

Does Gender Matter for Leaders' Behavior and Effectiveness? Insights from A Field Experiment

Simone Haeckl and Yuko Onozaka*

Abstract

This study examines gender differences in leadership behavior and effectiveness using a framed field experiment conducted in a large company. Leaders and followers in randomly assigned teams interacted in recorded online team meetings to discuss topics of strategic importance. Their behaviors were assessed through research assistant ratings and natural language processing, and effectiveness via external evaluations and follower surveys. We find that female leaders exhibited significantly more communal behaviors through elaboration on team members' ideas, frequent discussion contributions, and affirmative language than male leaders. However, these differences did not translate into superior team performance; male and female leaders showed comparable effectiveness, particularly in external evaluations. Follower evaluations were more responsive to leader gender, with evidence of a "communality bonus" whereby male leaders received disproportionately positive evaluations for communal behaviors. Higher-level leaders achieved better team performance, regardless of gender. These findings suggest that leadership effectiveness is more strongly associated with developed expertise than with gender per se. Organizations may thus benefit from broadly developing leadership capabilities, alongside implementing evaluation systems that mitigate gender biases.

Keywords: leadership, behavior, gender, RCT, effectiveness

*University of Stavanger, School of Business and Law, Department of Economics and Finance, Kjell Arholms gate 41, 4021 Stavanger, Norway;

simone.haeckl-schermer@uis.no; yuko.onozaka@uis.no (corresponding author);

We gratefully acknowledge the generous partnership with the corporation that provided field access, in addition to the financial support from the Research Council of Norway (325398) and the University of Stavanger. We thank the participants of the 3rd Leadership Conference at NEOMA Business School and the labor economics research group at the University of Stavanger School of Business and Law for their helpful comments. In addition, we thank several research assistants for their work on the video data and transcripts.

1 Introduction

Does gender matter for leaders' behavior and effectiveness? Understanding the impact of gender on leadership has critical implications for organizational design, performance, talent development, and efforts to address persistent workplace gender disparities. Still, this long-debated question remains unresolved. Traditionally, female leadership has been viewed as a disadvantage, rooted in the incongruence between communal gender norms and the agentic traits typically associated with leadership (Eagly and Karau, 2002). Social role theory (Eagly et al., 2000) postulates that people internalize or conform to gendered expectations—men to agency (assertiveness, competitiveness) and women to communion (nurturance, cooperativeness). Since leadership has historically aligned with masculine-coded, agentic traits, women often face a double-bind: they are perceived as violating prescriptive norms—whether they behave “too agentic” for a woman or “too communal” for a leader (Rudman and Glick, 2001; Rudman and Phelan, 2008). These dynamics not only impact how women behave as leaders but also how they are evaluated, frequently receiving lower ratings despite demonstrating comparable or superior performance (Heilman et al., 1995).

However, a shift into a globally-knit, knowledge-focused economy with diverse workforces has transformed the qualifications, practices, and culture surrounding leadership in modern organizations (Eagly et al., 2003). As the contemporary workplace increasingly emphasizes teamwork, a leader's ability to leverage each team member's talent and generate innovative ideas through collaborative interactions has become crucial. Consequently, desired leadership qualities now emphasize relational skills, coordination, coaching behavior, and inspirational qualities (Antonakis et al., 2010). As these leadership skills align well with the communal domain traditionally associated with women, the potential for a “female leadership advantage” has emerged (Eagly and Carli, 2003a,b; Vecchio, 2002, 2003). Meta-analyses find small but consistent patterns indicating that female leaders engage in more relation-oriented and transformational leadership styles (Eagly et al., 2003) and receive better evaluations (Paustian-Underdahl et al., 2024).

Women potentially seen as more effective leaders can also stem from the notion that women who achieve leadership status despite various hurdles must be exceptionally competent, as depicted in the theory of double standards of competence (Biernat and Kobrynowicz, 1997; Foschi, 2000). According to the theory of double standards of competence, women develop and adopt particularly effective leadership approaches—irrespective of implied agency or communion of such approaches—to overcome systemic undervaluation and demonstrate competence. Rosette and Tost (2010) indeed find that top female leaders are perceived as more agentic, communal, and effective than top male leaders.

Despite the extant literature, empirical evidence of *actual* behavioral differences between male and female leaders and how such differences manifest to influence leadership effectiveness is scarce (Buss et al., 2024). One reason is that leadership research continues to conflate evaluation with behavior, making it difficult to discern whether some leaders are less effective, perceived as less effective, or simply held to different standards than other leaders (Antonakis et al., 2010; Fischer et al., 2023; Hemshorn de Sanchez et al., 2022). As Banks et al. (2023) show in a comprehensive review, only 3% of leadership studies focus on actual observed behaviors. In contrast, most studies rely on evaluations by subordinates or third parties as proxies for behavior. As scholars have noted, this leads to theoretical and conceptual misalignment, where perceptions are treated as equivalent to behavior, and constructs such as leader effectiveness, leader style, and team performance are confounded with subjective judgments (Antonakis, 2017; Fischer et al., 2023; Stock et al., 2023). This confoundedness is particularly problematic in gender research, where evaluations are subject to bias held by perceivers (Eagly et al., 1995; Heilman et al., 1995). Yet, a recent meta-analysis indicates that only *one* study, out of 180 sources, employed an objective, rather than a subjective, measure in assessing the gender difference in the effectiveness of leaders (Paustian-Underdahl et al., 2024).

Another reason is that many studies rely largely on correlational evidence from study designs that fail to eliminate confounding effects or various types of selection bias (Antonakis et al., 2010). For instance, male and female leaders may select specific types of followers or vice versa; thus, the effects of the leader’s gender and the member selection cannot be separated. Experimental approaches offer a promising solution. A growing body of lab and artificial experiments has studied gender differences in follower behavior using coordination games (Grossman et al., 2019; Heursen et al., 2023; Reuben and Timko, 2018; Timko, 2017), public good games (Gangadharan et al., 2016, 2019) or rank allocation tasks (Chakraborty and Serra, 2023). While most of this literature does not find gender differences in leaders’ behavior, there is ample evidence that followers react differently to male and female leaders (Chakraborty and Serra, 2023; Grossman et al., 2019). As these lab experiments have limited external validity due to abstract settings and student samples, recent research has moved to framed field experiments. These studies have provided insights into gender stereotypes in follower responses—for example, De Paola et al. (2022) find female leaders improve team performance in college exam settings, but are still evaluated more negatively by male followers. Similarly, Macchiavello et al. (2020) show that female supervisors underperform initially due to bias but achieve parity over time.

This paper aims to address and mitigate the aforementioned challenges and provide

insight into whether and how gender matters in leadership effectiveness.¹ We utilize data from a randomized controlled trial (RCT) where leaders conducted and recorded a meeting with randomly matched employees from the same company.² The team task was designed as follows. During the meeting, the teams discuss a question of strategic importance to the company—specifically, how their company could become a better workplace. Teams selected the three best ideas from their discussion and summarized them in presentation slides. These summaries were then sent to and evaluated by the human resource department (HR) for originality and relevance.

The meeting recordings provide direct observations of leaders’ and followers’ behavior, from which we generate behavioral measures through independent raters and by employing natural language processing (NLP). The random assignment of leaders to randomly assembled teams eliminates the selection effects. Furthermore, we measure leadership effectiveness using a combination of subjective evaluations from followers, evaluations from independent raters, and ratings of output quality from experts who are blind to the leader’s gender. As the study is conducted in a real company with actual leaders, we can exploit experimental (exogenous) variation while avoiding potential bias from assessment studies (where leadership effectiveness is evaluated for individuals not selected as leaders) and still maintain the essence of organizational studies.³

Based on this rigorous methodological approach, we find some evidence that female leaders behave more communally to ensure engagement from followers (elaborate on others’ ideas, speak more frequently, and use more assent words) to a greater extent than male leaders. However, these behavioral differences do not translate into higher team performance. We find that male and female leaders demonstrate comparable effectiveness, especially when performance is evaluated externally. Followers’ evaluations are more perceptive to leaders’ gender, as we also find some evidence of a communality bonus for male leaders, where male leaders are extra rewarded for behaving communally. Furthermore, teams led by higher-level leaders, regardless of their gender, produce significantly better ideas, indicating that effective

¹We infer leaders’ gender from the personnel records. As this is how they are registered with their employer, we assume it is in line with how they view themselves.

²The purpose of the RCT was to evaluate the effect of a randomly assigned leadership training (see [Haeckl and Rege, 2025](#)). The analysis of the training was pre-registered in the AEA registry (AEARCTR-0007129) and the project received ethical approval from the Norwegian Agency for Shared Services in Education and Research. The random assignment of leaders to followers was not part of the original evaluation or pre-registration. The results should, therefore, be considered as exploratory. To mitigate the lack of pre-registration, we report all the variables we considered theoretically important in the estimation and correct for multiple hypothesis testing to account for the increased risk of false positives.

³These three types of studies, organizational, experimental, and assessment, have been found to systematically produce different results, though the direction of effects is not uniform across time ([Eagly and Johnson, 1990](#); [Paustian-Underdahl et al., 2014](#)).

leadership more strongly correlates with developed expertise than with gender-based traits.

This study contributes to the literature by providing rigorous empirical evidence on the relationship among leader gender, behavior, and effectiveness. The remainder of this paper unfolds as follows: First, we provide the context of the study, followed by our conceptualization of how and when gender plays a role in the team task processes. Then, we describe how we measure the key concepts. Finally, we present the estimation results, followed by the discussion and conclusion.

2 Context

In this research, we use data from an RCT conducted in 2021 at an industrial and multi-utility company in Norway (Haeckl and Rege, 2025). The company’s nearly 2000 employees were invited to participate in online group meetings to discuss a question of strategic importance to the company, namely what their employer could do to become an even better employer.⁴ Meetings were held on Microsoft Teams, and a randomly assigned leader led each randomly formed team of one to three employees (followers). The workforce was highly diverse in job functions and educational backgrounds. Employees in infrastructure roles were more likely to engage in fieldwork, while those in telecommunications or sales tended to hold office-based positions. Correspondingly, skill levels varied widely—from employees with high school education and on-the-job training to engineers with advanced university degrees. Thus, the random assignment generated teams with followers from various educational and professional backgrounds, as well as the random assignment of leader gender to teams.

Each meeting lasted one hour and followed a structured format as outlined in Figure 1 (meeting instructions are available in Appendix C). As team members did not know one another, the leaders started the meeting with a short round of introductions and an icebreaking task (Guilford, 1967). Then, the first step of the main task began, where teams were asked to spend 15 minutes finding as many ideas for what the company could do to become an even better employer. In the second step, teams were asked to spend the last 15 minutes of the meeting selecting and elaborating on the three ideas that they found most promising as a group. After the meeting was over, the leader sent a PowerPoint presentation including the team’s output and a team identifier to the researchers. To ensure unbiased evaluation, the researchers then compiled a document listing all ideas and sent it to the company’s HR department to be rated based on originality and relevance. In addition, followers and leaders

⁴At that time, most people were working from home due to the COVID-19 crisis, so online meetings like the one conducted in the experiment were the norm. The meetings were recorded with the participants’ consent.

answered a short survey one week prior to the meetings and right after the meeting.

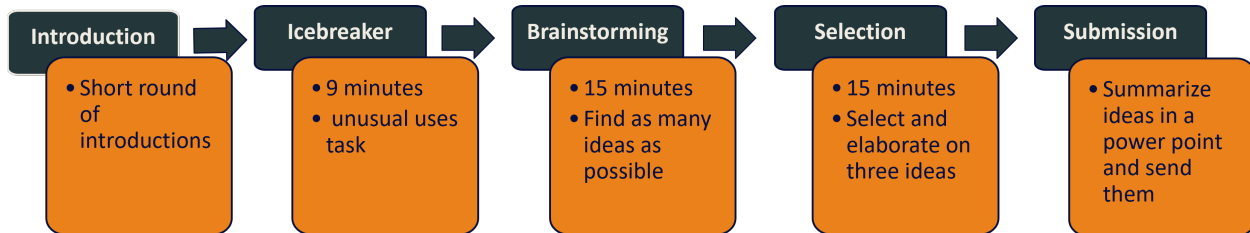


Figure 1: Meeting structure

This setting reflects key features of the modern workplace, where employees increasingly collaborate in multidisciplinary teams—often with unfamiliar colleagues. A recent survey reveals that as much as 71% of companies report that their employees work in teams and 28% report that they work in several different management-led teams (Eurofound and Cedefop, 2020). In addition, online meetings have become the norm since the pandemic. Based on data from Eurostat, half of EU enterprises report that they held online meetings in 2023 (Eurostat, 2023), and the share is even higher in Norway (77%), similar to other Scandinavian countries.

2.1 Sample

The recruitment was done through the HR department and the upper management. The HR department announced the research project on the company’s Microsoft Teams channels and sent out an invitation to participate by email. In addition, the company’s CEO posted a video message encouraging employees to participate. Approximately 70% of employees with personnel responsibilities (managers) and 30% of those without such responsibilities (non-managers) signed up. This resulted in 130 team meetings with a median group size of four (one leader and three followers). We obtained full meeting transcripts from 125 teams and survey responses from 302 followers.

Table 1 shows the sample characteristics of the leaders and the teams in the sample by leader gender. There are 37 female and 93 male leaders in our sample. Female leaders in our sample are, on average, 45 years old and have been working at the company for 11 years; the respective numbers for male leaders are 47 and 10 years and are not statistically significantly different. Seventeen percent of male and 27% of female leaders are higher-level leaders, i.e., have personnel responsibility for employees who have personnel responsibility themselves. This difference is not statistically significant. The average team size is 3.5 for both male and female leaders, and the average tenure of employees is eight years for male and ten years for female leaders. This difference is again not statistically significant. All in all, the male and

female leaders in our sample have very similar average characteristics. In addition, they are equally likely to have received leadership training in the RCT, as expected by randomization. We still control for receiving the leadership training in our empirical analysis, but do not report the coefficients in the main text.

Table 1: Team characteristics by leader gender

Variable	(1) Male leaders	(2) Female leaders	(3) Difference	(4) N
Leader tenure	9.892 (8.718)	11.054 (7.806)	1.162 (1.647)	130
Leader age	47.032 (9.359)	45.189 (7.612)	-1.843 (1.730)	130
Higher-level leader	0.172 (0.379)	0.270 (0.450)	0.098 (0.078)	130
Team size	3.559 (0.650)	3.514 (0.651)	-0.046 (0.126)	130
Average tenure followers	8.469 (5.652)	9.986 (6.570)	1.518 (1.152)	130
Leadership training	0.473 (0.502)	0.514 (0.507)	0.040 (0.098)	130
Observations	93	37	130	

Notes: The table shows the average by gender (columns (1) and (2)) and differences between male and female leaders (column (3)), with standard errors in parentheses. Leadership training indicates the proportion of leaders who went through the supportive leadership training as a randomized treatment in the RCT.

3 Conceptual framework and hypotheses

The theoretical underpinning as to why gender matters in leadership is attributed to two prominent theories: social role theory (SRT) and double standards of competence (DSC). According to SRT, the societal expectations and norms prescribed to each gender spill over to how men and women act as leaders (Eagly et al., 2000). More precisely, if leaders are socially conditioned to behave according to their prescribed gender norms, male leaders will behave more agentially (confident, competitive, assertive). In contrast, female leaders act more communally (kind, supportive, understanding). According to DSC, men may be evaluated more leniently than women because of the congruence of prescribed behavior for men and leaders (Biernat and Kobrynowicz, 1997; Foschi, 2000). Women in leadership positions face stricter evaluation standards and must demonstrate higher competence to receive equivalent recognition compared to their male counterparts. This evaluation bias may lead women to develop more effective leadership strategies as a compensatory mechanism while

paradoxically being seen as extra competent leaders who were able to overcome potential hurdles (Rosette and Tost, 2010). Thus, DSC predicts that female leaders employ the most effective strategies, irrespective of the nature of agency or communion associated with such strategies. Whether or not the two theories will predict the same or different behaviors for female leaders strongly depends on the setting in which the leaders are operating, as this defines which leadership behaviors are considered more effective.

There has been a tremendous interest in effective leadership behavior that enhances team outcomes within the literature, yet, surprisingly little is known about what effective leadership behavior is and how particular behavior influences team outcomes (Yukl, 2012). One reason is the sheer lack of studies investigating actual behavior (Banks et al., 2023). Another reason is that researchers tend to “lump up” behaviors into complex and aggregated constructs (e.g., transformational leadership), which often contain multiple dimensions. These aggregated constructs make it difficult to understand how particular behavior influences leaders’ effectiveness (Buss et al., 2024; Carton, 2022). In that regard, Yukl (2012) provides a helpful taxonomy of internal leadership behavior that is correlated to some measures of effectiveness and is sufficiently disaggregated. The taxonomy comprises three meta-categories: task-oriented, relation-oriented, and change-oriented behavior. In each meta-category, there are three to four components.⁵ Our approach is to use Yukl (2012) as inspiration and link leader behavior available in our specific setting to agency and communion.

Figure 2 outlines the team task processes and where we conceptualize that leader gender plays a role. Recall that, in the current study, the task is to come up with three ideas to make the company an even better workplace. After random assignment into teams, each team engages in cognitively demanding discussions to produce novel ideas (① and ② in Figure 2) and collaboratively evaluate each suggestion to reach a consensus on the best options and produce an output (③ in Figure 2). The team task involves a highly interdependent environment where team members are required to actively participate in generating, critiquing, and prioritizing ideas. Below, we discuss our conceptualization of how leaders’ gender can affect multiple stages within the process.

① Leader behavior

Leaders coordinate and facilitate collective effort for the success or survival of a group by exerting influence and providing guidance (Kaiser et al., 2008). While leaders are provided with instructions on the overall structure of the meeting, they retain the autonomy to choose which leadership practices would enable them to lead effectively. Given the highly structured

⁵In addition, there is also a category for external leadership, but this category is not relevant in our setting. We, therefore, focus on internal leadership.

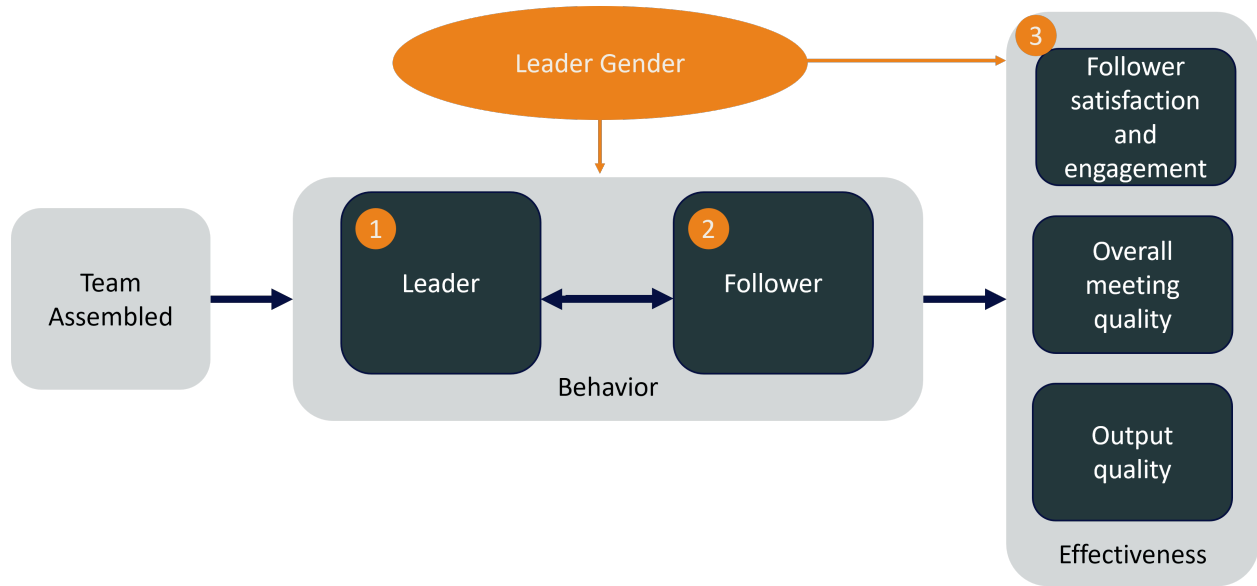


Figure 2: Output generating team processes

task setting and short time span, not all of the listed categories or components of internal leadership in Yukl (2012) are relevant in this case. To facilitate predictions about expected gender differences in leadership behavior, we select behaviors we believe are relevant and further classify them into communal and agentic dimensions. Communal leadership behavior reflects actions that are people-focused, relationship-enhancing, supportive, and friendly (Abele et al., 2008). Accordingly, communal leader behavior is consistent with relations-oriented (supporting, developing, recognizing, and empowering) and some of the change-oriented domains (encouraging innovation and facilitating collective learning) of effective leadership behavior taxonomy. Agentic behavior stems from agency, which refers to “a person’s striving to be independent, to control one’s environment, and to assert, protect and expand one’s self (Abele et al., 2008).” In the leadership context, this manifests as behaviors characterized by exerting influence, authority, and control over followers (Eagly and Karau, 2002).⁶

In the context of the current study, collaboration and innovation are essential, and the desirable leadership characteristics that foster such practices are relational skills, coordination, coaching behavior, and inspirational qualities (Antonakis et al., 2010). In addition, previous literature suggests that relation-oriented behaviors are more effective in facilitating a virtual team with high task interdependence (Brown et al., 2021). Accordingly, we hypoth-

⁶Agentic leadership may be associated with more task-oriented leadership (Paustian-Underdahl et al., 2024), which could be linked to behavior such as clarifying, planning, monitoring operations, and problem-solving in the effective leadership taxonomy. However, such behaviors are not very relevant in our structured and well-defined team tasks.

esize that female leaders display more communal behavior as it aligns with their prescribed gender norms and are expected to be more effective:

Hypothesis 1. Female leaders exhibit more communal leadership behavior than male leaders.

② Follower behavior

We further hypothesize that gender-specific leader behavior also affects the behavior of followers. More specifically, if female leaders are more communal, followers in a female-led team should have more active participation and engagement than those in male-led teams, e.g., by speaking more words, speaking more frequently, and contributing more ideas.

Hypothesis 2. Followers in female-led teams show more active participation than followers in male-led teams.

③ Leader effectiveness

When evaluating the effectiveness of leader behavior, leadership scholars typically care about two different types of outcomes: effects on the followers (e.g., engagement and motivation) and effects on productivity (Fischer et al., 2023). If female leaders act more communally by exerting more effort to facilitate active engagements by followers, we expect a higher evaluation by followers for their own engagement and satisfaction, as well as a higher overall meeting quality in female-led teams compared to male-led teams.

Hypothesis 3. Female leaders make followers feel more engaged and satisfied than male leaders.

Hypothesis 4. Meetings led by female leaders are rated as more engaging and overall better than meetings led by male leaders.

Given the collaborative and innovative nature of the task, it is possible that communal leadership behavior is more effective in stimulating ideas from the followers, leading to the production of better ideas. It also follows from the evidence based on virtual team performance that communal, relation-oriented leadership behavior is effective (Brown et al., 2021). Since the evaluators of ideas are blind to leader gender, we expect no systematic gender bias in this process.

Hypothesis 5. Female-led teams generate ideas that are rated higher than male-led teams.

4 Measures and empirical specifications

In this section, we discuss in detail how we measure the constructs introduced in the previous section. Our measures are also summarized in Table 2.⁷

Table 2: Overview of measures and their characteristics

Measure	Evaluator	Source	Concept
Rephrase others' ideas (C)	RA	Transcript	Leader behavior
Elaborate on others' ideas (C)	RA	Transcript	Leader behavior
Contribute new ideas (A)	RA	Transcript	Leader behavior
Supportive behavior score-external (C)	RA	Recording	Leader behavior
Supportive behavior score-follower (C)	Follower	Survey	Leader behavior
Leader no. contribution	NLP	Transcript	Leader behavior
Leader word count	NLP	Transcript	Leader behavior
Communion dictionary words (C)	NLP	Transcript	Leader behavior
Agency dictionary words (A)	NLP	Transcript	Leader behavior
High status words (A)	NLP	Transcript	Leader behavior
Authentic words (C)	NLP	Transcript	Leader behavior
Categorical-Dynamic Index (A)	NLP	Language	Leader behavior
Assent (C)	NLP	Transcript	Leader behavior
Rephrase others' ideas	RA	Transcript	Follower behavior
Elaborate on others' ideas	RA	Transcript	Follower behavior
Contribute new ideas	RA	Transcript	Follower behavior
Follower no. contribution	NLP	Transcript	Follower behavior
Follower word count	NLP	Language	Follower behavior
Follower evaluation: discussion quality	Follower	Survey	Effectiveness
Follower evaluation: own engagement	Follower	Survey	Effectiveness
Overall impression: leader supportiveness	RA	Recording	Effectiveness
Overall impression: leader engagement	RA	Recording	Effectiveness
Idea originality score	HR	Output	Effectiveness
Idea relevance score	HR	Output	Effectiveness

Notes: The table shows the overview of measures we employ (Measure), who evaluates them (Evaluator), the source of information (Source), and the intended concept we aim to measure (Concept). Evaluators are either research assistants (RAs), followers, human resource experts (HR), or natural language processing (NLP). Leaders' behavioral measures are classified as communal (C) or agentic (A). In addition, we consider the number of leaders' contributions and word count as a set. Precisely, a high number of contributions *and* a high word count as dominance (A), wherein a high number of contributions *without* high word count as engagement (C).

⁷We focus mainly on the transcripts rather than video recordings. People participated from home offices with varied audiovisual settings, resulting in a large variation in sound and video quality. In addition, Microsoft Teams only captures the faces of the participants (often from varied angles). These conditions prohibit a more thorough investigation of pitches, facial expressions, body language, and other non-verbal behaviors.

4.1 Leader behavior

Given the controlled setup we investigate, with pre-specified topics and meeting structures that are constant across teams, the scope of behaviors leaders can engage in is limited; for example, very little planning and monitoring behavior would be necessary in the short and pre-structured meetings and using incentives to influence followers would be difficult. To capture meaningful numerical information on leader behavior that allows us to gain insight into behavioral differences between male and female leaders, we identify concrete behaviors that might be effective in our context. For this purpose, we exploit the context of the RCT that focused on specific supportive leader behaviors to enhance follower engagement.

Subjective measures

We identify three target behaviors that leaders could engage in to foster engagement and create measures by making the numerical count each time a leader (1) rephrases others' ideas, (2) elaborates on others' ideas, and (3) comes up with a new idea themselves.⁸ For this measure, a team transcript was coded by one rater.⁹ Following the conceptualization, we consider rephrasing and elaborating on others' ideas to be communal behavior, as it shows that leaders are actively contributing to collaborative team interactions. Contributing new ideas can be neutral (neither communal nor agentic), but we interpret this behavior as agentic. The reason is our specific context, where meetings are conducted within a limited time frame. Thus, if leaders share many of their own ideas, this will necessarily take away opportunities for followers to share their ideas, making this behavior more in line with assertiveness, dominance, and autocracy.

Next, we consider measures of the overall impression of leaders' supportiveness during the meeting evaluated by followers in a survey immediately after the meeting is complete. Followers evaluated leaders on a Likert scale from 1 to 5 in responding to five questions: "To what extent did the leader engage in the following behaviors: 1) Encourage you to think outside the box? 2) Make you feel safe and secure? 3) Listen to the suggestions you were making? 4) Provide ideas oneself? 5) Make you feel comfortable participating in the discussion?"

In addition, four research assistants, who were blind to the leader's training status,

⁸As the study is situated in a rather specific one-hour online meeting in a company, we opted to employ this self-defined set of behaviors to capture relevant leadership behaviors that fit into our context. In this respect, our approach is different from interaction coding, which typically codes each line of the interaction using pre-existing coding.

⁹To estimate the inter-rater reliability, we asked a second rater to go through 20% of the meetings and repeat the task. The average agreement between raters is 94% and the agreement is significantly larger than expected by chance in all tested transcripts. The Cohen's Kappa varies between 0.48 and 1.

were asked to indicate whether the leader overall engaged in the following 10 behaviors: 1) Emphasize that each group member’s ideas are important because they all have their own experience and it is important to come up with a variety of ideas. 2) Acknowledge that some might feel nervous about contributing to the group discussion, but clarify that there is no need to be shy and that it is important that all of them contribute. 3) Emphasize that there are no stupid suggestions. 4) Look into the camera when others speak. 5) Write down their suggestions when they see it. 6) Ask clarification questions. 7) Emphasize that there is also room for more radical ideas. 8) Share one very radical and one more realistic idea yourself to show the group members that any type of contribution is welcome. 9) Allow group members to be critical but honor the potential of each idea. 10) Emphasize that it is critical that they all speak up in order to select and elaborate on the three best ideas. These behavioral categories were adopted from the study design in [Haeckl and Rege \(2025\)](#), whose main motivation was to understand the causal effect of supportive leader behavior. We count the number of behaviors the leader engaged in to generate the measure for supportive leader behavior.¹⁰

We consider the above measures as *subjective* as there is room for interpretation and bias from raters, although one might argue that the three behavioral counts are more objective than the supportiveness scores by measuring one specific behavior each.

Objective measures

In contrast to the subjective measures we discussed above, the following measures are based on word counts and existing dictionaries and do not depend on who analyzes the text. Thus, we consider these measures *objective*. We consider both simple word counts and measures rooted in psychometrics. How much a leader speaks (in frequency and word count) is a measure of engagement and influence. Next, we employ the communion and agency dictionaries developed in [Pietraszkiewicz et al. \(2019\)](#).¹¹ The original dictionaries in English were translated into Norwegian first by AI (Claude.ai) and then checked by a native Norwegian speaker for accuracy and suitability for our specific contexts. The communion dictionary comprises

¹⁰Prior to starting the evaluation tasks, the research assistants received a two-hour training. After evaluating each of the leaders individually, the research assistants were grouped into two teams of two, discussed their ratings, and collectively decided which rating to give the leader. We use this harmonized score in our analyses. In addition, we have information on how each research assistant individually rated the leaders. These ratings are highly correlated with $\rho=0.67$, $N=108$, and $p < 0.001$.

¹¹Dictionaries in NLP context are the predefined set of words or terms that are related to certain concepts or domains. For instance, sentiment analysis, a commonly used simple NLP applied to Media content, is conducted by counting the occurrence of target terms in the library related to sentiment, such as “happy” and “sad.”

187 root terms,¹² including words like *care*, *group*, *help*, and *trust*. These words are relational by construction. The agency dictionary comprises 194 root terms, including words such as *accomplish*, *goal*, *risk*, and *success*. These words are goal-oriented by construction. Note that communion/agency dictionaries are not inherently linked to gender but could be linked to task characteristics. Pietraszkiewicz et al. (2019) reports that male-dominated jobs are advertised using more agency words, while female-dominated jobs are advertised with more communion words.

The next set of language-based variables employs Linguistic Inquiries and Word Count (LIWC; Pennebaker et al. (2015)), where we investigate specific linguistic styles.¹³ LIWC is the leading tool to account for psychological constructs and processes in language (Tausczik and Pennebaker, 2010), and is widely used in psychometric language analysis in multiple disciplines.¹⁴ As our transcripts are in Norwegian, we employ a Norwegian translation of LIWC by Arnold Goksøyr, which is only available in the 2007 version of LIWC. We generate a set of language style summary variables inspired by Pennebaker et al. (2015): the language of power and authority (*status*), honesty (*authenticity*), and logical and hierarchical thinking (*Categorical-Dynamic Index, CDI*).¹⁵

First, we consider the language of power and authority. An individual’s sense of power affects one’s language. For instance, a person with high self-perceived power tends to use a more positive emotional tone and less tentative language (Körner et al., 2024). Kacewicz et al. (2014) find that those with higher social standing use more first-person plural pronouns (we-words), while those with lower social standing tend to use more first-person singular pronouns (i-words). Language that signals power may also be strategically employed as a form of impression management. For instance, activist investors are more likely to succeed in their campaigns when they use language rich in words associated with confidence and authority (Brauer et al., 2023). We link the language of power and status to agency. Accordingly, we construct the measure by adding we-words and social words and subtracting i-words, negations, and swear words (Brauer et al., 2023).

¹²For example, a root term *agree** includes all variations of the term, including *agree*, *agreed*, *agreement*, *agreeing*, etc.

¹³We refrain from using Large Language Model Artificial Intelligence, such as Chat-GPT, for the language-based analysis, due to the data confidentiality concern. LIWC was installed and run on a secure local server.

¹⁴LIWC is employed in more than 30,000 scientific studies, based on a Google Scholar search on “liwc” OR “linguistic inquiry and word count” accessed in December 2024.

¹⁵More recent versions of LIWC have composite summary measures of *clout*, *authenticity*, and *analytical thinking* that our measures of *status*, *authenticity*, and *logical/analytical thinking* correspond to. Unfortunately, these measures are not available in the 2007 version, for which the Norwegian translation is available. Instead, we construct these measures based on the original studies from which the current composite summary variables are derived. LIWC does not disclose the exact composition of these summary variables, which prevents us from checking the match between the original and the summary measures, but the correlation between the two is generally high (Brauer et al., 2023).

Next, we consider the language of authenticity, which is “associated with a more honest, personal, and disclosing text” (Pennebaker et al., 2015). In Newman et al. (2003), where the authenticity measure originates, researchers find that authentic communication requires more cognitive resources as speakers process and convey genuine experiences rather than constructed narratives. The validated linguistic markers shown in Newman et al. (2003) as positive indicators of authentic language are first-person singular pronouns, third-person pronouns, and exclusive words, while negative emotion words and motion verbs are negatively linked to authentic communication. We associate authentic language with communion, as this style is more humble and personable (Pennebaker et al., 2015).

Finally, we consider the language of analytical versus narrative thinking. Analytical thinking “reflects formal, logical, and hierarchical thinking,” whereas narrative thinking represents a more informal, story-based style (Pennebaker et al., 2015). To measure analytical thinking, we use the Categorical-Dynamic Index (CDI) based on Pennebaker et al. (2014). Analytical linguistic style has a formal and precise language, characterized by the frequent use of articles and prepositions. This is because articles and prepositions are necessary for referencing precise objects (e.g., data, facts, events, organizations, and statistics). In contrast, the narrative-based style requires time-based, person-focused words to tell the story. Thus, this style of language is characterized by the frequent use of pronouns, conjunctions, auxiliary verbs, and adverbs. CDI is constructed by contrasting these two distinct styles by summing prepositions, articles, and pronouns, then subtracting pronouns, auxiliary verbs, conjunctions, adverbs, and negations. A high value in CDI indicates the dominance of categorical language, and a low value indicates that dynamic language dominates. We associate high analytical thinking with task-orientation and agency and low analytical thinking with people-orientation and communion. A study by Pitt (2021) finds that female CMOs (Chief Marketing Officers) use less authoritative and more authentic words than male CMOs, but no significant differences in analytic thinking—demonstrating that there could be systematic gender differences in language styles between male and female leaders.

In addition, we consider the language of *assent* (e.g., yeah, yes, okay) to specifically account for encouraging and reassuring behavior, using the corresponding LIWC category.

4.2 Follower behavior

To measure the behaviors of followers, we use similar behavioral classifications as leaders, evaluated by RAs. The targeted behaviors are the same as those in the leader behavior. The raters count the number of times followers (1) rephrase others’ ideas, (2) elaborate on others’ ideas, and (3) come up with a new idea themselves. These measures represent their

participation and engagement in the team discussion. In addition, we also use the number of contributions and words per follower. We are interested in investigating whether these behaviors by followers are linked to the leader’s gender.

4.3 Effectiveness

As described in the conceptual framework, we define effectiveness in terms of productivity as well as follower outcomes. We measure leaders’ effectiveness by considering the quality of the output generated during the meeting, employee self-evaluation of satisfaction with and engagement in the meeting, external evaluation of the meeting quality, and external evaluation of the ideas generated by teams.

Follower engagement and satisfaction

Follower engagement and satisfaction are measured based on survey data collected right after the meeting. To measure self-evaluated engagement in the group discussion, we construct a measure based on three questions, all rated on a five-point Likert scale: “During the group discussion I . . . ” 1) saw myself as an important part of the discussion, 2.) did my best to come up with good ideas, and 3) experienced the discussion as interesting. Internal consistency was considered adequate (Cronbach’s alpha = 0.73; see Table A.3 in the appendix). To measure the self-evaluated satisfaction by followers, we use participants’ responses on a five-point Likert scale to the following question: “All in all, I was very satisfied with the discussion I had with my colleagues about our company’s future.” This single-item measure is an adapted version of the one-item measure by [Dolbier et al. \(2005\)](#).¹⁶

Overall meeting quality

The same four trained RAs who rate leader behavior also evaluate the overall quality of meetings if they perceive the meeting to be *engaging* (yes, no) and *good* (yes, no). After rating the meeting individually, they agree on a rating, which is what we use in the analysis. For each meeting, we have the individual score of two raters. For engagement, the raters agree in 80% of the meetings. Concerning the overall meeting quality, they agree in 95% of the meetings.

¹⁶We used this single-item measure, which performs similarly to a multi-item measure, in order to keep our survey short. See [Wanous et al. \(1997\)](#) for a discussion of the appropriateness of using single-item measures to elicit work satisfaction.

Output quality

To measure the quality of ideas, four HR experts of the company evaluate the suggested ideas based on their value for the strategic focus of the company. These evaluators receive a document in which all ideas from all teams are listed in an anonymized way, so they have no information on who the leaders or followers were or who suggested the ideas. The evaluators then rate each idea based on its relevance to the strategy discussion and its originality on a scale from 1 to 5.¹⁷ After evaluating each idea individually, the raters agree on a common rating for each idea. We use this harmonized rating in our analysis.

4.4 Empirical specifications

We investigate gender differences in behaviors and effectiveness using the following general regression setting:

$$Y_i = \alpha + \gamma Female_i + \beta' \mathbf{X}_i + \epsilon_i, \quad (1)$$

where Y_i is the behavior/performance measure for leader/team i ; α is the constant term; γ captures the estimated gender difference for female leaders, with $Female_i$ as an indicator for whether the leader is female; \mathbf{X}_i is a vector of the control variables. The included control variables are team size, leader tenure, leader age, average tenure of team members, and the hierarchical level of the leader—a binary variable where a higher-level leader indicates that the person has personnel responsibility for those who themselves have personnel responsibilities.¹⁸ β is the vector of coefficients for the control variables; and ϵ_i is the error term. We present results of OLS regressions in the main part of the paper but show robustness to using other regression models to account for the data structure in the appendix. For follower-level outcomes, we estimated Equation (1) at the follower level, clustering standard errors at the team level to account for intra-team correlations.

5 Results

In this section, we present the results from estimating the empirical model depicted generally in Equation (1). We also report p -values adjusted for multiple hypothesis testing using the Westfall and Young procedure (Jones et al., 2019), accounting for the large number of

¹⁷A detailed description of each category is provided in the appendix in Table A.1.

¹⁸Additionally, we control for a leadership training indicator, a binary variable for leaders who went through the supportive leadership training as a randomized treatment in the RCT. Haeckl and Rege (2025) also control for leaders' coaching behavior evaluated by their real followers. We abstain from including this control variable as it is not a good predictor of outcomes. Adding it, however, does not change our results.

measures. This procedure corrects for the increased risk of Type I error associated with testing multiple hypotheses. We first present results for behaviors of leaders and followers and then investigate differences in leader effectiveness. Lastly, we also test for interaction effects with the leadership training employed in the RCT by analyzing the group receiving the leadership training and the group that did not receive it separately.

5.1 Leader behavior

To investigate the gender differences in leader behavior, we use the outcome measures labeled as “Leader Behavior” in Table 2. We first focus on the measures that we consider subjective and then look at the objective, language-based measures.

Subjective leader behavior

Subjective leader behavior comprises leaders’ actions to enhance follower engagement. On average, leaders rephrase others’ ideas four times in a meeting (ranging from 0 to 16), elaborate on others’ ideas twice (ranging from 0 to 8), and contribute four ideas themselves (ranging from 0 to 11). The distributions of these variables are provided in Figure A.1 in the appendix. Regression results on these behavioral measures are shown in columns (1) - (3) in Table 3.¹⁹ After controlling for multiple hypothesis testing, we only find statistically significant gender differences in the number of times leaders elaborate on others’ ideas. Female leaders elaborated on followers’ ideas, on average, 0.9 times more often than male leaders during a 30-minute discussion.

On average, followers rate their leaders’ supportiveness with 4.4 out of 5 (ranging from 2.2 to 5, see also Figure A.2 in the appendix), and there is no statistically significant difference between male and female leaders (see column (4) in Table 3). As there is evidence that female leaders receive lower evaluations from male followers than from female followers (De Paola et al., 2022), we test whether this is also the case in our study. In Column (1) of Table A.6 in the appendix, we estimate the model with an interaction term for follower gender and find no gender difference in how male and female followers evaluate their leaders.

Concerning the measure based on RAs’ evaluation of the overall supportiveness of leaders, we see that, on average, leaders engage in 4.5 of these behaviors (see Figure A.2 in the appendix). The fifth column of Table 3 shows the estimation results, and we do not find any significant gender differences in this type of behavior. The point estimate is close to zero, indicating that male and female leaders are equally supportive during the meetings.

¹⁹In Table A.4 in the appendix, we show that the results remain qualitatively the same if we use Negative Binomial regressions to account for the fact that we have count data instead.

Table 3: Subjective measures of leader behavior observed during the discussion

	(1)	(2)	(3)	(4)	(5)
	Rephrase Idea	Elaborate Idea	New Idea	Supportive Behavior (fol)	Supportive Behavior (RA)
Female	0.268 (0.682)	0.945** (0.377)	0.853* (0.500)	0.043 (0.065)	0.118 (0.340)
Team size	0.157 (0.519)	-0.664*** (0.236)	-0.971*** (0.297)	-0.062 (0.046)	0.137 (0.247)
Leader tenure	-0.072 (0.237)	-0.123 (0.148)	0.097 (0.200)	-0.014 (0.031)	-0.105 (0.153)
Team tenure	0.096 (0.244)	0.223 (0.184)	0.285 (0.227)	-0.032 (0.029)	0.168 (0.150)
Leader level	0.623 (0.831)	-0.647 (0.395)	-0.654 (0.636)	-0.030 (0.077)	-0.441 (0.384)
Constant	3.332* (1.977)	4.380*** (0.909)	7.437*** (1.140)	4.531*** (0.175)	3.514*** (0.914)
Observations	125	125	125	302	128
Adjusted R^2	-0.036	0.103	0.083	0.017	0.063

Notes: OLS regressions. Numbers in parentheses show robust standard errors. Leader level is a binary indicator for higher-level leaders. We also include an indicator for the leadership training. After correcting for multiple hypothesis testing, the result for the female coefficient remains significant at the 10% level in column (2).*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Objective leader behavior

Next, we investigate objective measures of leader behavior. The estimation results are shown in Table 4 and the distributions of all dependent variables are provided in Figure A.3 in the appendix. The first column shows the number of times the leader speaks during the meeting, and the second column shows the number of words the leader speaks. On average, leaders speak 116 times (ranging from 42 to 248 times) and use 2152 words (ranging from 585 to 5007 words). Together, we interpret them as a measure of influence and engagement. If a leader speaks many times and uses many words, it shows the dominance of the leader as the leader occupies much of the talking space. However, if the leader speaks often but not many words, this is more consistent with higher engagement, where leaders may speak to encourage and invite others to contribute. The latter is indeed what we find: female leaders speak significantly more often (28 times more on average) but not necessarily more words (spoke about 48 words more, but it is not statistically significant).

Columns (3) and (4) show the results for agency and communion words. Overall, leaders use an equal share of agency and communion words. On average, 6% of the words used by leaders are agentic (ranging from 4% to 8%) and 6% of the words are communal (ranging from 4% to 9%). The estimation results show no significant differences in to use of words from the agency and communion dictionaries.

Table 4: Objective measures of leader behavior observed during the discussion

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Number of Contributions	Number of Words	Agency	Communion	Status	Authenticity	CDI	Assent
Female	28.028*** (7.888)	47.551 (121.623)	-0.069 (0.173)	-0.286 (0.188)	-0.589 (0.440)	0.153 (0.278)	-0.111 (0.539)	1.094** (0.429)
Team size	-8.526* (5.102)	-257.561** (107.238)	-0.047 (0.110)	0.311** (0.128)	0.569* (0.288)	0.067 (0.176)	0.240 (0.347)	0.221 (0.263)
Leader tenure	-1.532 (3.267)	37.768 (51.307)	-0.011 (0.071)	0.063 (0.095)	-0.074 (0.224)	0.305** (0.141)	0.154 (0.216)	-0.137 (0.168)
Team tenure	3.636 (2.898)	36.570 (53.013)	0.135* (0.077)	0.120 (0.082)	-0.006 (0.177)	0.038 (0.127)	-0.002 (0.253)	-0.050 (0.123)
Leader level	-15.994** (7.421)	-144.376 (155.042)	0.072 (0.213)	0.433* (0.225)	0.961** (0.430)	0.408 (0.342)	1.383** (0.578)	-0.611* (0.347)
Constant	138.805*** (19.018)	2,895.052*** (404.369)	6.184*** (0.398)	4.969*** (0.454)	7.384*** (1.025)	8.923*** (0.678)	4.783*** (1.340)	3.191*** (1.034)
Observations	125	125	125	125	125	125	125	125
Adjusted R^2	0.126	0.107	-0.006	0.043	0.016	0.018	0.000	0.048

Notes: OLS regressions. The dependent variables in columns (3), (4) and (8) are word counts based on dictionaries. The dependent variables in columns (5) - (7) are based on language styles. Numbers in parentheses show robust standard errors. Leader level is a binary indicator for higher-level leaders. We also include an indicator for the leadership training. After correcting for multiple hypothesis testing the result for the female coefficient remains significant at the 5% level in columns (1) and (8).*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Next, we look at measures of status, authenticity, and the Categorical-Dynamic Index (CDI) in columns (5)-(7), where we find no statistically significant differences.²⁰

Lastly, we look at words of assent. On average, leaders show assent with 4% of the words they use (ranging from 1% to 12%). We find that female leaders use more assent words (column (7)), and this effect is statistically significant. Higher usage of assent words is consistent with communal behavior, as assent words provide support and encouragement to followers. The coefficient is 1.09, which indicates that female leaders use about one percentage point more assent words than male leaders. This difference translates to about 22 more assent words in the 30-minute discussion (the average word count for female leaders is 2,200).

Overall, we find some support for the hypothesis that female leaders behave more communally than male leaders (H1). While we do not find differences in the overall supportive behavior score, we find that female leaders speak more frequently, are more likely to elaborate on followers' ideas, and use more assent words.

5.2 Follower behavior

We investigate the behavior of followers using regression analysis on the individual level. Here, our interest is in looking into whether followers behave differently based on the leader's gender by hypothesizing that followers show more active participation when the team is led by a female leader (H2). We present the estimation results in Table 5 and the distributions of the variables in Figure A.4. While we find that followers in female-led teams contribute more often and elaborate on teammates' ideas more frequently, these differences are no longer statistically significant once we control for multiple hypothesis testing. In addition, we find no significant differences in word counts, the number of times they rephrase others' ideas, or the number of ideas contributed by the followers. We, therefore, conclude that we do not find support for H2.

5.3 Effectiveness

To evaluate leader effectiveness, Table 6 summarizes regression results on follower-reported engagement and satisfaction, overall quality of the meeting as perceived by RAs, and output quality evaluated by HR experts. We also show the distribution of each of the non-binary variables in Figure A.5 in the appendix.

²⁰On average, leaders use high-status language in 9% of the words (ranging from 1% to 17%), 9% of the words used by leaders are authentic (ranging from 6% to 13%) and 6% of the words are logical/categorical (ranging from 0% to 14%).

Table 5: Measures of follower behavior observed during the discussion

	(1) Contributions follower	(2) Words follower	(3) Rephrase follower	(4) Elaborate follower	(5) Ideas follower
Female leader	9.780** (4.758)	99.762 (117.625)	0.240 (0.204)	0.568* (0.307)	0.121 (0.257)
Leader tenure	-3.710* (2.007)	16.114 (47.824)	-0.058 (0.076)	-0.111 (0.114)	-0.083 (0.104)
Female	15.066*** (4.177)	-13.868 (43.090)	0.199 (0.166)	-0.027 (0.222)	0.378 (0.300)
Team size	-24.040*** (3.657)	-232.527** (94.107)	0.035 (0.141)	-0.824*** (0.190)	-0.962*** (0.245)
Team Tenure	4.559** (1.862)	9.155 (31.165)	0.083 (0.093)	0.301** (0.125)	0.146 (0.123)
Leader level	-8.211* (4.677)	-157.899 (143.753)	-0.405** (0.187)	-0.190 (0.351)	-0.159 (0.300)
Constant	159.722*** (14.995)	2,795.882*** (359.921)	1.241** (0.557)	5.616*** (0.783)	8.037*** (0.934)
Observations	318	318	318	318	318
Adjusted R^2	0.203	0.122	0.001	0.077	0.060

Notes: OLS regressions. Leader level is a binary indicator for higher-level leaders. We also include an indicator for the leadership training. Numbers in parentheses indicate standard errors clustered at the team level. After correcting for multiple hypothesis testing, the result for the female coefficient is no longer statistically significant.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Follower engagement and satisfaction

On average, followers report being fairly engaged in and satisfied with the meetings (average score of 4, ranging from 3 to 5 for engagement and 2 to 5 for satisfaction). Columns (1) and (2) of Table 6 show no differences in follower self-reported engagement and satisfaction by leader gender. Thus, we do not find evidence in support of H3, when we look at the average effect. In Section 5.4, however, we show that this null effect at the average level is caused by interaction effects with the leadership training provided in the RCT. We also test if followers' engagement and satisfaction are affected by follower gender but we find no significant effect (see Table A.6 in the appendix).

Table 6: Output Quality and Effectiveness

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Engagement follower	Satisfaction follower	Good Meeting (RA)	Engaging Meeting (RA)	Relevance	Originality	Total Score
Female	-0.017 (0.080)	0.134 (0.091)	0.022 (0.036)	0.155** (0.060)	-0.015 (0.114)	-0.031 (0.110)	-0.023 (0.104)
Team size	-0.063 (0.058)	-0.032 (0.075)	0.016 (0.023)	-0.006 (0.047)	0.061 (0.096)	0.031 (0.103)	0.046 (0.094)
Leader tenure	0.035 (0.031)	-0.029 (0.039)	-0.043 (0.030)	-0.043 (0.043)	-0.061 (0.042)	-0.051 (0.049)	-0.056 (0.043)
Team tenure	-0.055* (0.028)	-0.043 (0.036)	-0.010 (0.015)	0.002 (0.039)	-0.013 (0.046)	0.023 (0.049)	0.005 (0.044)
Leader level	0.118 (0.071)	-0.006 (0.093)	0.055** (0.025)	0.002 (0.080)	0.398*** (0.109)	0.337*** (0.125)	0.368*** (0.108)
Constant	4.466*** (0.233)	4.424*** (0.307)	0.905*** (0.088)	0.727*** (0.179)	2.528*** (0.322)	2.359*** (0.356)	2.443*** (0.321)
Observations	302	302	128	128	130	130	130
Adjusted R^2	0.010	0.019	0.020	0.023	0.038	0.012	0.033

Notes: OLS regressions. Leader level is a binary indicator for higher-level leaders. We also include an indicator for the leadership training. Robust standard errors in parentheses. In columns (1) and (2) standard errors are clustered at the team level. After correcting for multiple hypothesis testing, the result for the female coefficient is no longer statistically significant.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Meeting quality

Concerning the overall rating of the meeting, there is little variation in the sample to begin with. The research assistants rate 96% of the meetings as good and 80% as engaging. Still, the share of engaging meetings is 16 percentage points higher in meetings with a female leader compared to meetings with a male leader (92% compared to 76%, p -value = 0.011, see column (4) in Table 6). However, this difference is not large enough to remain statistically significant after we account for multiple hypothesis testing. We conclude that we do not find support for H4.

Output quality

Lastly, we look at the rating of HR experts. On average, they rate ideas as 3 out of 5 in terms of originality (ranging from 1 to 4) and 3 out of 5 in terms of relevance (ranging from 1 to 5). The combined performance measure also has an average of 3 and ranges from 1 to 4. There are no significant differences between male and female-led teams in terms of the quality of the ideas they contribute, and the estimates are very close to zero.²¹ Thus, we do not find support for H5.

Leader levels

In general, we do not find any significant effect of leader gender on effectiveness. An interesting observation, though, is that the level of the leader has a significant influence on the effectiveness, such that the teams with higher-level leaders produce significantly better ideas, in terms of both originality and relevance. In addition, meetings led by higher-level leaders are more likely to be evaluated as “good meetings” by research assistant raters. Looking at the objective measures of leader behavior in Table 4, we see some significant behavioral differences. We will further elaborate on this point in the discussion section.

5.4 Interactions between leader gender and the RCT intervention

Using data from an RCT provides us not only with the opportunity to investigate randomly composed teams but also allows us to investigate whether leadership training mitigates gender differences. The intervention in the RCT was a 20-minute leadership training aiming at increasing leaders’ supportiveness. Supportive leadership training arguably increases the communal behavior of leaders and previous research indicated that the training was particularly effective for male leaders (see [Haeckl and Rege, 2025](#), for details). It is plausible that

²¹The results do not change when we use alternative measures, such as the best idea from each team. Results are available upon request from authors.

male leaders in the leadership-training group behave extra-communally. Obviously, such effects also make it more difficult to identify gender differences in leader behavior, looking at the pooled sample. To account for this concern and gain deeper insight into the role of leadership training in mitigating gender differences, we replicate the analysis above separately by leadership-training status in Appendix B.

This analysis brings two interesting insights. First, we find that the observed significant gender differences in objective measures are driven by the group not receiving the training (Table B.2: Panel B). For example, absent the leadership training, female leaders speak, on average, 45 times more often and use words of assent two percentage points more frequently than male leaders. The respective numbers in the treated group are 14 and 0.2, and these numbers are not statistically different from those of the male leaders. As the group without any influence of leadership training represents the default setting to observe gender differences, this subgroup analysis provides further support for H1. There are no significant differences by leadership-training status and gender using the target behaviors identified by RAs (see columns (1)-(3) and (5) in Table B.1).

Second, followers' perceptions of leaders' supportiveness exhibit the opposite signs depending on training status (Table B.1). Followers evaluate female leaders to be significantly more supportive than male leaders when leaders have not received the training, but significantly less supportive when they have received the training. A similar pattern emerges for followers' self-reported engagement and satisfaction (Table B.4). This shows an interesting nuance. Absent leadership training, female leaders exhibit more communal behavior from the objective measures (more frequent contribution with higher affirmative words), and their followers respond to this by reporting higher engagement and satisfaction. However, for the group led by leaders receiving the training, male leaders are evaluated as more supportive, even though we find little behavioral difference between male and female leaders (weak evidence that female leaders are more supportive, if any). Providing leaders with a 20-minute supportive leadership training appears to significantly influence how followers evaluate their male leaders' supportiveness and their own engagement and satisfaction. We interpret these results as an indication of male communality bonus (Hentschel et al., 2018) where male leaders are disproportionately rewarded for communal behaviors, even though they are not necessarily more supportive than female leaders based on the objective measures.

6 Discussion and conclusion

This paper examines whether and how leaders' gender impacts their behavior and leadership effectiveness. Using data from an RCT in a real-world business context, we observe male

and female leaders who were randomly assigned to teams to work on a strategic exercise in a digital meeting. The setting of this study brings a number of advantages. First, the random assignment of leaders to teams mitigates potential biases, such as selection effects – where male and female leaders might choose different team members (or are chosen by them). Second, anonymous evaluation of the output mitigates potential evaluation biases, where leader gender could influence perceptions of performance and effectiveness. Third, we use the recording of the meetings to observe actual leader behavior, thereby accounting for what leaders do rather than recounts and perceptions. Fourth, we include followers’ subjective evaluation of leaders to bring insights into the gap between actual performance and perceived performance. Fifth, study participants are real employees and team leaders are actual leaders who hold genuine managerial responsibilities, bringing realism and enhanced validity to our study.

We anchor our investigation on gender differences in leader behavior on two prominent theories: social role theory (SRT) and double standards of competence (DST). Considering the highly collaborative nature of the given task, we generally hypothesize that communal leadership behavior would be more effective. Based on theoretical arguments, we hypothesize that female leaders act more communally, which also leads to higher engagement of followers measured in behaviors, followers’ self-evaluations, and external evaluation. We further hypothesize that female-led teams produce better ideas than male-led teams as a result of effective leadership.

Based on a series of regressions, we find that female leaders act to facilitate high engagement by elaborating on ideas from team members, speaking more frequently (but not more words), and using assent words. Elaborating on team members’ ideas shows that the leader is paying attention and valuing the contribution of the members. Speaking more frequently (but not more words) with higher usage of assent words may indicate that leaders provide assurance, make encouraging comments after one member speaks, and invite the team member or others to contribute further. Such behavior may be an indication that female leaders make sure that everyone is engaged and participating, which is consistent with our suggestive empirical finding that each member contributes more in female-led teams than in male-led teams.

However, we do not find evidence that these facilitation efforts by female leaders lead to higher team performance.²² In particular, the ideas submitted by female-led teams are neither more original nor more relevant. The evaluation of overall meeting quality by RAs

²²Recent literature shows that more engagement does not necessarily lead to better outcomes and suggests that in some settings it might actually be even beneficial to increase the cost of contributing (Charness et al., 2020)

also does not exhibit any differences by leader gender. Thus, external evaluations suggest that male and female leaders are equally effective.²³

The evaluation by followers shows more complex patterns, especially when we also consider interactions with the leadership training from the RCT design (a 20-minute supportive leadership training). Among the group without the leadership training, female leaders exhibit behaviors that facilitate engagement. Unlike the average effects, these women are indeed rewarded by receiving higher evaluations from followers (significantly more supportive), and followers themselves evaluate their engagement and satisfaction higher in this group. This changes drastically in the group receiving leadership training, where female leaders are evaluated as less supportive by followers, and followers also report lower engagement for themselves than in male-led teams. [Haeckl and Rege \(2025\)](#) reported that supportive leadership training was more effective for male leaders than for female leaders. We extend the analysis by showing that supportive leadership training 1) significantly reduced gender differences in communal behaviors and 2) reversed a female advantage in perceived supportiveness and effectiveness in making followers feel engaged and satisfied. We interpret this as the communality bonus for male leaders, such that male leaders receive disproportionately high evaluations by showing high communality ([Hentschel et al., 2018](#)).

The single most influential factor that affected output quality was the level of the leaders. Teams with higher-level leaders produce ideas that are evaluated as highly original and relevant. It is possible that higher-level leaders know more about the company's strategies and, therefore, are able to provide better ideas. In that case, these leaders may be contributing their own ideas without eliciting ideas from others. However, we do not observe such behavioral patterns. We find suggestive evidence that they use more words related to communion, more higher-status language, more logical/categorical language, and fewer assent words, while speaking less frequently. The language styles that higher-level leaders employ contain both communal and agentic behavior, implying that effective leaders can flexibly deploy both types of behavior to deliver high-quality output. It is also possible that other factors are at play, e.g., there may be unobserved characteristics common among high-level leaders (they were promoted to high-level leaders, after all), or having higher-level leaders changes the motivation and mindset of followers. As current research tends to focus on the top management team (e.g., CEOs) when investigating higher-level leaders, not many studies investigate the effectiveness of leadership in smaller units ([Buss et al., 2024](#)). Thus, it could be interesting for future research to further decode the effective leadership behavior of higher-level managers (those with responsibility for managers but not necessarily at the

²³This finding is also in line with [Berle et al. \(2024\)](#), who show that having more women on business committees/boards makes deliberation more thorough and engaged, while it does not affect decisions.

top echelon).

Our results challenge the notion of advantage or disadvantage of female leadership, as we find comparable effectiveness between male and female leaders. We also do not find any systematic downward bias in the evaluation of female leaders by followers. The evidence on male communality bonus also goes against SRT that men are expected to display agency-coded leader behavior, as we find that men are rewarded for being perceived as communal. As the context of the current study is limited, further research may shed light on how communal behavior by male leaders plays a role in modern organizations, and whether this creates an uneven playing field where female leaders, for whom communal behavior is expected, receive lower evaluations than male leaders by exhibiting the same communal behaviors.

Like all framed field experiments, our study is set in specific contexts. Following suggestions in previous research (Eagly, 2007; Eagly et al., 1995), we, therefore, discuss how contextual factors might have influenced our results. First, the nature of the team task employed in this study was collaborative and high in dependency. The meetings are short (one hour) and targeted, where leaders are instructed to lead the team with members who do not know one another. Therefore, leaders are expected to take the initiative and facilitate the discussion. In such settings, it may be more natural for leaders to engage in relation-oriented behaviors than task-oriented behaviors, and fewer gender differences in leader behavior may be expected. In addition, both male and female leaders are assigned exactly the same tasks in this study. It is not the case in many organizational studies where male and female leaders occupy different leader ranks, so their tasks, and therefore behaviors, are inherently different (Eagly et al., 2003).

Second, the country/cultural setting of this study may be another contextual factor influencing gender differences — an experiment run with a company in Norway. Norway is consistently one of the most gender-equal countries in the world (e.g., ranked third in the 2024 Global Gender Gap Report by the World Economic Forum) with a strong egalitarian value. Thus, the workplace culture may be more gender-neutral in the sense that male and female leaders behave similarly and more communally. The Norwegian gender-equal context, combined with the fact that participants in leadership roles are actual company leaders, might have contributed to the adequate legitimization of leadership authority, which female leaders sometimes struggle to gain in other contexts (Sidhu et al., 2021). It is also worth noting that the Norwegian labor force is highly gender segregated, such that the majority (63%) of workers in the private sector are men (Statistics Norway, 2024). The company we collaborated with was a multi-utility company with more male employees in general. Thus, there may be a self-selection of more agentic women in private sector jobs and into leadership roles in private firms. Such self-selection may also contribute to similar behavior between

male and female leaders.

Third, past studies find that independent effectiveness measures or objective measures show a weaker effect on leader effectiveness compared to measures based on surveys (Burke et al., 2006; Kaiser et al., 2008). In line with these findings, our survey-based measures (follower evaluation) yield more significant effects (especially with leadership-training interactions) than external evaluations (no or weak gender difference in effectiveness).

It is interesting to note that the contextual factors discussed above would generally tend to minimize gender differences in leadership. The gender-egalitarian Norwegian culture, the specific collaborative task structure, the legitimacy afforded to established leaders, and the employed measurements are all expected to reduce gender-based behavioral variations. Despite these mitigating factors, we still observe a number of significant gender differences in leadership behaviors. Conversely, these same contextual elements may help explain why these behavioral differences did not translate into a significant difference in leadership effectiveness. Combined, these findings suggest that gender differences in leadership behavior may be more persistent than differences in leadership outcomes, particularly in environments designed to promote equality.

Our study design offers both substantial advantages (clean identification, comparable teams, observability of actual behavior) but also limitations (specific task, specific country context). To overcome these limitations future research should explore longer-term outcomes and consider diverse organizational and cultural contexts. Despite these limitations, this study makes valuable contributions to evidence-based leadership development and organizational decision-making. Our findings challenge the notion of substantial gender differences (either advantage or disadvantage) in leadership. The results suggest that effective leadership is more strongly linked to developed expertise than to inherent gender-based differences. These insights have important implications for leadership development and organizational policies, suggesting that organizations may benefit from focusing on building leadership capabilities broadly while implementing balanced evaluation approaches that minimize potential gender biases in leadership assessment.

References

- Abele, A. E., Uchronski, M., Suitner, C., and Wojciszke, B. (2008). Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology*, 38(7):1202–1217.
- Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly*, 28(1):5–21.
- Antonakis, J., Bendahan, S., Jacquart, P., and Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(6):1086–1120. Leadership Quarterly Yearly Review.
- Banks, G. C., Woznyj, H. M., and Mansfield, C. A. (2023). Where is “behavior” in organizational behavior? a call for a revolution in leadership research and beyond. *The Leadership Quarterly*, 34(6):101581.
- Berle, E. C., Kavajecz, K., and Onozaka, Y. (2024). Effect of gender composition of committees. *Human Relations*, 77(5):622–649.
- Biernat, M. and Kobrynowicz, D. (1997). Gender-and race-based standards of competence: lower minimum standards but higher ability standards for devalued groups. *Journal of personality and social psychology*, 72(3):544.
- Brauer, M., Wiersema, M., and Binder, P. (2023). “Dear CEO and Board”: How activist investors’ confidence in tone influences campaign success. *Organization Science*, 34(4):1487–1508.
- Brown, S. G., Hill, N. S., and Lorinkova, N. N. M. (2021). Leadership and virtual team performance: A meta-analytic investigation. *European Journal of Work and Organizational Psychology*, 30(5):672–685.
- Burke, C. S., Stagl, K. C., Klein, C., Goodwin, G. F., Salas, E., and Halpin, S. M. (2006). What type of leadership behaviors are functional in teams? a meta-analysis. *The leadership quarterly*, 17(3):288–307.
- Buss, M., Andler, S., and Tiberius, V. (2024). Female leadership: An integrative review and research framework. *The Leadership Quarterly*, page 101858.
- Carton, A. M. (2022). The science of leadership: A theoretical model and research agenda. *Annual Review of Organizational Psychology and Organizational Behavior*, 9(1):61–93.

- Chakraborty, P. and Serra, D. (2023). Gender and leadership in organisations: the threat of backlash. *The Economic Journal*, 134(660):1401–1430.
- Charness, G., Cooper, D. J., and Grossman, Z. (2020). Silence is golden: Team problem solving and communication Costs. *Experimental Economics*, 23(3):668–693.
- De Paola, M., Gioia, F., and Scoppa, V. (2022). Female leadership: Effectiveness and perception. *Journal of Economic Behavior and Organization*, 201:134–162.
- Dolbier, C. L., Webster, J. A., McCalister, K. T., Mallon, M. W., and Steinhardt, M. A. (2005). Reliability and validity of a single-item measure of job satisfaction. *American Journal of Health Promotion*, 19(3):194–8.
- Eagly, A. H. (2007). Female leadership advantage and disadvantage: Resolving the contradictions. *Psychology of Women Quarterly*, 31(1):1–12.
- Eagly, A. H. and Carli, L. L. (2003a). The female leadership advantage: An evaluation of the evidence. *The Leadership Quarterly*, 14(6):807–834.
- Eagly, A. H. and Carli, L. L. (2003b). Finding gender advantage and disadvantage: Systematic research integration is the solution. *The Leadership Quarterly*, 14(6):851–859.
- Eagly, A. H., Johannesen-Schmidt, M. C., and Van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: a meta-analysis comparing women and men. *Psychological Bulletin*, 129(4):569.
- Eagly, A. H. and Johnson, B. T. (1990). Gender and leadership style: A meta-analysis. *Psychological Bulletin*, 108(2):233.
- Eagly, A. H. and Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3):573.
- Eagly, A. H., Karau, S. J., and Makhijani, M. G. (1995). Gender and the effectiveness of leaders: a meta-analysis. *Psychological Bulletin*, 117(1):125.
- Eagly, A. H., Wood, W., and Diekmann, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. *The Developmental Social Psychology of Gender*, 12(174):9781410605245–12.
- European Foundation for the Improvement of Living and Working Conditions, European Centre for the Development of Vocational Training (2020). European company survey, 2019. <http://doi.org/10.5255/UKDA-SN-8691-1> [Accessed: (01.11.2024)].

- Eurostat (2023). Online meetings and remote access to enterprise resources - statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Online_meetings_and_remote_access_to_enterprise_resources_-_statistics[Accessed: (01.11.2024)].
- Fischer, T., Hambrick, D. C., Sajons, G. B., and Van Quaquebeke, N. (2023). Leadership science beyond questionnaires. *The Leadership Quarterly*, 34(6):101752. Beyond the ritualized use of questionnaires: Toward a science of actual behaviors and psychological states.
- Foschi, M. (2000). Double standards for competence: Theory and research. *Annual Review of Sociology*, 26(1):21–42.
- Gangadharan, L., Jain, T., Maitra, P., and Vecci, J. (2016). Social identity and governance: The behavioral response to female leaders. *European Economic Review*, 90:302–325.
- Gangadharan, L., Jain, T., Maitra, P., and Vecci, J. (2019). Female leaders and their response to the social environment. *Journal of Economic Behavior and Organization*, 164:256–272.
- Grossman, P. J., Eckel, C., Komai, M., and Zhan, W. (2019). It pays to be a man: Rewards for leaders in a coordination game. *Journal of Economic Behavior & Organization*, 161:197–215.
- Guilford, J. P. (1967). Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, 1(1):3–14.
- Haeckl, S. and Rege, M. (2025). Effects of supportive leadership behaviors on employee satisfaction, engagement, and performance: An experimental field investigation. *Management Science*, 71(1):347–365.
- Heilman, M. E., Block, C. J., and Martell, R. F. (1995). Sex stereotypes: Do they influence perceptions of managers? *Journal of Social behavior and Personality*, 10(4):237.
- Hemshorn de Sanchez, C. S., Gerpott, F. H., and Lehmann-Willenbrock, N. (2022). A review and future agenda for behavioral research on leader–follower interactions at different temporal scopes. *Journal of Organizational Behavior*, 43(2):342–368.
- Hentschel, T., Braun, S., Peus, C., and Frey, D. (2018). The communality-bonus effect for male transformational leaders–leadership style, gender, and promotability. *European Journal of Work and Organizational Psychology*, 27(1):112–125.

- Heursen, K., Reuben, E., and Rott, C. (2023). Are women less effective leaders than men? evidence from experiments using coordination games. University of Zurich, Unpublished Manuscript.
- Jones, D., Molitor, D., and Reif, J. (2019). What do W workplace wellness programs do? Evidence from the Illinois workplace wellness study. *The Quarterly Journal of Economics*, 134(4):1747–1791.
- Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., and Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143.
- Kaiser, R. B., Hogan, R., and Craig, S. B. (2008). Leadership and the fate of organizations. *American Psychologist*, 63(2):96.
- Körner, R., Overbeck, J. R., Körner, E., and Schütz, A. (2024). The language of power: Interpersonal perceptions of sense of power, dominance, and prestige based on word usage. *European Journal of Personality*, 38(5):812–838.
- Macchiavello, R., Menzel, A., Rabbani, A., and Woodruff, C. (2020). Challenges of change: An experiment promoting women to managerial roles in the bangladeshi garment sector. Technical report, National Bureau of Economic Research.
- Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.
- Paustian-Underdahl, S. C., Sockbeson, C. E. S., Hall, A. V., and Halliday, C. S. (2024). Gender and evaluations of leadership behaviors: A meta-analytic review of 50 years of research. *The Leadership Quarterly*, page 101822.
- Paustian-Underdahl, S. C., Walker, L. S., and Woehr, D. J. (2014). Gender and perceptions of leadership effectiveness: a meta-analysis of contextual moderators. *Journal of Applied Psychology*, 99(6):1129.
- Pennebaker, J. W., Booth, R., Boyd, R. L., and Francis, M. E. (2015). Linguistic inquiry and word count: Liwc2015. www.LIWC.net.
- Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., and Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS One*, 9(12):e115844.

- Pietraszkiewicz, A., Formanowicz, M., Gustafsson Sendén, M., Boyd, R. L., Sikström, S., and Sczesny, S. (2019). The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology*, 49(5):871–887.
- Pitt, C. (2021). Gender and the cmo: do the differences make a difference? *Journal of Strategic Marketing*, 29(4):301–315.
- Reuben, E. and Timko, K. (2018). On the effectiveness of elected male and female leaders and team coordination. *Journal of the Economic Science Association*, 4(2):123–135.
- Rosette, A. S. and Tost, L. P. (2010). Agentic women and communal leadership: How role prescriptions confer advantage to top women leaders. *Journal of Applied Psychology*, 95(2):221.
- Rudman, L. A. and Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, 57(4):743–762.
- Rudman, L. A. and Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior*, 28:61–79.
- Sidhu, J. S., Feng, Y., Volberda, H. W., and Van Den Bosch, F. A. (2021). In the shadow of social stereotypes: Gender diversity on corporate boards, board chair’s gender and strategic change. *Organization Studies*, 42(11):1677–1698.
- Statistics Norway (2024). Indicators for gender equality in municipalities. Accessed: 2025-04-19.
- Stock, G., Banks, G. C., Voss, E. N., Tonidandel, S., and Woznyj, H. (2023). Putting leader (follower) behavior back into transformational leadership: A theoretical and empirical course correction. *The Leadership Quarterly*, 34(6):101632.
- Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24–54.
- Timko, K. (2017). Gender, communication styles, and leader effectiveness. MPRA Paper 77021, University Library of Munich, Germany.
- Vecchio, R. P. (2002). Leadership and gender advantage. *The Leadership Quarterly*, 13(6):643–671.

- Vecchio, R. P. (2003). In search of gender advantage. *The Leadership Quarterly*, 14(6):835–850.
- Wanous, J. P., Reichers, A. E., and Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, 82(2):247–252.
- Yukl, G. (2012). Effective leadership behavior: What we know and what questions need more attention. *Academy of Management perspectives*, 26(4):66–85.

Appendix

A Additional figures and tables

Table A.1: Scales used by the evaluators to rate the quality of ideas.

Originality	
5	The proposed idea is novel, unique, exciting, brings up a new topic.
4	The proposed idea is unusual, exciting, brings up a new aspect.
3	The proposed idea is interesting.
2	The proposed idea is interesting but has already been discussed in the company (is not new).
1	The proposed idea is common, mundane, boring.
Relevance	
5	We will definitely include this idea in the strategy discussion.
4	We might include this idea in the strategy discussion.
3	We will not include this concrete idea but the general topic in the strategy discussion.
2	We will probably not consider this idea or the general topic in the strategy discussion.
1	We will definitely not consider this idea or the general topic in the strategy discussion.

Table A.2: Satisfaction

	Mean	SD	N
Satisfaction discussion	4.42	0.66	302

Notes: Descriptive statistics for the survey measure for satisfaction with the group task.

Table A.3: Engagement

	Mean	SD	N	Cronbach's Alpha
Engagement	4.30	0.57	302	0.73
Engagement1	4.13	0.80	302	
Engagement2	4.28	0.73	302	
Engagement3	4.49	0.60	302	

Notes: Descriptive statistics for the survey measure for follower engagement in the group task and its components Engagement1-3.

Table A.4: Subjective measures of leader behavior observed during the discussion - Negative binomial regression

	(1) Rephrase Idea	(2) Elaborate Idea	(3) New Idea	(4) Supportive Behavior
Female	0.063 (0.153)	0.397*** (0.140)	0.194* (0.107)	0.024 (0.072)
Team size	0.036 (0.122)	-0.286*** (0.091)	-0.227*** (0.066)	0.030 (0.053)
Leader tenure	-0.014 (0.061)	-0.059 (0.070)	0.022 (0.043)	-0.024 (0.034)
Team tenure	0.020 (0.055)	0.095 (0.073)	0.061 (0.048)	0.037 (0.032)
Leader level	0.140 (0.175)	-0.326* (0.185)	-0.158 (0.156)	-0.097 (0.085)
Constant	1.239*** (0.459)	1.700*** (0.343)	2.174*** (0.246)	1.278*** (0.198)
Observations	125	125	125	128

Notes: Negative binomial regressions. Leader level is a binary indicator for high-level leaders. We also include an indicator for the leadership training. Numbers in parentheses show robust standard errors. After correcting for multiple hypothesis testing, the result for the female coefficient remains significant at the 10% level in column (3).*** p<0.01, ** p<0.05, * p<0.1

Table A.5: Measures of follower behavior observed during the discussion - Negative binomial regression

	(1) Nr Rephrase follower	(2) Nr Elaborate follower	(3) Nr Ideas follower
Female leader	0.176 (0.137)	0.213** (0.107)	0.027 (0.057)
Leader tenure	-0.041 (0.059)	-0.044 (0.047)	-0.019 (0.024)
Female	0.149 (0.116)	-0.000 (0.087)	0.084 (0.066)
Team size	0.026 (0.099)	-0.320*** (0.072)	-0.217*** (0.053)
Team Tenure	0.061 (0.061)	0.109*** (0.041)	0.034 (0.027)
Leader level	-0.342** (0.161)	-0.086 (0.132)	-0.034 (0.069)
Constant	0.215 (0.398)	2.095*** (0.290)	2.293*** (0.199)
Observations	318	318	318

Notes: Negative binomial regressions. Leader level is a binary indicator for high-level leaders. We also include an indicator for the leadership training. Numbers in parentheses indicate standard errors clustered at the team level. *** p<0.01, ** p<0.05, * p<0.1

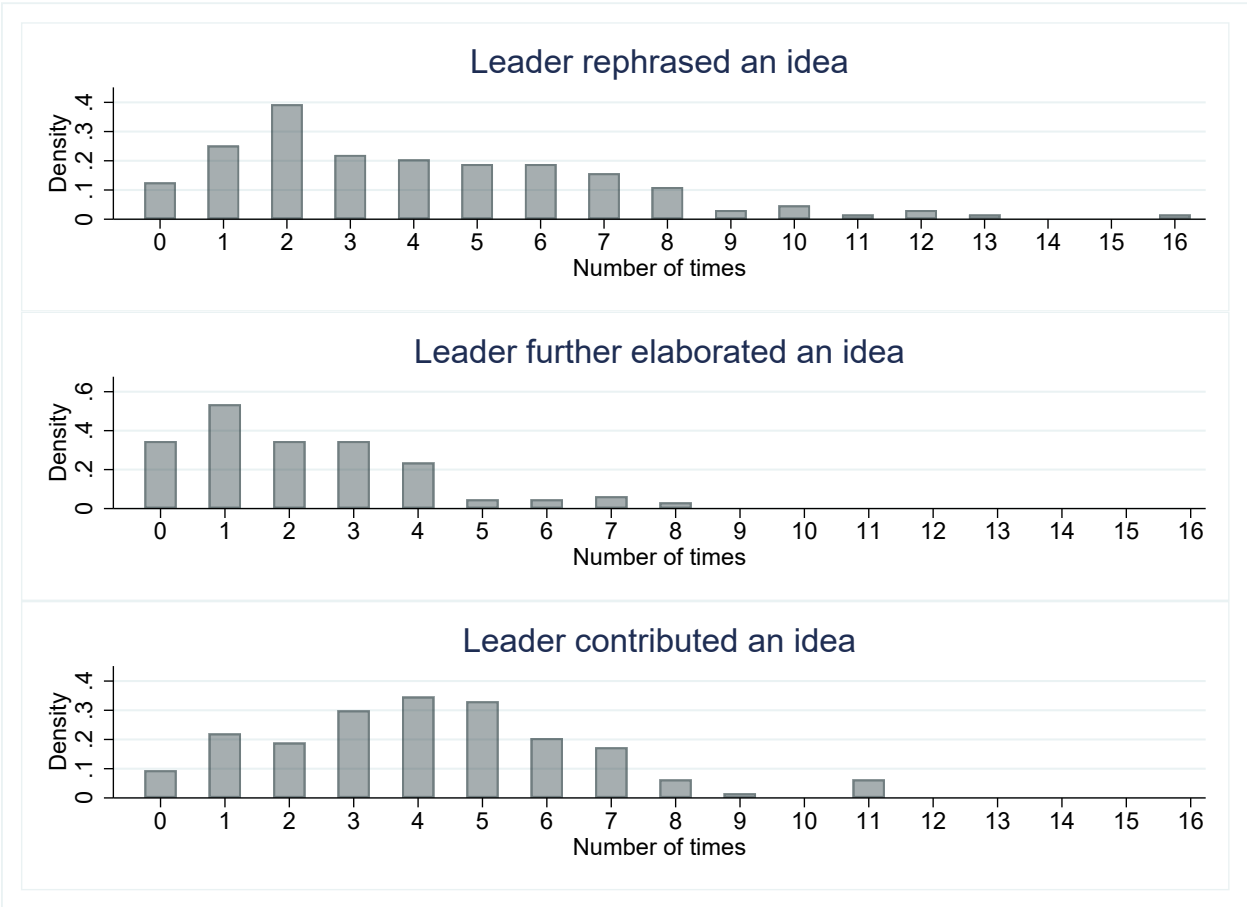


Figure A.1: Number of times leaders engaged in certain behaviors

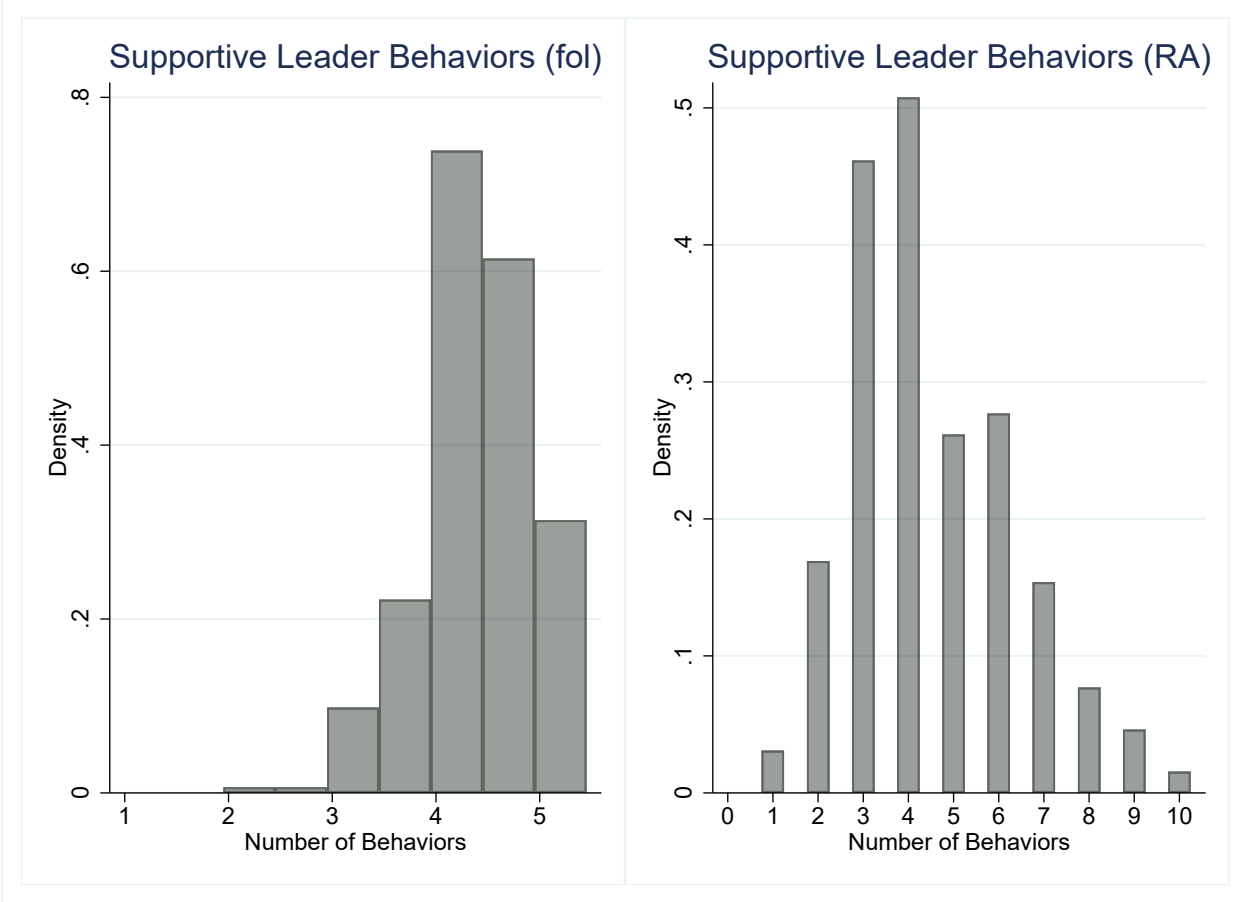


Figure A.2: Supportive behavior score

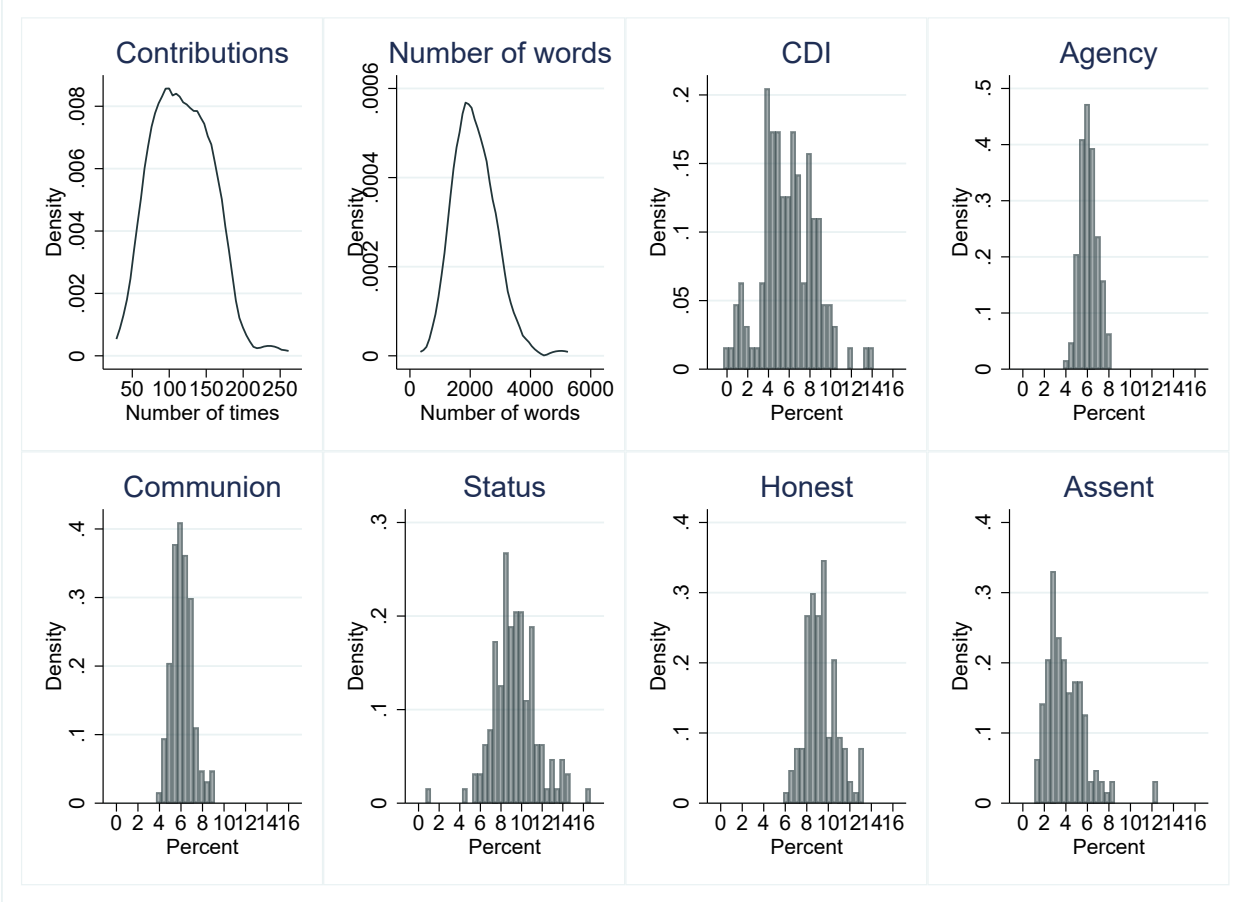


Figure A.3: Leader language

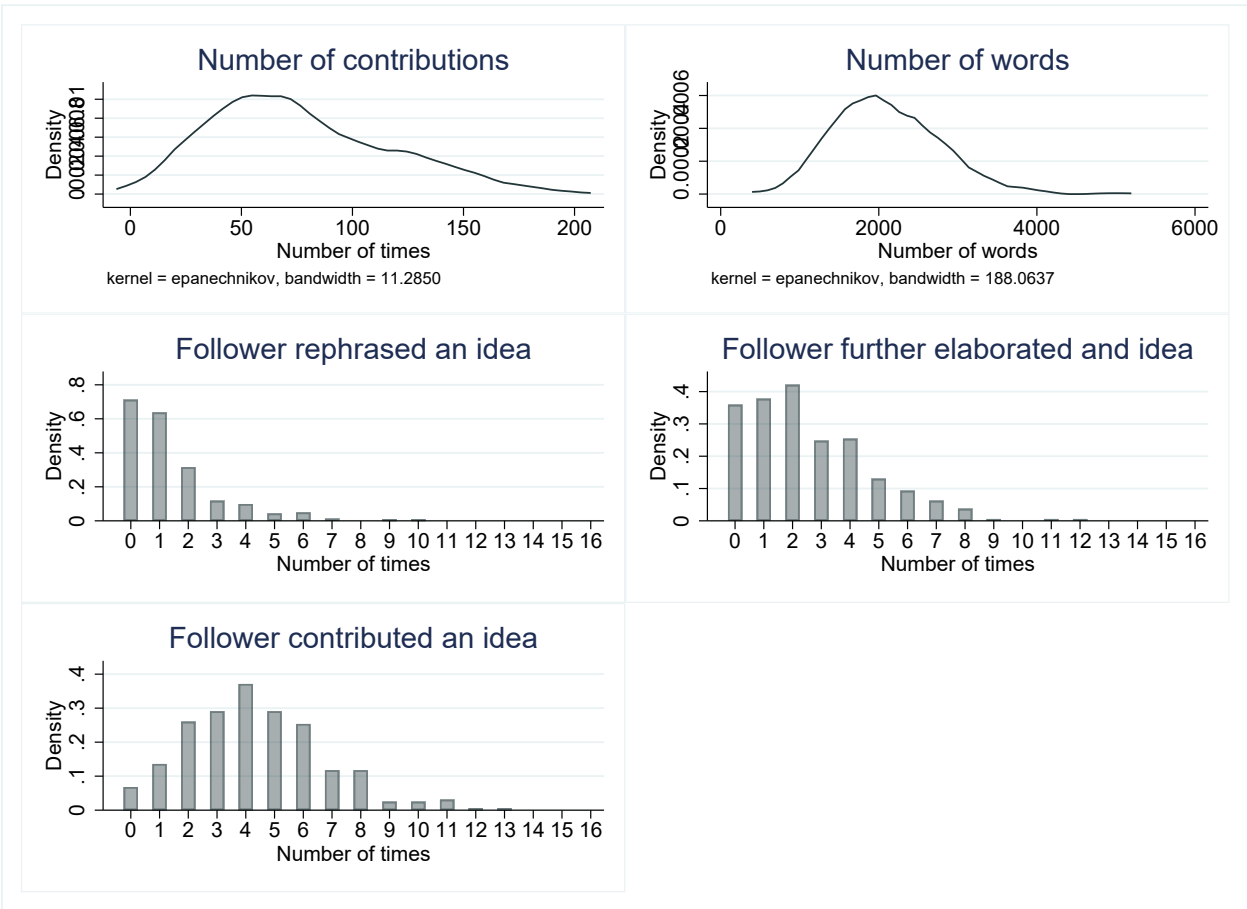


Figure A.4: Follower behavior

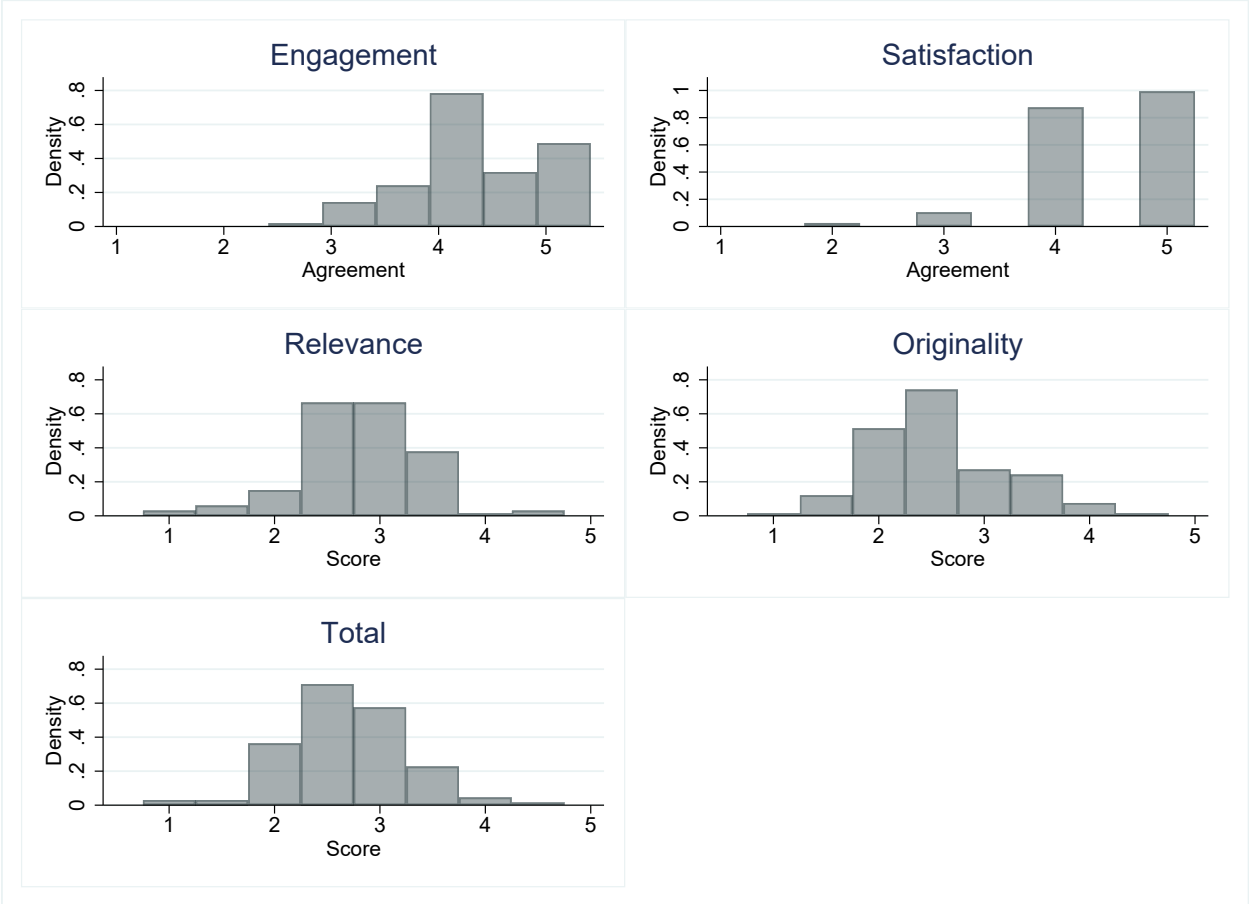


Figure A.5: Leader effectiveness

Table A.6: Differences in evaluations by follower gender

	(1)	(2)	(3)
	Supportive Behavior (follower)	Engagement Follower	Satisfaction Follower
Female leader	0.029 (0.056)	-0.079 (0.096)	0.115 (0.111)
Female follower	-0.034 (0.050)	-0.015 (0.078)	0.126 (0.084)
Female leader \times Female follower	0.053 (0.097)	0.165 (0.175)	0.041 (0.165)
Team size	-0.052* (0.032)	-0.056 (0.058)	-0.023 (0.075)
Leader tenure	-0.013 (0.020)	0.035 (0.031)	-0.027 (0.039)
Team tenure	-0.033 (0.021)	-0.055* (0.028)	-0.043 (0.035)
Leader level	-0.033 (0.021)	-0.055* (0.028)	-0.043 (0.035)
Constant	4.518*** (0.127)	4.447*** (0.235)	4.350*** (0.310)
Observations	328	302	302
Adjusted R^2	0.037	0.008	0.023

Notes: OLS regressions. Leader level is a binary indicator for high-level leaders. We also include an indicator for the leadership training. Numbers in parentheses indicate standard errors clustered by team.

B Analysis by leadership-training status

Table B.1: Subjective measures of leader behavior by training status

Panel A: Leaders who received the training					
	(1)	(2)	(3)	(4)	(5)
	Rephrase Idea	Elaborate Idea	New Idea	Supportive Behavior (fol)	Supportive Behavior (RA)
Female	0.505 (0.719)	1.141** (0.551)	1.648** (0.681)	-0.141** (0.066)	0.135 (0.512)
Constant	0.305 (2.127)	5.063*** (1.229)	6.883*** (1.539)	4.622*** (0.150)	3.781*** (1.281)
Observations	60	60	60	153	63
Adjusted R^2	0.119	0.170	0.086	0.030	-0.030
Panel B: Leaders who did not receive the training					
	(1)	(2)	(3)	(4)	(5)
	Rephrase Idea	Elaborate Idea	New Idea	Supportive Behavior (fol)	Supportive Behavior (RA)
Female	0.043 (1.157)	0.802 (0.540)	0.226 (0.756)	0.222*** (0.064)	-0.076 (0.453)
Constant	6.409** (3.183)	3.498*** (1.224)	7.936*** (1.608)	4.528*** (0.192)	4.815*** (1.125)
Observations	65	65	65	175	65
Adjusted R^2	-0.019	-0.003	0.102	0.061	-0.063

Notes: OLS regressions. Numbers in parentheses show robust standard errors. In column (1), the dependent variable is measured at the follower level and we cluster standard errors at the team level. Leader level is a binary indicator for high-level leaders. We also include controls for team size, leader and team tenure and leader level.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table B.2: Objective measures of leader behavior by training status

Panel A: Leaders who received the training								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Number of Contributions	Number of Words	Agency	Communion	Status	Authenticity	CDI	Assent
Female	13.869 (11.891)	51.021 (173.560)	-0.094 (0.231)	-0.183 (0.251)	-0.361 (0.519)	0.443 (0.382)	-0.580 (0.750)	0.179 (0.394)
Constant	145.571*** (28.287)	3,139.652*** (710.798)	6.207*** (0.514)	5.376*** (0.527)	8.064*** (1.069)	8.839*** (0.820)	4.716*** (1.535)	3.519*** (1.182)
Observations	60	60	60	60	60	60	60	60
Adjusted R^2	0.026	-0.040	-0.034	0.095	0.036	0.074	0.002	0.042
Panel B: Leaders who did not receive the training								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Number of Contributions	Number of Words	Agency	Communion	Status	Authenticity	CDI	Assent
Female	44.829*** (9.489)	74.781 (165.120)	-0.089 (0.282)	-0.297 (0.275)	-0.720 (0.727)	0.176 (0.334)	0.451 (0.842)	2.040** (0.804)
Constant	136.559*** (24.826)	3,000.681*** (430.381)	6.385*** (0.649)	4.302*** (0.667)	6.149*** (1.620)	8.639*** (1.059)	4.695** (2.120)	2.289 (1.549)
Observations	65	65	65	65	65	65	65	65
Adjusted R^2	0.225	0.066	-0.052	0.026	0.015	0.105	-0.044	0.123

Notes: OLS regressions with robust standard errors in parentheses. The dependent variables in columns (3), (4) and (8) are word counts based on dictionaries. The dependent variables in columns (5) - (7) are based on language styles. We also include controls for team size, leader and team tenure and leader level.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table B.3: Measures of follower behavior by training status

Panel A: Leaders who received the training					
	(1)	(2)	(3)	(4)	(5)
	Nr Contributions follower	Nr Words follower	Nr Rephrase follower	Nr Elaborate follower	Nr Ideas follower
Female leader	1.050	74.233	0.221	-0.020	0.271
	(6.378)	(163.287)	(0.244)	(0.364)	(0.342)
Constant	175.974***	3,092.824***	1.573**	6.065***	7.692***
	(19.321)	(534.823)	(0.638)	(0.823)	(1.114)
Observations	148	148	148	148	148
Adjusted R^2	0.292	0.011	0.026	0.096	0.083
Panel B: Leaders who did not receive the training					
	(1)	(2)	(3)	(4)	(5)
	Nr contributions follower	Nr Words follower	Nr Rephrase follower	Nr Elaborate follower	Nr Ideas follower
Female leader	22.374***	153.213	0.320	0.931*	0.135
	(7.535)	(153.612)	(0.353)	(0.479)	(0.410)
Constant	132.002***	2,870.651***	0.452	5.263***	7.609***
	(21.720)	(482.444)	(0.839)	(1.220)	(1.518)
Observations	170	170	170	170	170
Adjusted R^2	0.141	0.069	-0.019	0.080	0.014

Notes: OLS regressions. Numbers in parentheses indicate standard errors clustered at the team level. We also include controls for team size, leader and team tenure and leader level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table B.4: Output quality and effectiveness by training status

Panel A: Leaders who received the training							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Engagement Follower	Satisfaction Follower	Good Meeting (RA)	Engaging Meeting (RA)	Relevance	Originality	Total Score
Female	-0.233** (0.114)	-0.088 (0.127)	0.001 (0.057)	0.116 (0.077)	0.021 (0.168)	0.094 (0.166)	0.058 (0.155)
Constant	4.700*** (0.255)	4.843*** (0.391)	0.854*** (0.141)	0.745*** (0.204)	1.985*** (0.354)	2.052*** (0.446)	2.018*** (0.369)
Observations	140	140	63	63	63	63	63
Adjusted R^2	0.040	-0.005	0.139	0.027	0.077	0.018	0.063
Panel B: Leaders who did not receive the training							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Engagement Follower	Satisfaction Follower	Good Meeting (RA)	Engaging Meeting (RA)	Relevance	Originality	Total Score
Female	0.239*** (0.090)	0.421*** (0.101)	0.039 (0.031)	0.267** (0.108)	-0.068 (0.159)	-0.141 (0.155)	-0.105 (0.148)
Constant	4.206*** (0.327)	4.054*** (0.420)	0.930*** (0.122)	0.732** (0.318)	3.251*** (0.541)	2.819*** (0.541)	3.035*** (0.519)
Observations	162	162	65	65	67	67	67
Adjusted R^2	0.010	0.037	-0.048	0.005	0.070	0.003	0.042

Notes: Robust standard errors in parentheses. In columns (1) and (2) standard errors are clustered at the team level. We also include controls for team size, leader and team tenure and leader level. *** p<0.01, ** p<0.05, * p<0.1

C Instructions (translated from Norwegian)

Welcome!

Thank you for contributing to *companyname*'s future.

Please provide your email address, so we know which team you belong to (only the researchers will have access to this information to merge the data).

Enter your email address:

Confirm your email address

We will now explain to you how the group discussion you will have in Microsoft Teams will be structured.

It is important that you read these instructions carefully.

IMPORTANT!

We will guide you through the group discussion step by step on this page during the meeting. It is therefore important that you do not close this browser window after reading the instructions, but continue to click through the instructions during the meeting. You are not done before you have finished the short survey at the end.

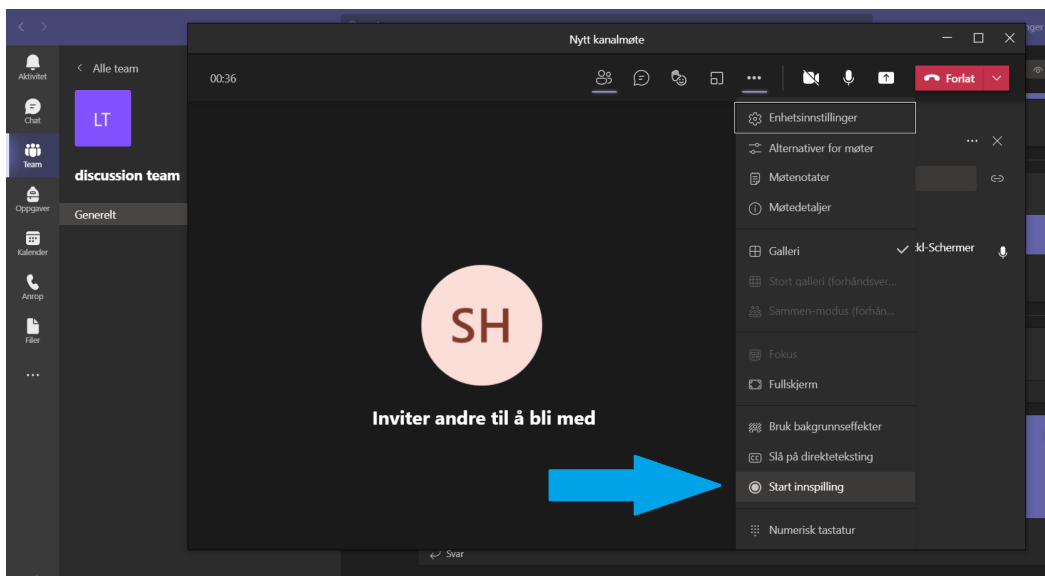
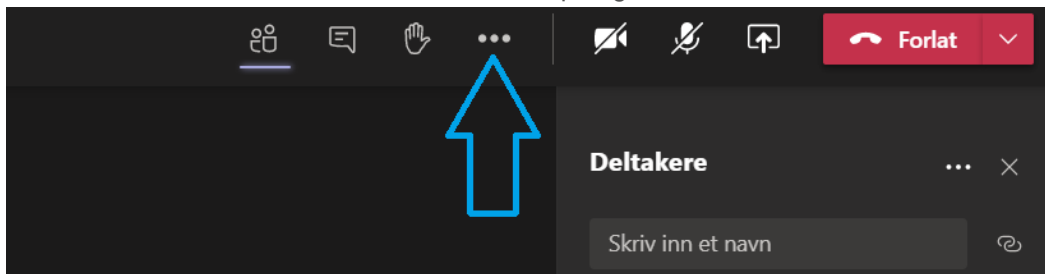
The steps of the meeting are:

- At the beginning of the group discussion: **Activate recording**
- When the meeting has started:
 - Introduce yourself to the other participants and also let them introduce themselves. (Please also make sure that they have activated their cameras.)
- Briefly introduce the agenda of the meeting
- Ice breaker task
- Group discussion:
 - Phase 1: Brainstorm
 - Phase 2: Prioritize and make recommendations
- Post discussion survey for group leaders and members

We will now provide more details for each step.

At the beginning of the group discussion: **Activate recording**. If you are working for Viken or Signal, please ask one of the other participants to start the recording for you.

Click on the three dots and then select „Start Inspilling“



If you have been successful, you will see a red dot in the upper left corner of your meeting screen.



When the meeting has started, please introduce yourself to the two other participants and also let them introduce themselves. Also, provide a short overview of the meeting agenda.

After your introduction we have a short ice breaker task prepared for you and your group. To work on this task, you have to open this browser window again.

Here is a short preview for you:

Task Description:

The ice breaker task consists of three rounds. In each round, you will be asked to think about unusual uses of an object. Please enter as many of these “uses” as your group can come up with in the survey form that you will see on this page. You have three minutes for each of the objects.

Example of the Task:

Please list as many, as different, and as unusual uses for a rubber tire as you can think of. Do not restrict yourself to a specific size of a tire. You can also list uses that require several tires. Do not restrict yourself to uses you are familiar with, but think of as many new uses as possible! After three minutes the next task will appear.

Example Answers:

Please describe the possible use in a few words if necessary. Using the example of the rubber tire: “sled” or “flower box” are clear answers, whereas “target” would require further explanation, such as “ball game with tire as target. Try to think of original answers: An answer is considered (very) original if only (very) few people think of it. Furthermore, try to think about different categories: Using the example of the rubber tire: “car tire” and “bicycle tire” belong to the category “tires as wheels” and the answer “swing seat” is a different category (category “toys”).

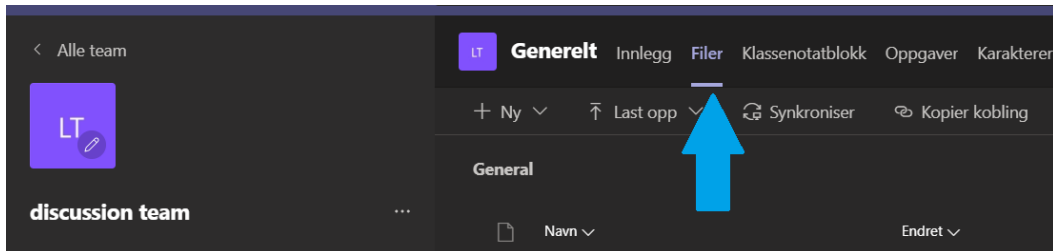
After completing the ice breaker task, please start the discussion with your group members.

During the group discussion, we would like you to discuss *companyname*'s future with a focus on a specific question, we we tell you more about when the meeting has started. As a meeting leader you are also expected to participate in the discussion.

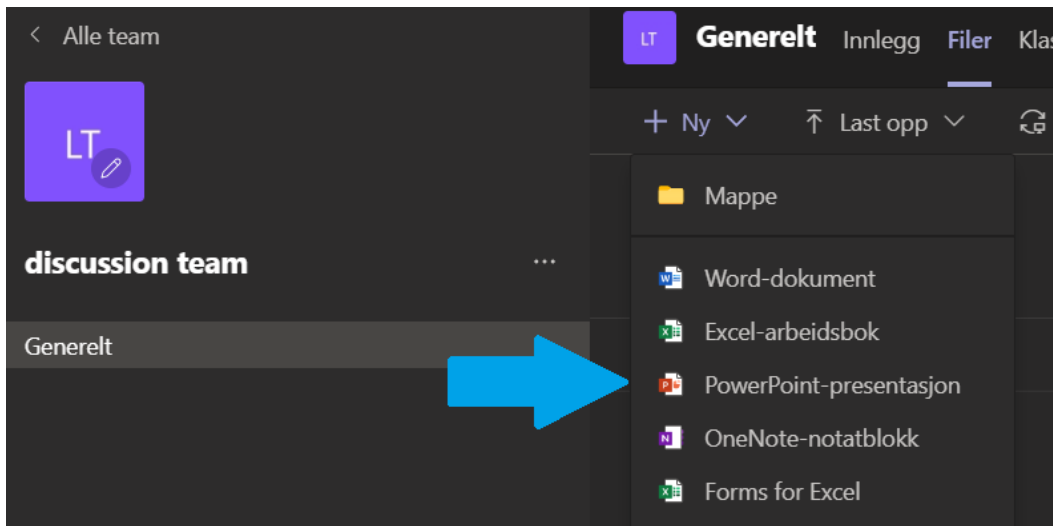
You have 30 minutes to discuss. First, please take 15 minutes to collect ideas. Please collect all your ideas on one PowerPoint slide (just a simple list is enough). After you have prepared a list of potential ideas, please take the next 15 minutes to choose the 3 ideas you find most promising. Elaborate on each of these ideas. Make a new slide for

each idea and describe it in 3 bullet points.

To open a PowerPoint slide in Microsoft Teams click on “Filer”.



Next click on “Ny” and select “Power Point-presentasjon”



Enter “Ideas Team X” with X being your Team number as the file name. Your changes to the Power Point Presentation will be saved automatically.

Before you continue, please try if you are able to open a Power Point Presentation and save it in Microsoft Teams.

The Treatment group will see the leadership training here

Thank you for completing the introduction. The meeting will start soon
Here we have summarized again the outline of the group discussion.
(Download the summary so you can access it during the meeting.)

- Start with a short round of introductions. Also, briefly explain the agenda of the meeting.
- Open your browser and do the “ice-breaking” tasks with your group.
- Explain the topic of the group discussion to your group. We will display the topic of the discussion in this browser when you reach this part.
- Open a PowerPoint Presentation in Teams.
- First 15 minutes: Brainstorm and list all ideas that your group has (on the first PowerPoint slide)
- Second 15 minutes: Pick 3 ideas and describe each of them on one PowerPoint slide. Be brief - 3 bullet points are enough.

Please note that all the information will be repeated here once your meeting has started, so you can see the instructions again step by step. (Click “Meeting has started” when the meeting has started.)

Start the meeting at the set time. We hope that everyone will show up as agreed upon. However, please start the meeting nonetheless, if someone does not show up, as long as at least one person shows up.

Good luck!

Great, your meeting has started.

Now please make sure you **activated the recording**. If you are working for Viken or Signal, please ask one of your group members to start the recording for you.

Please also check that **everyone has turned on their cameras**.

If you have forgotten how to start the recording you can look at the instructions again [here](#).

- recording is activated
- all cameras are turned on

Introduce yourself to the two other participants and also let them introduce themselves. Afterwards, briefly explain the meeting agenda.

Breaking the ice

You can now start the ice breaker task. Please read the following out loud to your group members:

“The ice breaker task consists of three rounds. In each round, you will be asked to think about unusual uses of an object. Please enter as many of these “uses” as your group can come up with in the survey form that you will see on this page. You have three minutes for each of the objects.

Example of the Task:

Please list as many, as different, and as unusual uses for a rubber tire as you can think of. Do not restrict yourself to a specific size of a tire. You can also list uses that require several tires. Do not restrict yourself to uses you are familiar with, but think of as many new uses as possible! After three minutes the next task will appear.”

If you and your group members are ready please press “**Ready**” to see the first object.

Round one:

In this round your object is **Paper**.

Please list the "uses" your group could think of in the box below

0300

Round two:

In this round your object is **Tin Can**.

Please list the "uses" your group could think of in the box below



0300

Round three:

In this round your object is **Cord**.

Please list the "uses" your group could think of in the box below



0300

Great!

You have now completed the ice breaker task.

Now you can start the discussion about *companyname*'s future.

Please read the following instructions out loud:

“*companyname* has an ambitious goal in its Strategy 2030 that says that we want to be among the top ten employers. We would like you to discuss what *companyname* can do to become an even more attractive workplace. The meeting leader is also expected to participate in the discussion.

You have 30 minutes to discuss. Please take 15 minutes to brainstorm and collect ideas. Write down all ideas in one PowerPoint slide (a simple list is enough). After you have prepared a list of potential ideas, take the next 15 minutes to choose the 3 ideas you find most promising. Elaborate on each of these ideas. Make a new PowerPoint slide for each idea and describe it in three bullet points..”

If you are done, please press “**Start discussion**”

Discussion

What can *companyname* do to become an even more attractive workplace?

Please use the first 15 minutes to brainstorm ideas. Collect your ideas on the first slide of your PowerPoint Presentation. The researchers will not forward this first slide to *companyname*'s strategy discussion.

1 5 0 0

Please use the next 15 minutes to choose the 3 best ideas and generate a PowerPoint slide for each idea. Describe each idea in three bullet points. These slides will be used in *companyname*'s strategy discussion.

IMPORTANT: Remember to come back to this website after you are done with the discussion and the PowerPoint presentation, so you can receive the link to the final survey for leaders and employees.

1 5 0 0

You have now finished the group discussion. We kindly ask each of you to take 5 minutes to respond to the survey.

Please share the following link with your group members by posting it in the chat in Teams (NOTICE: You as a leader, should not follow the same link).

[link](#)

The link leads to a short survey to provide feedback on the group discussion. You can see the survey for leaders if you press continue.

Powered by Qualtrics