using more communal language than male dominated roles, while male roles are described using more agentic language than female roles. For the full list of words see Appendix C.

Exploratory outcome measures
Predictability of gender: An exploratory outcome measure was the predictability of gender within treatment and control conditions, measured by the area under the receiver operating characteristic curve (referenced in this document as the AUC, the 'area under the curve'). AUC measures the performance of a prediction algorithm on a scale from 0 to 1. A higher AUC indicates a more discriminating algorithm, where, for example, AUC = 1 in a gender-prediction algorithm indicates 100% correct identification of gender.

[11] Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3), 255.
[12] Advice model from the doc2concrete package by Yeomans, M. Concreteness, Concretely. (2020 - currently in revision and resubmission). A Case Study for Open Science in Natural Language. *Organizational Behavior and Human Decision Processes*.
[13] 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.

Table 1 - Outcome measures

Outcome measure	*Measured by*
Primary: Specificity of language	Scores from doc2concrete package's pre-trained advice model (these scores were then standardised to a mean of zero)[14]
Secondary: Use of potentially gender stereotyped language	Average number of words from Gaucher, Friesen & Kay's (2011) 'agentic' and 'communal' language dictionaries used per document
Exploratory: Predictability of gender within treatment and control conditions	Area under the curve (AUC)

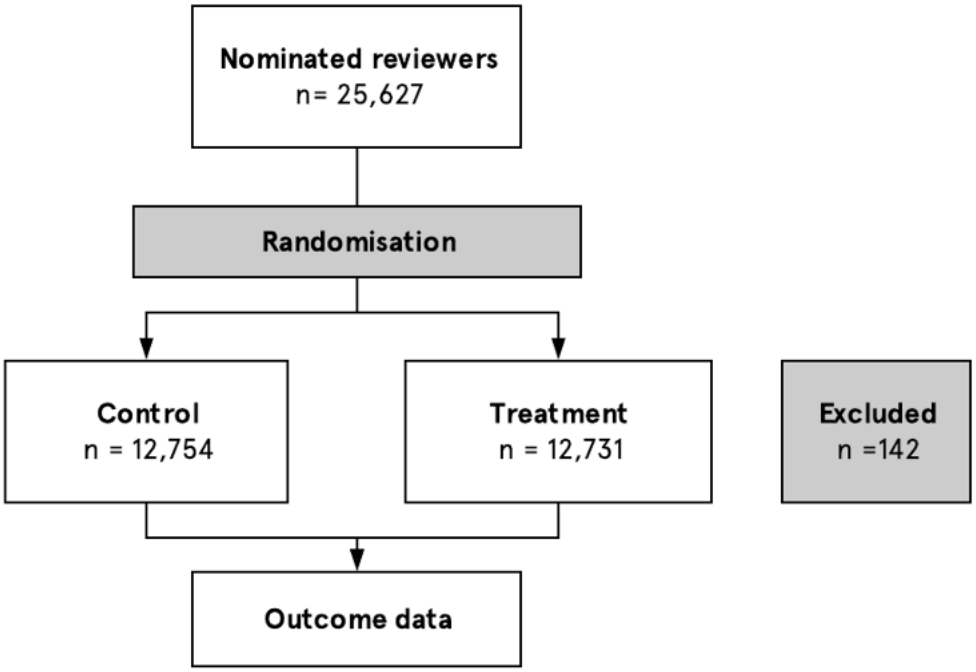
2.4 Randomisation

We conducted the study as a randomised controlled trial (RCT). Each reviewer who was nominated to provide feedback was randomly assigned to either a treatment or a control group. Individuals were identified by unique email addresses. Once assigned to the treatment group, the reviewer would see the same version (treatment or control) of the feedback form for all of their review requests. As a result, we can causally infer that differences in outcomes between the two reviewer groups are due to the redesigned feedback form that the treatment group was exposed to.

A total of 12,754 reviewers were assigned to the control group and 12,731 were assigned to the treatment group (see Figure 1). An additional 142 were labelled as having been included in both conditions (due to having been nominated via two email addresses) and were dropped from the analyses. The final sample included 45,693 reviews carried out by 25,485 reviewers for 4,328 review subjects.

[14] Advice model from the doc2concrete package by Yeomans, M. Concreteness, Concretely. (2020 - currently in revision and resubmission). A Case Study for Open Science in Natural Language. Organizational Behavior and Human Decision Processes.

Figure 1 - Trial design



2.5 Review subject characteristics

Only review subjects were asked to report their gender. We do not have gender data on the nominated reviewers. The gender composition of the review subjects was:

43.2% of reviews were for women,

49.6% of reviews were for men,

7.2% of reviews were for people who did not disclose their gender or selected ‘Other’.

Review subjects of different genders nominated a similar distribution of reviewers (i.e. line managers, peers, direct reports and ‘others’). Balance checks also confirmed that there were no systematic differences in the review subject gender composition of the control and treatment groups.

2.6 Intervention design

The intervention made alterations to the instructions for the open text response sections. The scope of this intervention was constrained by the fact that for each review subject, once their nominated reviewers had completed their reviews, an automated process combined the different reviews into one report. This meant that the questions across the two forms (control and treatment) had to be compatible in terms of question numbers and expected content. So while the

nature of the question could not be changed significantly, there was scope to remind reviewers in the treatment group to be more specific in their feedback.

The intervention primarily focused on the three open text response questions at the end of the review form, each of which asked about a different characteristic of the review subject:

- The first, 'Strengths', asks for information about the review subjects' strengths.
- The second, 'Development', asks about areas for improvement.
- The third, 'Overall', asks for a more general assessment of the review subjects' leadership performance.

Table 2 below shows how the language in the instructions for these three open text questions across the two forms differed. Key differences were that questions in the treatment condition explicitly asked reviewers to:

- Focus on both communal and task-oriented skills,
- Provide concrete examples,
- Provide suggestions on how they could improve.

Table 2 - Instructions provided for the open text response questions across the treatment and control groups

Open text question	*Control*	*Treatment*	*Rationale behind treatment*
Question 1: 'Strengths'	What are their main strengths as a leader and why? Please include up to three examples.	What are their main strengths as a leader? Please include examples of **both** how they relate to others **and** their leadership of their team's objectives.	Based on the findings that women are more likely to be described in communal terms and men in agentic terms,[15] we aimed to remind reviewers to consider both communal as well as task-oriented skills, to counter possible use of gender stereotyped language.

[15] 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(3-4), 159-171.

Question 2: ‘Development’	What are their main areas to develop as a leader? Please include up to three examples.	What are their main areas to develop as a leader? Please include concrete examples of what they could do to achieve this.	Based on the findings that specific and actionable feedback is more useful for development,[16] we aimed to remind reviewers to provide concrete examples as well as advice to increase the actionability of the feedback.
Question 3: ‘Overall’	Overall, please state if you feel they provide a good role model as a [Leader in this organisation] and how they demonstrate this?	Overall, are they a good role model as a [Leader in this organisation]? Please provide specific examples to explain your answer and if needed what actions they could take to improve.	As above, we aimed to remind reviewers to provide specific examples as well as advice to increase the actionability of the feedback.

[16] Reyt, J. N., Wiesenfeld, B. M., & Trope, Y. (2016). Big picture is better: The social implications of construal level for advice taking. *Organizational Behavior and Human Decision Processes*, 135, 22-31; Kraft, M. A., & Rogers, T. (2015). The underutilized potential of teacher-to-parent communication: Evidence from a field experiment. *Economics of Education Review*, 47, 49-63

Trial results

3. Trial results

3.1 Response rates

As shown in Table 3, reviewers in the treatment group were statistically significantly less likely to start a review and were less likely to provide open text feedback. These results imply the treatment condition raised the expected effort of the writing task.

Table 3 - Response rates across treatment groups

Reviewer response rate	*Treatment*	*Control*	*Statistical significance of the difference*
Did not start the assigned review	23.9%	22.8%	Statistically significant difference $X^2(1) = 7.7$, $p = .005$)
Of those who did access the feedback system, those that did not provide open text feedback	19%	15.4%	Statistically significant difference $X^2(1) = 108$, $p < .001$)

Table 4 summarises the average number of words present in the open text feedback for each of the three questions. Feedback about strengths was on average the longest, whilst overall feedback was the shortest. When considering gender differences, men received statistically significantly shorter feedback about their strengths as well as shorter feedback on the overall question. There was no difference in the length of the feedback received by men and women for the development question.

Table 4 - Response rates per question type

Outcome	*Question type*		
	Strengths	*Development*	*Overall*
Average length of response overall (conditional on response including at least one word)	52.7 words	48.2 words	34.2 words

Gender difference in the number of words used	Male subjects received fewer words (β = 3.32 fewer words for males, SE = .85, t(30572) = 3.9, p < .001).	Not statistically significantly different (β = 1.01 fewer words for males, SE = .81, t(27271) = 1.2, p = .216).	Male subjects received fewer words (β = 1.45 fewer words for males, SE = .51, t(28934) = 2.8, p = .005).
Difference in the number of words used, by treatment	The treatment condition had a higher number of very short responses. However, this was driven entirely by non-responses.		
Non-response, by treatment	35.0 % (treatment) vs 31.2% (control)	42.5% (treatment) vs 38.1% (control)	38.7% (treatment) vs. 34.7% (control)

3.2 Numeric ratings

Before the open text response questions, respondents rated the review subjects on a series of numeric scales (1-5) intended to assess their leadership qualities. A full list of numeric questions asked and a summary of those ratings by gender is given in Appendix D. Overall, the ratings all tended to correlate with one another, suggesting a general difference in performance quality that explained most of the variation in the numeric ratings. A factor analysis revealed that the questions focused on results delivery tended to be more correlated with each other, while the questions focused on relational and communication performance correlated more tightly with one another. Interestingly, we found that women were rated more highly than men, both on the overall average ratings and when looking at averages for the relational and communication performance questions and for the results delivery questions separately (all $p < .001$).

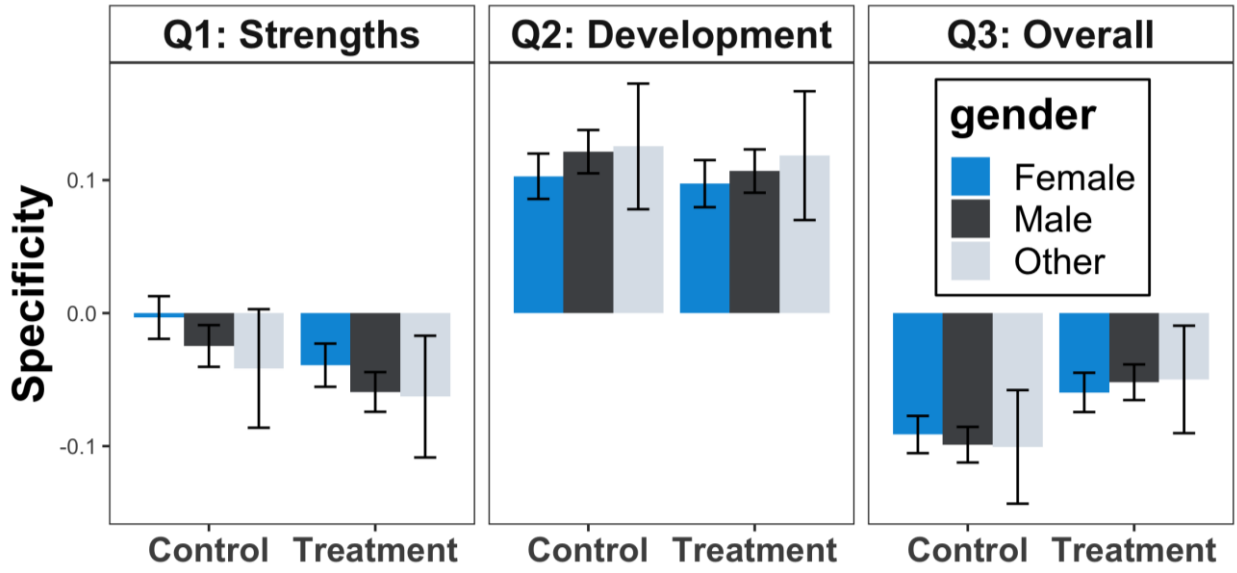
The numeric questions were identical across conditions, with one exception. The second question related to empowering teams (part of the relational and communication performance grouping) was worded as “Being visible and approachable at all times” in the control condition, but changed to “Being accessible and approachable” in the treatment condition. The treatment question was given higher ratings on average than the control question ($p < .001$). However, there was no interaction with gender ($p = .288$), and like the other numeric questions, women tended to get higher ratings on this question than men across both conditions ($p = .015$). Still we do find some interesting results for the treatment version of the question. That is, the treatment version was more highly correlated with the average of the other numeric ratings than the control version (interaction effect: $p = .007$), suggesting that ‘accessible’ is a more relevant trait for overall job performance than ‘visible’. We speculate that this distinction seems particularly important for remote and flexible work, where accessibility may be more achievable than visibility. We also note that one could question the value of accessibility (a job ‘input’) compared to questions which instead ask about what a leader delivers (an ‘output’). Unfortunately we could not compare to a third arm which did not ask about inputs at all, nor could we link data with individuals’ actual working patterns in order to make further

inferences. We caution against over-extrapolation of this finding as this analysis was exploratory and not pre-registered.[17]

3.3 Specificity of feedback

We measured specificity of the language used across the three open text response questions using a pre-trained NLP algorithm.[18] These scores were then standardised to a mean of zero. In Figure 2, we display the overall means for each open text question, analysed separately for each group.

Figure 2 - Average standardised specificity scores for all three text responses, by gender and treatment assignment



Average standardised specificity scores for all three text responses, by gender and treatment assignment. Each point represents a group mean and 95% CI.

Specificity at baseline
In the control group, women received more specific feedback than men for the question about their strengths (p = .002), and less specific feedback about their areas for development (p = .016), with no difference between the genders in feedback for the overall question (p = .146).

In a subgroup analysis, we analysed the data based on the reported relationship between the review subjects and their reviewer. We found the largest effect among direct reports - in the control condition they gave less specific feedback to their male

[17] You can view the pre-registration at this link

[18] Advice model from the doc2concrete package by Yeomans, M. Concreteness, Concretely. (2020 - currently in revision and resubmission). A Concrete Example of Construct Construction in Natural Language. Organizational Behavior and Human Decision Processes.

managers than female managers for both the strengths-based question (p = 0.0017) and the overall question (p = 0.0032)

Overall, our findings are inconsistent with the hypothesis that women systematically receive less specific feedback than men.

Effect of treatment on specificity
Across all genders, the treatment had a mixed impact on specificity of the language used across the three different question types. For the strengths question, counter to our expectation, the treatment reduced the specificity of the feedback provided (p < .001). For the overall question, it increased the specificity of feedback written (p < .001), and it had no effect on the specificity of feedback for the development question (p = .488). These findings were even more pronounced when looking at the subgroup of direct reports providing feedback to their line managers.

The treatment did not interact with gender on any of the three open text questions. This means that the treatment did not have a different effect on specificity of the language used to give feedback to men and women.

3.4 Gender biased language

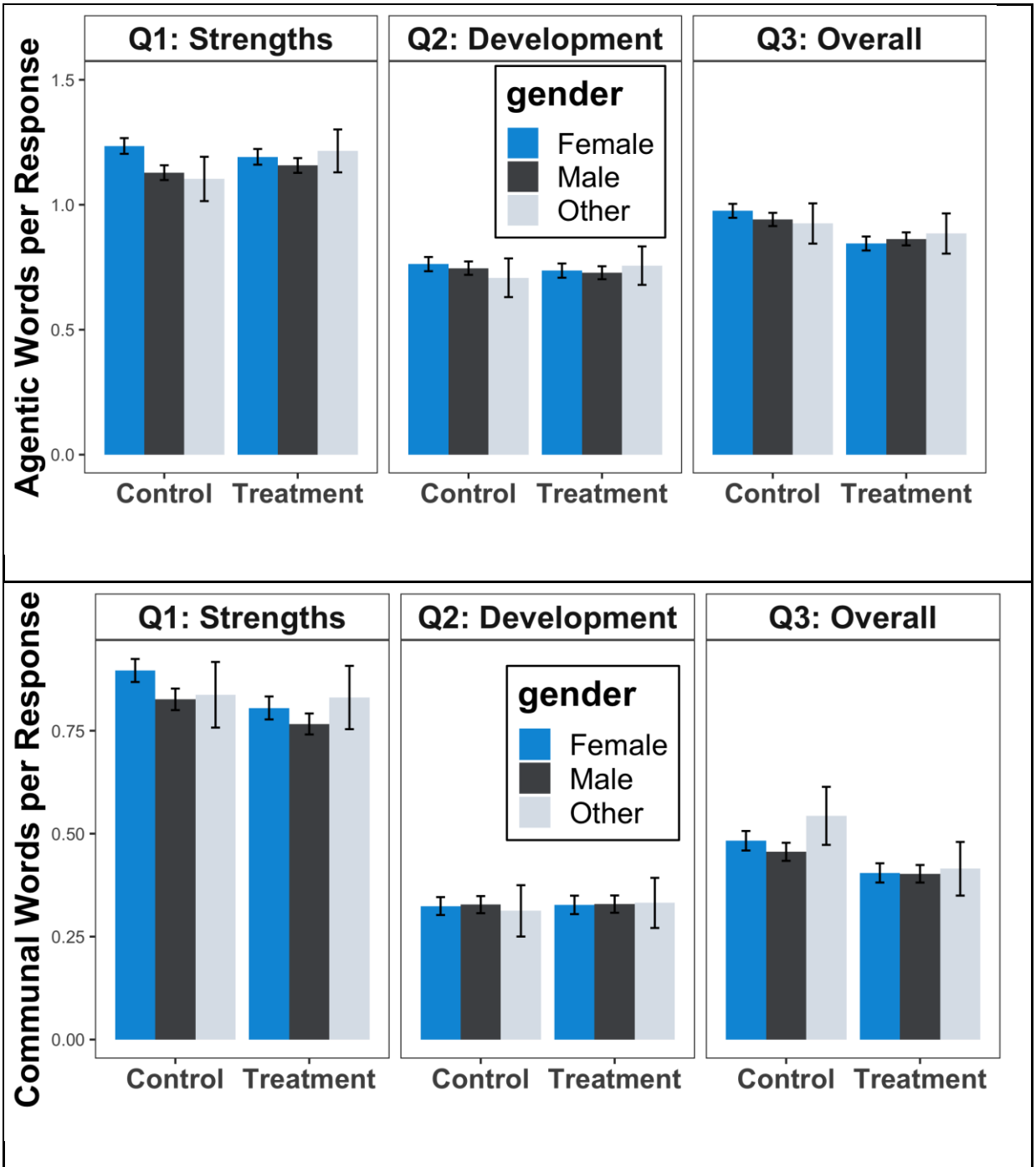
We also analysed the feedback to test for the use of 'agentic' and 'communal' language using dictionaries created by Gaucher, Friesen & Kay (2011)[19].

In Figure 3, we show the average number of words from each of these dictionaries used per review, separately for each treatment group and for each of the three questions. Our results suggest limited support for this underlying model of gender bias. While women received nominally more communal words in their feedback about their strengths than men (p = .001), they also received more agentic words in their strengths-based feedback than men did (p < .001). The gender differences in agentic or communal language used in the other questions were not significant. This is likely driven by the fact that strengths-based feedback for women contains significantly more words than it does for men. More tentatively, this is also in line with the recent research that finds that counter to traditional thinking, women leaders are characterised as possessing both agentic and communal traits.[20]

Overall, we did not find robust treatment effects on the use of gendered language, nor an interaction between treatment and gender. This means that the treatment did not change the relative use of agentic or communal words, for either men or women, to a statistically significant extent for any of the three questions.

[19] 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.
[20] Griffiths, O., Roberts, L., & Price, J. (2019). Desirable leadership attributes are preferentially associated with women: A quantitative study of gender and leadership roles in the Australian workforce. *Australian Journal of Management*, 44(1), 32-49; Berkery, E., Morley, M., & Tiernan, S. (2013) Beyond gender role stereotypes and requisite managerial characteristics: From communal to androgynous, the changing views of women. *Gender in Management: An International Journal* 28: 278–298.

Figure 3 - Group averages for the rate of agentic and communal words in response to each question



Each point represents a group mean and 95% CI.

3.5 Predictability of review subject's gender based on the language used in the feedback

An exploratory outcome measure in this study was predictability of the review subjects' gender within treatment and control conditions. We trained a machine learning model to estimate the gender of the subjects using the text of each response. The results suggest that gender was indeed identifiable within the text,

and to roughly equal extent for each question, due to linguistic differences including the following examples:

- Feedback about women contained greater use of object pronouns (“her / hers”), whereas feedback about men contained greater use of subject pronouns (“he”).

- Feedback about women contained greater use of subjective phrasing (using the first person pronoun, rather than making definitive statements, and using hedging phrases, such as “maybe” and “sometimes”).

- Women’s strengths-based feedback was more likely to include mentions of “her team”, while men’s strengths-based feedback was more likely to include terms such as “knowledge”, “strategy”, “calm”, and “thinker”.

However, the treatment did not appear to impact the predictability of review subjects’ gender: we found almost no difference in the predictability of gender between the two conditions for the strengths question (treatment: (AUC = .608, 95% CI= [.597,.618]; control: AUC = .603, 95% CI= [.592,.613]), the development question (treatment: AUC = .610, 95% CI= [.599, .620]; control: AUC = .596, 95% CI= [.586,.610]), or the overall question (treatment: AUC = .605, 95% CI= [.594,.616]; control: AUC = .609, 95% CI= [.599,.620]). This suggests that the treatment did not reduce the total amount of this gender bias in reviews.

3.6 Benevolent sexism

As mentioned above, we found that women did not in general elicit less specific reviews - it was only on the subject of their weaknesses where reviewers were less willing to be specific compared to reviews submitted for men. This pattern is in fact consistent with the literature on benevolent sexism.[21] To explore this further, we processed each text response using the politeness R package[22] and found that in both the control and the treatment condition women consistently received more positive emotion words than men in all three questions of their reviews - this persisted even for the question that specifically requested development feedback. Additionally, on the development question respondents employed subjective language, using the first person pronoun, rather than making definitive statements, and often included hedges to soften their claims, e.g. "I think that Anna needs to improve..." vs. "Andrew needs to improve...". We summarise these results in Appendix E.

---

[21] Dardenne, B., Dumont, M., & Bollier, T. (2007). Insidious dangers of benevolent sexism: consequences for women’s performance. *Journal of personality and social psychology*, 93(5), 764; Dumont, M., Sarlet, M., & Dardenne, B. (2010). Be too kind to a woman, she’ll feel incompetent: Benevolent sexism shifts self-construal and autobiographical memories toward incompetence. *Sex Roles*, 62(7-8), 545-553; Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology*, 70(3), 491; Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, 56(2), 109; Hideg, I., & Ferris, D. L. (2016). The compassionate sexist? How benevolent sexism promotes and undermines gender equality in the workplace. *Journal of Personality and Social Psychology*, 111(5), 706; Jones, K., Stewart, K., King, E., Morgan, W. B., Gilrane, V., & Hylton, K. (2014). Negative consequence of benevolent sexism on efficacy and performance. *Gender in Management: An International Journal*.

[22] Yeomans, M., Kantor, A., & Tingley, D. (2018). The politeness Package: Detecting Politeness in Natural Language. *R Journal*, 10(2).

In an effort to understand what was underlying these findings, we explored a hypothesis that women were simply performing better than men. We analysed the numeric ratings at the beginning of the feedback survey using a factor analysis and found that performance quality indeed had a strong effect on many of these measures, increasing the amount of positivity in a person's reviews. However, we also found that the gender bias was robust when controlling for performance. That is, women who were judged to be performing equally well as men still received more positive praise and less direct feedback. This finding is consistent with the theory that benevolent sexism was affecting women's reviews.

Discussion

4. Discussion

Many of our findings from this analysis go counter to our main expectations. In Table 5 below we summarise our expectations and our findings.

Table 5 - Expectations and findings

Expectation	***Findings from analysis***
Specificity of feedback	
Feedback for women is less specific than for men in the control.	Findings were **mixed**. Women received more specific feedback on strengths, but less specific feedback on development compared to men.
The treatment makes feedback more specific.	Findings were **mixed**. The treatment did increase specificity for the overall question, did not affect the specificity of the development question, and decreased specificity for the strengths-based question.
If women receive less specific feedback, the treatment will work especially well for women at making the feedback more specific.	The data **did not confirm** this expectation. The treatment did not have a different effect on specificity of the language used for men and women.
Gendered language	
Women receive more communal words in their reviews than men in the control condition.	The data **partly confirmed** this expectation. Women received more communal words in their strengths-based feedback, but no differences were observed in the other feedback questions.
Men receive more agentic words in their reviews than women in the control condition.	The data **did not confirm** this expectation. Women received more agentic words than men in their strengths-based feedback, but no differences were observed in the other feedback questions.
The treatment will reduce gender differences in the use of the communal language.	The data **did not confirm** this expectation.
The treatment will reduce gender differences in the use of the agentic language.	The data **did not confirm** this expectation.

The treatment will make review subjects' gender harder to predict based on the language used in the reviews.	The data **did not confirm** this expectation.
Exploratory analysis of benevolent sexism (not pre-registered)	
Data showed that women received more positive words in their praise and 'hedges' in their development feedback than men even when controlling for performance quality.	

4.1 Specificity of feedback

The fact that the treatment reduced specificity of the strengths-based question was surprising. One explanation could be the existence of a complex interplay between the valence (positive versus negative) of feedback, its specificity, and politeness. Prior research shows that positivity in reviews is generally correlated with a lack of specificity – that is, in general people tend to be either polite or direct.[23] This is consistent with our finding that feedback for the development question (negative valence) was much more specific than the strengths-based (positive valence) feedback on average. This may partly explain why our treatment was less effective at increasing the specificity of the strengths-based feedback and may also partly explain why the treatment was not effective at increasing the specificity of the developmental feedback, as it was already relatively high in specificity.

On the other hand, the changes to the instructions of the open text feedback may also go some way to explain these findings. For the strengths-based question, the original text asked reviewers to "include up to three examples", while the treatment text did not ask for a specific number of examples but instead asked reviewers to focus on providing feedback across a range of attributes ("please include examples of **both** how they relate to others **and** their leadership of their team's objectives"). We posit that asking people to list "up to three examples" in the control condition may have already encouraged specificity. By changing this wording we may have inadvertently reduced the extent to which the question encouraged specificity.

Similarly, when we consider the instructions for the developmental feedback question, the control text again asked reviewers to "include up to three examples", while the treatment text asked reviewers to "include concrete examples of what they could do to achieve this". The two conditions may have included approximately equal levels of encouragement to be specific.

Finally, for the overall question, where the treatment *did* significantly increase specificity, the control text did not ask for specific examples but the treatment text did ("Please provide specific examples to explain your answer and if needed what actions they could take to improve."). Asking for specific examples therefore appears to be an important factor in successfully increasing specificity of the feedback.

[23] Correll, S., & Simard, C. (2016). Vague feedback is holding women back. Harvard Business Review; Advice model from the doc2concrete package by Yeomans, M. Concreteness, Concretely. (2020 - currently in revision and resubmission). A Case Study for Open Science in Natural Language. Organizational Behavior and Human Decision Processes.

Overall, we detected a pattern of specificity scores interacting with gender differently by question type – women received more specific feedback than men for the question about their strengths, and less specific feedback about their areas for development. Increased specificity in the strengths-based feedback for women is somewhat in line with the work by Jampol and Zayas indicating that reviewers give more positive feedback when they are told the subject is a woman.[24] Reduced specificity in the development-based feedback is also in line with work finding that women receive vaguer feedback on what they need to do to improve.[25] However, our treatment failed to reduce this gender-specificity interaction in both questions.

4.2 Gender biased language

We found that alterations to the instructions for the open text responses did not change the use of agentic and communal qualities in the reviews regardless of the review subjects' gender. Additionally, the exploratory analysis using machine learning to predict review subjects' gender based on the review text was no less accurate in the treatment reviews than in control reviews. This suggests that the treatment did not reduce the total amount of linguistic gender bias in the reviews.

Overall, we found that women were more likely than men to receive strengths-based feedback that reviewed both their communal and agentic qualities. As discussed, this may reflect changing stereotypes, whereby women (counter to traditional thinking) are perceived as possessing both communal and agentic traits.[26] This may be especially the case in the UK public sector which explicitly seeks to increase the diversity of its workforce at this level and supports balancing of work and care responsibilities at senior levels. Moreover, at the time this data was collected in 2018, the this particular workforce was 43.1% female and likely less male-dominated than other workplaces. Prior research has suggested that women working in more male-dominated work contexts (such as the US military or tech firms) are seen as deficient in agentic qualities.[27]

4.3 Exploratory analyses

Our exploratory analysis seeking to predict review subjects' gender based on feedback found that gender was indeed revealed within the text. Women were more likely to be described using object (rather than subject) pronouns, receive more subjective phrasing and receive strength-based feedback related to their 'team' more often, whereas men received more feedback about their 'knowledge' or 'strategy'. These findings suggest that there do appear to be subtle linguistic differences in reviews for women versus men, and that women appear to receive more tentative and relationship-oriented feedback, which was not picked up by the specificity algorithm or the communal / agentic dictionaries.

---

[24] Jampol, L., & Zayas, V. (2020). Gendered White Lies: Women Are Given Inflated Performance Feedback Compared With Men. *Personality and Social Psychology Bulletin*, 0146167220916622; McKinsey (2016). Women in the Workplace 2016. *Lean In and McKinsey & Company*. Retrieved February, 27, 2017.

[25] Correll, S., & Simard, C. (2016). Vague feedback is holding women back. *Harvard Business Review.*

[26] Griffiths, O., Roberts, L., & Price, J. (2019). Desirable leadership attributes are preferentially associated with women: A quantitative study of gender and leadership roles in the Australian workforce. *Australian Journal of Management*, 44(1), 32-49; Berkery, E., Morley, M., & Tiernan, S. (2013) Beyond gender role stereotypes and requisite managerial characteristics: From communal to androgynous, the changing views of women. *Gender in Management: An International Journal* 28: 278–298.

[27] 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(3-4), 159-171.

Our finding that women did not in general receive less specific reviews compared to men and that reviewers were only less willing to be specific in relation to their areas for development is consistent with the literature on benevolent sexism, which suggests that women are given less constructive feedback in an attempt to 'protect' them from negative feelings upon receiving critical feedback - ultimately holding them back. To explore this we undertook additional exploratory analysis that was not pre-registered, and found that overall women got more positive emotion words than men in their feedback for all questions. This was the case even when they were judged to be performing equally well based on the numeric ratings collected in the feedback survey. Counterintuitively, this pattern held even in the development-based question where respondents made less definitive statements and used hedges more to soften their claims when providing feedback to women.

Finally, in another exploratory analysis we found that asking people to assess others on being ‘visible’ versus ‘accessible’ made a significant difference in the scores given. In our study, the control group were asked to rate if someone was 'visible and approachable' and the treatment group were asked to rate whether someone was 'accessible and approachable'. Women got higher scores than men in both conditions but the treatment condition produced a) higher scores for everyone for this question compared to the control and b) a score that better reflected the overall performance rating for the review subject. This suggests that ‘accessibility’ was a closer match to the ‘true’ performance quality than ‘visibility’. We think that asking about 'visibility' could start triggering irrelevant aspects of someone's interactions at work (e.g. are they available full time or during anti-social hours). We also question the validity of ‘input’ related questions, be it focused on visibility or accessibility, as they could introduce potential bias against remote and flexible workers. Instead we would advise focusing on work outputs rather than inputs (i.e. hours) in performance reviews.

4.4 Limitations

There were a number of limitations to this research.

The dataset was anonymised very carefully, which created limitations: the dataset did not include review subjects’ grade, department, ethnicity, age, job role, and other characteristics that may be important influences on the language used. It also did not include the numeric performance ratings that reviewers gave to review subjects.

In addition, we were unable to investigate how the language in feedback may relate to progression outcomes, as this outcome data was not available. Similarly, we were not able to ascertain how the language used may be linked to perceived usefulness of the feedback from the perspective of the review subjects. For example, the treatment may have increased the usefulness of the feedback by making it more actionable, but this may not be measured by the specificity algorithm.

Third, because different questions were re-written in different ways, the treatment design makes question-by-question analysis complicated. Overall, the changes to the treated reviews seem to have increased the expected effort of the writing task, which reduced reviewers’ response rates.

Finally, the UK public sector is likely to be more progressive in terms of gender equality than other UK employers and so we are unable to generalise these findings to all other workplace contexts.

Conclusions and future research directions

5. Conclusion and future research directions

In the context of this large public sector employer, our findings indicate that women receive more specific feedback about strengths and less specific feedback than men on areas to develop, which is in line with existing research.

The fact that women are more likely to receive strengths-based feedback about both communal and agentic attributes may reflect changing stereotypes about women in leadership positions, particularly in this relatively progressive workplace.

However, our preliminary evidence that women are more likely than men to receive tentative (less direct) and relationship-oriented feedback warrants further investigation. Indeed, this study has found evidence of benevolent sexism in women's reviews, which manifests in more positive and less direct feedback given to women compared to men. Despite its good intentions this in fact may hold women back in their careers as they have less actionable feedback to draw upon. This result suggests different interventions may be more effective at closing the gender gap in performance reviews. For example, interventions directed at the respondents' directness, or honesty, may bring feedback for women more in line with the feedback that men receive.

The intervention we tested did not systematically increase specificity of the feedback. This may reflect the fact that changes to the instructions in the treatment condition were attempting to fulfil a number of aims: notably, increase actionability and also reduce gender bias by encouraging feedback on both communal and agentic traits. In asking for a defined number of examples, the control condition may also have provided a similar degree of encouragement to be specific as the treatment condition. Future research should build on this knowledge to further test different ways of structuring performance feedback forms in order to get reviewers to give more actionable feedback. One such avenue of enquiry could look into whether asking people to provide future oriented advice rather than feedback (which tends to be focused on past performance) leads to more specific and actionable suggestions.

Finally, future research is needed to demonstrate how language used in feedback may be linked to the perceived usefulness of the feedback from the review subjects' perspective, as well as how it may be linked to career progression outcomes.

Tentatively, this study also found a notable increase in the performance scores both men and women received (with women higher in both arms) when people were asked to review whether they were 'visible and approachable' rather than 'accessible and approachable at all times'. We suggest that suppliers of 360 degree software and employers using such software should consider whether they need to ask about 'inputs' (in terms of hours in the office or in direct contact with a team) rather than 'outputs' (in terms of individual and team delivery against objectives). There is a risk that asking about inputs disadvantages remote or part-time workers in performance reviews and instead rewards a long-hours culture, which is less inclusive for all.

Appendices

Appendices

Appendix A - Control and treatment open text question instructions

Figure 4 - Control questions

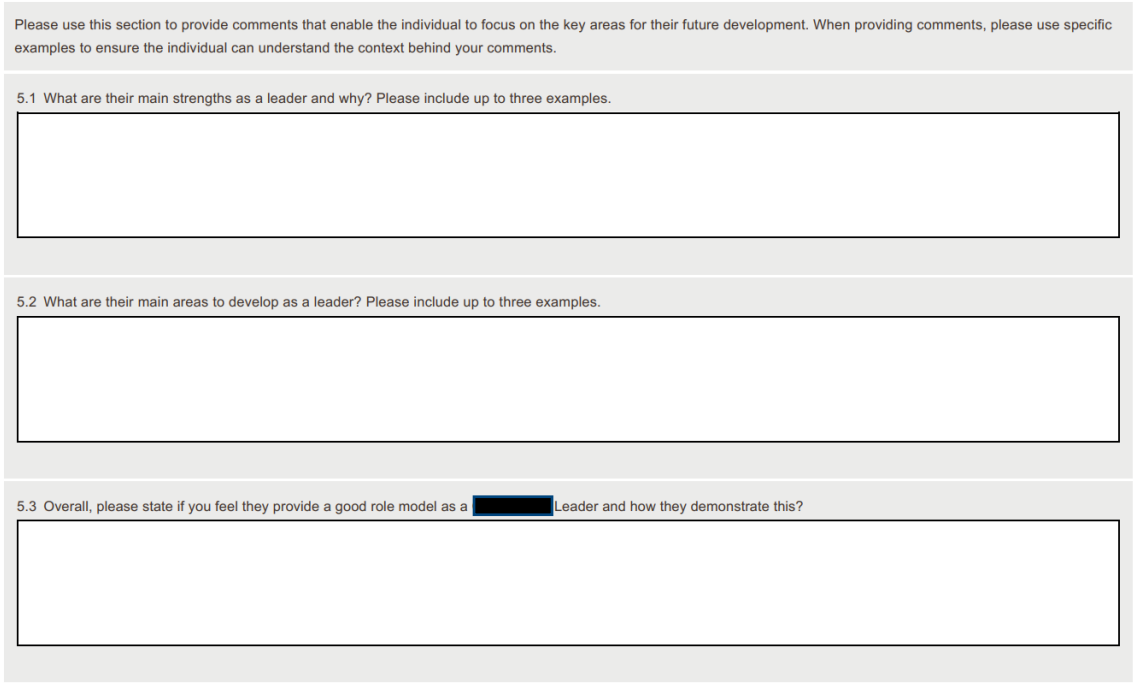
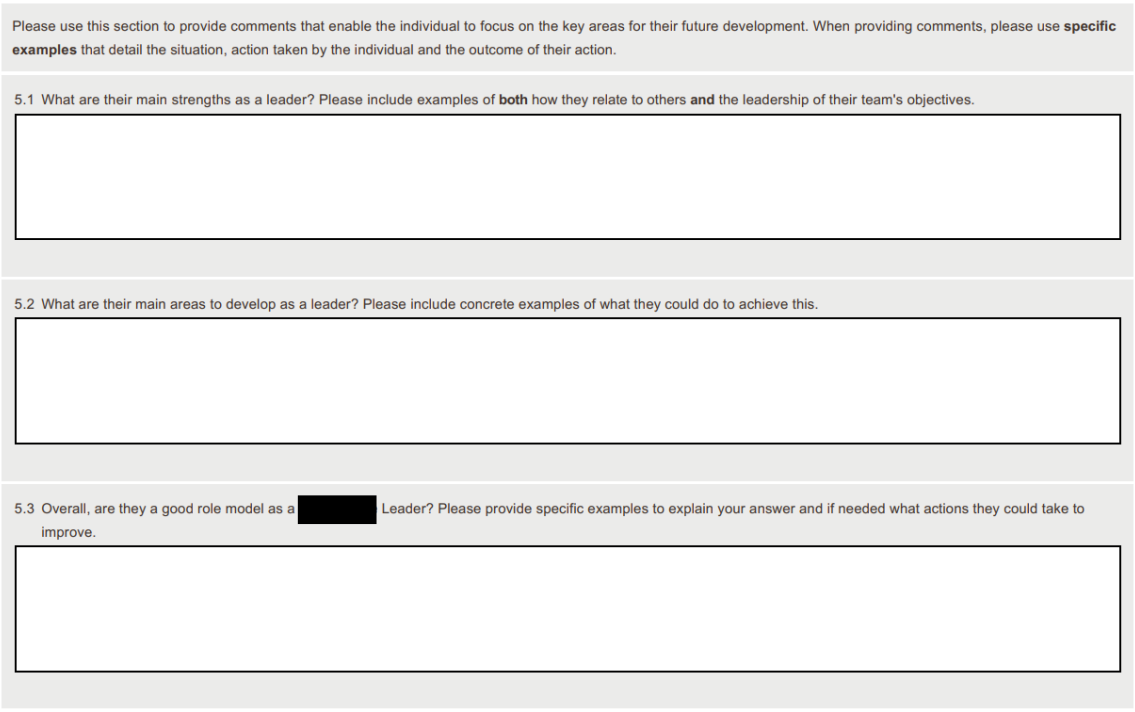


Figure 5 - Treatment questions



Appendix B - How doc2concrete works

Doc2concrete is an analysis function that uses three types of data to create its measure of concreteness of a piece of advice. First, it takes six real-world datasets where various types of participants (e.g. colleagues, teachers, peers) gave others advice. In five of those datasets, human participants annotated (scored) the datasets for their concreteness (i.e. the extent to which writers provide concrete, specific details to recipients). In the sixth, concreteness was inferred from the treatment condition.[28] The table below provides details of the datasets.

Table 6 - Doc2concrete algorithm training data sets

Dataset	*Sample size*	*Reference*
Workplace feedback - Similar to the 360 feedback that was the subject of this analysis	1334	Blunden, H., Green, P., & Gino, F. (2018). The Impersonal Touch: Improving Feedback- Giving with Interpersonal Distance. Academy of Management Proceedings, 2018(1).
Teacher feedback - Feedback from teachers to parents about their children	304	Rogers & Kraft, 2015
Personal feedback - Feedback to a friend about any 'recent task' the friend performed	171	Blunden, H., Green, P., & Gino, F. (2018). The Impersonal Touch: Improving Feedback- Giving with Interpersonal Distance. Academy of Management Proceedings, 2018(1).
Task tips - Advice on games (e.g. darts, boggle)	228	Levari, D.E., Wilson, T.D, & Gilbert, D.T. (2019) Advice from top performers feels (but is not) more helpful. Working Paper.
Letter advice - Advice improving a cover letter	951	Yoon, J., Blunden, H., Kristal, A. & Whillans, A. (2019). Seeking Constructive Feedback? Ask for Advice Instead. Working Paper.
Life goals - Advice on how to live a happy life	301	Zhang, T., North, M. (2019). Wunderkind wisdom: Younger advisers discount their impact. Working Paper.

Doc2concrete uses the association between n-grams (phrases that include from one to 'n' words) in a piece of feedback and the human annotators' concreteness score for that

[28] In the dataset where concreteness was inferred from the treatment condition, teachers were randomly allocated to give either 'positive feedback' or feedback related to 'improvement'. Afterwards, a research assistant blind to the treatment condition confirmed that the improvement feedback was more actionable than the positive feedback (89% vs. 8%). The creators of doc2concrete used treatment assignment as a measure of concreteness, e.g. assuming that reviews about improvement were concrete and that positive feedback reviews were not concrete.

piece of feedback. It also uses two scores that come from ascribing concreteness scores to individual words based on predefined dictionaries – one by Brysbaert, Warriner, and Kuperman[29], and another by Paetzold and Specia[30]. In other words, doc2concrete uses 1) the n-grams in a piece of feedback, 2) the feedback's concreteness score based on the words in it according to the Brysbaert dictionary, and 3) the Paetzold dictionary concreteness score. It then produces a summary score based on these three methods.

In the development of the doc2concrete function, its creators tested its 'out-of-sample' performance by rebuilding it six times, each time leaving out one of the six datasets of feedback. They found that the function performed best in classifying concreteness in the teacher feedback dataset and the workplace feedback datasets. This gives us confidence that doc2concrete provides meaningful outputs in examining this 360 feedback dataset, despite this feedback being a new context for doc2concrete.

[29] Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46(3),* 904-911.
[30] Paetzold, G., & Specia, L. (2016). Inferring psycholinguistic properties of words. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 435-440).

Appendix C - Agentic and communal dictionaries

Table 7 - Agentic and communal dictionaries

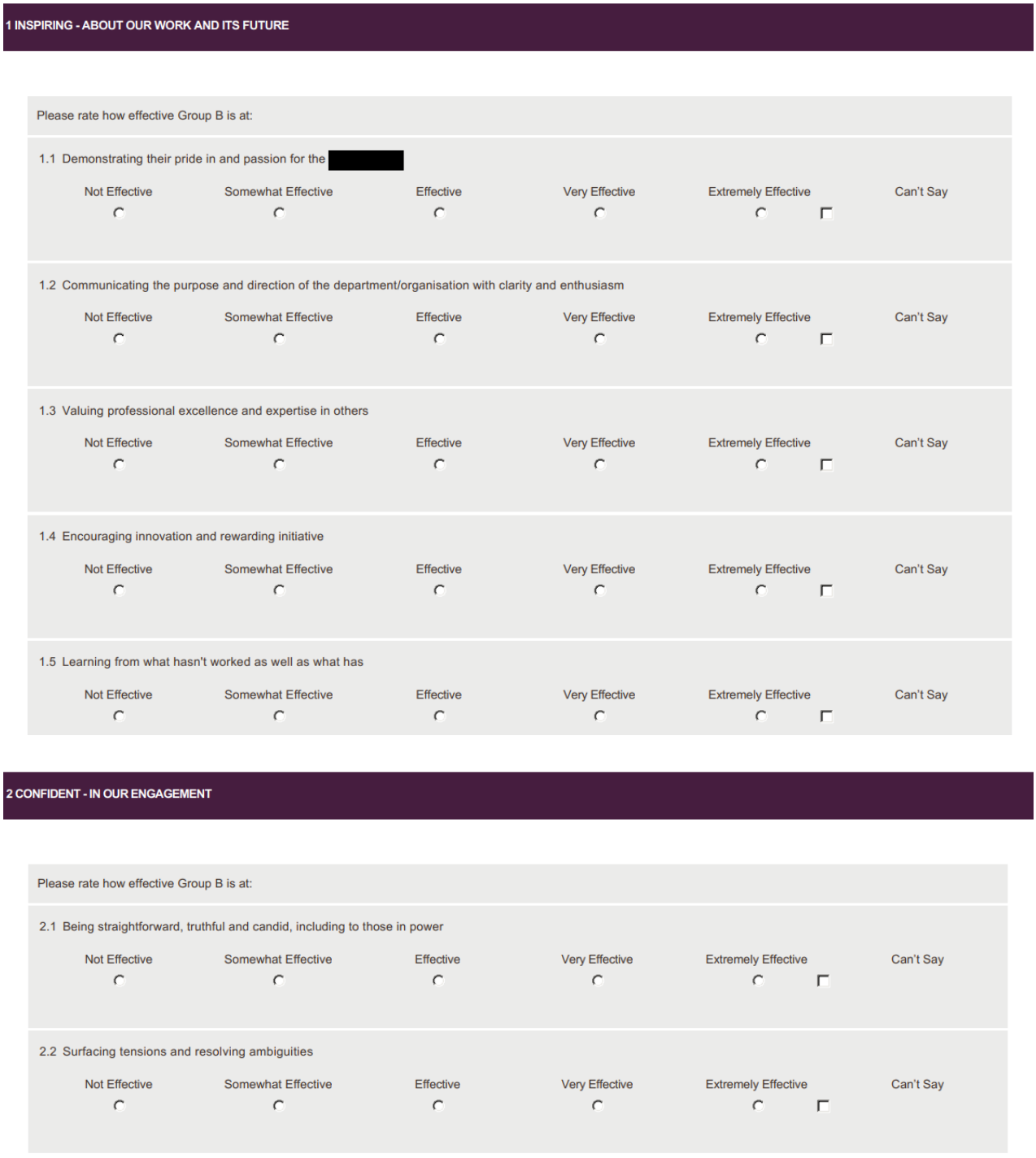
Male-affiliated ('agentic') words	*Female-affiliated ('communal') words*
Active	Affectionate
Adventurous	Child*
Agress*	Cheer*
Ambitio*	Commit*
Analy*	Communal
Assert*	Compassion*
Athlet*	Connect*
Autonom*	Considerate
Boast*	Cooperat*
Challeng*	Depend*
Compet*	Emotiona*
Confident	Empath
Courag*	Feminine
Decide	Flatterable
Decisive	Gentle
Decision	Honest
Determin*	Interpersonal
Dominat*	Interdependen*
Force*	Interpersona*
Greedy	Kind
Headstrong	Kinship
Hierarch*	Loyal*
Hostil*	Modesty
Impulsive	Nag
Independen*	Nurtur*
Individual*	Pleasant*
Lead*	Polite
Logic	Quiet*
Masculine	Respon*
Objective	Sensitiv*
Opinion	Submissive
Outspoken	Support*
Persist	Sympath*
Principle*	Tender*
Reckless	Together*
Stubborn	Trust*
Superior	Understand*
Self-confiden*	Warm*
Self-sufficien*	Whin*
Self-relian*	Yield*

Note: The asterisk (*) denotes the acceptance of all letters, hyphens, or numbers following its appearance.

Appendix D - Gender differences in numeric ratings

Numeric rating scales were the same across both conditions with one exception: question 3.2 was worded as “Being visible and approachable at all times” in the control condition, but changed to “Being accessible and approachable” in the treatment condition.

Figure 6 - Numeric scales



2.3 Giving clear and honest feedback to help staff succeed

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

2.4 Addressing performance concerns resolutely, fairly and promptly

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

2.5 Being a team player and working collaboratively

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

2.6 Encouraging others to take managed risks and learn from their mistakes

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

2.7 Building credibility and influence to strengthen relationships with [redacted] external stakeholders

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

2.8 Developing and maintaining positive, productive relationships with junior staff

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

3 EMPOWERING - OUR TEAMS TO DELIVER

Please rate how effective Group B is at:

3.1 Giving others the space and authority to deliver

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

3.2 Being accessible and approachable

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

3.3 Demonstrating receptiveness to being challenged, however uncomfortable

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

3.4 Championing difference, recognising the value it brings

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

3.5 Demonstrating commercial awareness in decision making

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

3.6 Embedding flexible and responsive ways of working including digital where possible

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

3.7 Investing time in and showing commitment to their own development

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

3.8 Developing talent and investing in the development of others to be effective now and in the future

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

4 DELIVERY LEADERSHIP

Please rate how effective Group B is at:

4.1 Taking responsibility for delivery of [redacted] priorities

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

4.2 Developing strategies that focus on delivering results that have a long term impact

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

4.3 Showing resilience in making tough strategic decisions

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

4.4 Prioritising a consistently high quality customer outcome

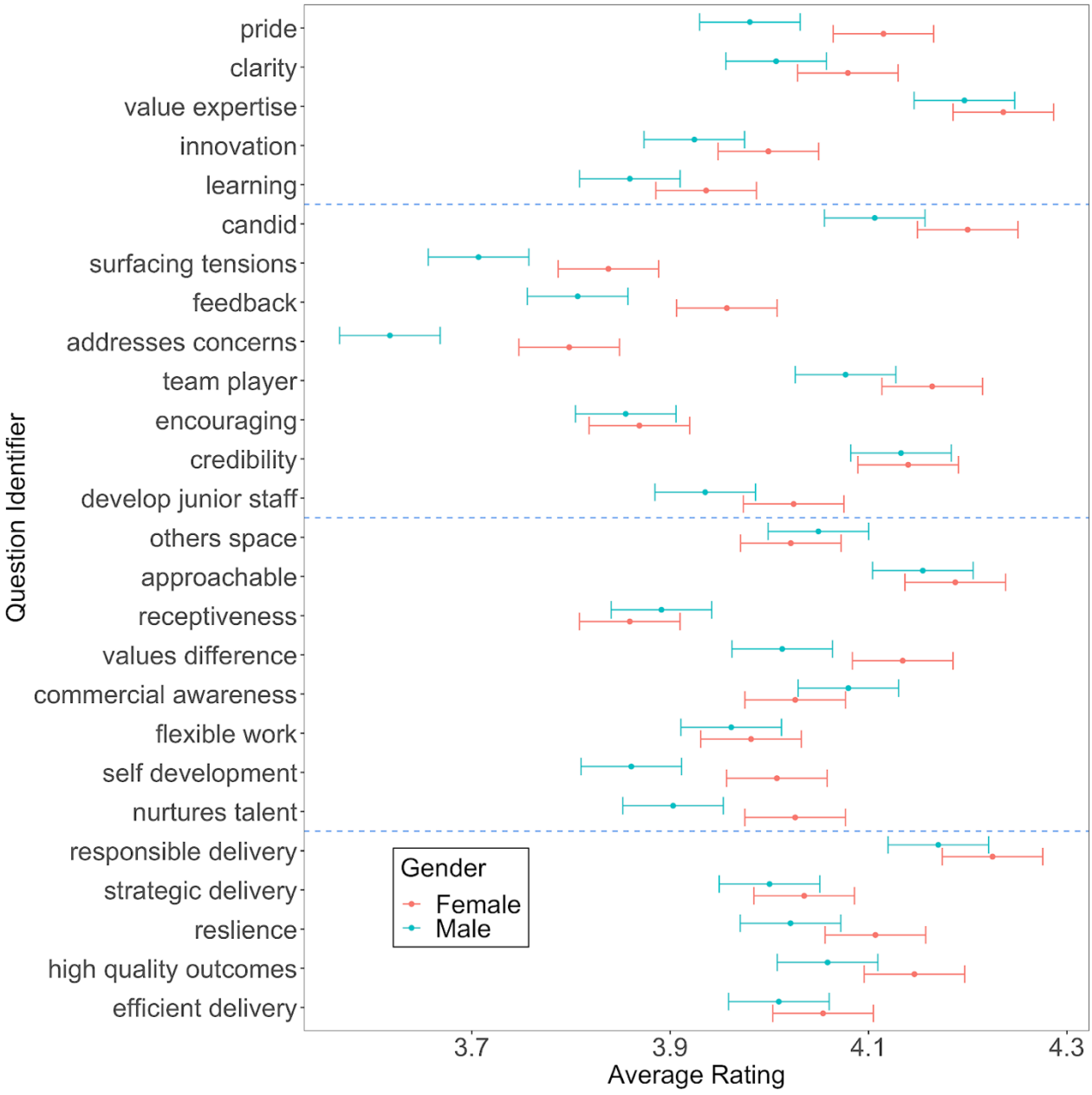
Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

4.5 Maintaining real focus on delivering efficiency and value for money

Not Effective	Somewhat Effective	Effective	Very Effective	Extremely Effective	Can't Say
○	○	○	○	○	☐

The graph below shows the average ratings for men and women on each of the numeric questions. In general, ratings across questions tended to be correlated, and women tended to be rated more positively than men.

Figure 7 - Average numeric ratings by gender



Appendix E - Regression estimates of benevolent sexism

All models use the gender of the subject as the outcome variable in a linear probability model, and cluster standard errors at the subject and respondent level. Performance ratings are the first principal component extracted from the numeric ratings in each feedback response; all other measures are extracted from responses to the open-ended questions.

Table 8 - Benevolent sexism regression estimates

'Strengths' question	***Model 1***	***Model 2***		
Positive Emotion	-0.240*** (0.055)	-0.196*** (0.054)		
Performance Ratings	-	0.503*** (0.028)		
df	21,789	21,788		
'Development' question	***Model 3***	***Model 4***	***Model 5***	***Model 6***
Positive Emotion	-0.103** (0.039)	-0.120** (0.040)	-	-
Performance Ratings	-	-0.113*** (0.020)	-	-0.001 (0.010)
Hedges	-	-	-0.044* (0.020)	-0.045* (0.020)
df	21,742	21,741	21,742	21,741
'Overall' question	***Model 7***	***Model 8***		
Positive Emotion	-0.183*** (0.039)	-0.156*** (0.038)		
Performance Ratings	-	0.286*** (0.020)		
df	21,701	21,700		

Appendix F - Further details on the main results

Table 9 - Specificity regression estimates

Outcome	*Question type*		
	Strengths	*Development*	*Overall*
Difference in the number of words used, by treatment	The treatment had no impact on the number of words used in the Strengths review text (β = -.125, SE = .988, t(30572) = -0.127, p = .899).	The treatment increased the number of words used in the Development review text (β = 3.90, SE = .927, t(27271) = 4.20, p < .001).	The treatment had no impact on the number of words used in the overall review text (β = -.289, SE = .664, t(28934) = -.435, p = .663).
Specificity as measured by standardised Doc2concrete score	See Figure 2 for means by treatment for each question, by gender.		
Gender difference in specificity as measured by Doc2concrete, control condition	Female subjects received more specific Strengths review text (β = .021 , SE = .007, t(23077) = 3.0, p = .002).	Female subjects received less specific Development review text (β = -.019 , SE = .008, t(23047) = 2.4, p = .016).	Female subjects received no less or more specific Overall review text (β = .008 , SE = .005, t(22961) = 1.5, p = .146).
Gender difference in specificity as measured by Doc2concrete, treatment condition	The treatment decreased the specificity of the Strengths review text (β = -.036, SE = .007, t(23097) = 4.9, p < .001).	The treatment condition had no effect on the specificity of the Development review text (β = -.006, SE = .008, t(23047) = 0.7, p = .488).	The treatment caused reviewers to write more specific overall review text (β = .116, SE = .021, t(22961) = 5.4, p < .001).
Treatment * gender interaction	We find no statistically significant treatment * gender interaction on any of the three questions, for specificity and for word count.		

Table 10 - Gender biased language regression estimates

Outcome	***Question type***		
	Strengths	***Development***	***Overall***
Gender differences in the number of communal words	Female subjects received more communal words in their Strengths review text than men (β = .065 words, SE = .020, t(23077) = 3.2, p = .001).	Gender differences in the number of communal words were not significant in the Development or Overall questions.	
Gender differences in the number of agentic words	Female subjects received more agentic words in their Strengths review text (β = .105 words, SE = .026, t(23077) = 4.0, p < .001).	Gender differences in the number of communal words were not significant in the Development or Overall questions.	
Differences in the number of agentic or communal words, by treatment	We found no treatment effects, or interactions between treatment and gender, for either dictionary (agentic or communal words) across any of the three questions.		

© Crown copyright 2019

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit nationalarchives.gov.uk/doc/open-government-licence/version/3

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned.