Organizational Behavior and Human Decision Processes 160 (2020) 115–130

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

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

Gender, power and emotions in the collaborative production of knowledge: A large-scale analysis of Wikipedia editor conversations☆

Jana Gallus[a,1,*], Sudeep Bhatia[b,1]

[a] UCLA, Anderson School of Management, 110 Westwood Plaza, Los Angeles, CA 90095, United States
[b] University of Pennsylvania, Department of Psychology, Solomon Labs, 3720 Walnut Street, Philadelphia, PA 19104-6018, United States

ARTICLE INFO

Keywords:
Conversations
Gender
Power
Emotionality
Wikipedia

ABSTRACT

This paper studies the conversations behind the operations of a large-scale, online knowledge production community: Wikipedia. We investigate gender differences in the conversational styles (emotionality) and conversational domain choices (controversiality and gender stereotypicality of content) among contributors, and how these differences change as we look up the organizational hierarchy. In the general population of contributors, we expect and find significant gender differences, whereby comments and statements from women are higher-valenced, have more affective content, and are in domains that are less controversial and more female-typed. Importantly, these differences diminish or disappear among people in positions of power: female authorities converge to the behavior of their male counterparts, such that the gender gaps in valence and willingness to converse on controversial content disappear. We find greater sorting into topics according to their gender stereotypicality. We discuss mechanisms and implications for research on gender differences, leadership behavior, and conversational phenomena arising from such large-scale forms of knowledge production.

1. Introduction

Collaborative work would be unthinkable absent people's ability to converse in order to share information and to coordinate and motivate efforts. Conversations influence work, for instance through their effects on productivity and creativity (Huang, Gino, & Galinsky, 2015; Wu, Waber, Aral, Brynjolfsson, & Pentland, 2008). At the same time, conversations are also shaped by work processes. Expressions of emotions in natural collaborative production processes offer an important window into the psychology of work. They can inform our understanding of the differential motivations and experiences of various subgroups of workers, and how their presence might influence the broader organizational climate and culture (Cross & Madson, 1997; Schein, 2004). Women in positions of power are one important subgroup of workers on which our knowledge is still limited, largely due to the unavailability of data.

Research has shown that men and women in the general population differ in their choices (Kugler, Reif, Kaschner, & Brodbeck, 2018), preferences (see Croson & Gneezy, 2009), and personality traits (Costa Jr., Terracciano, & McCrae, 2001; Feingold, 1994), but little is known about gender differences higher up in organizational hierarchies (Adams & Funk, 2012) and how they compare to gender differences at lower levels of the same organization. Our paper addresses this gap by observing conversations between individuals who jointly and voluntarily work on one of the largest knowledge production platforms, Wikipedia. Specifically, we address the following questions: Are there systematic differences in the expression of emotions by women and men, and in their choice of conversational topics in terms of domain gender stereotype and topic controversiality? Do possible gender differences persist as we shift our perspective to people in positions of authority? Are they amplified, or are they attenuated?

The responses to these questions are important from a gender perspective because they speak to the more fundamental question of whether some of the main effects of gender that have been found in previous research might result from confounding gender with underlying power differentials (see, e.g., Johnson & Helgeson, 2002). The answers matter from an organizational perspective because they allow us to shed further light on how conversations and verbal interactions are related to emotional experience and motivation in organizations (Herring, 2000; Pennebaker, Mehl, & Niederhoffer, 2003). We put a

☆ This article is part of the special issue "The Psychology of Conversation," Edited by Dr. Alison Wood Brooks."

* Corresponding author.
E-mail addresses: jana.gallus@anderson.ucla.edu (J. Gallus), bhatiasu@sas.upenn.edu (S. Bhatia).

[1] The authors contributed equally. They wish to thank Alan P. Fiske, Myriam Benzarti, as well as the reviewing team and editors for their most helpful input and ideas.

https://doi.org/10.1016/j.obhdp.2020.02.003
Received 17 March 2019; Received in revised form 17 February 2020; Accepted 24 February 2020
Available online 01 May 2020

special emphasis on the role of an important and growing subgroup of workers: women who advance to the core of collaborative knowledge production processes. While in the early days of computer-mediated work, there had been hopes that gender anonymity or at least the reduction of gender cues might reduce or eliminate gender differences, research suggests that gendered power differentials often carry over into virtual online contexts, and that women and girls are more likely than their male counterparts to experience dissatisfaction and other adverse consequences (Guiller & Durndell, 2007; 2000; Lee, 2007; Prinsen, Volman, Terwel, 2007). It is therefore interesting to analyze whether gender differences found in non-virtual contexts also obtain in our online context, and how they relate to power differences. Finally, answers to the above questions can lay the ground for developing interventions to address persistent gender discrepancies in organizations and online communities (Bohnet, 2016).

It is difficult to observe natural conversations and topic selection as they occur at work without being invasive and potentially distorting people's behavior. The problem is compounded if one's interest lies in gender differences across different hierarchy levels due to the lack of observations on women in positions of power. We address this problem by utilizing a large-scale, publicly available online dataset of conversations between Wikipedia contributors (also called editors or users).

Wikipedia is a prime example of a large-scale online production system where millions of contributors voluntarily establish and curate a global public knowledge good (Benkler, 2006; Gallus, 2017; Lih, 2009; Zhang & Zhu, 2011). At present, it appears to epitomize what is widely seen as a novel form of organizing, which is increasingly garnering attention by scholars in diverse fields, such as organization, economics, innovation, and strategy (e.g., Faraj, Jarvenpaa, & Majchrzak, 2011; Klapper & Reitzig, 2018; Lakhani & Von Hippel, 2003; Lerner, Pathak, & Tirole, 2006; Levine & Prietula, 2014): Wikipedia is built on principles of open collaboration, designed to reduce social hierarchies and increase decentralization among decision-makers, and it is fueled by a wide range of motivations, going well beyond "standard economic incentives" such as money and career concerns. This organization has given rise to the most comprehensive encyclopedia in history, with more than 51 million articles in over 290 languages, and an information resource that has continuously ranked among the top 10 most popular websites worldwide.[2] The increasing importance of this mode of producing and innovating (von Hippel, 2017), and the success of Wikipedia specifically, make understanding its organization important in its own right (Gallus, 2017).

But even in traditional sectors, notably in the sciences and knowledge economy, more and more work is being conducted in teams (Lazear & Shaw, 2007; Wuchty, Jones, & Uzzi, 2007), and there has been an increasing predominance of large teams in particular (Wu, Wang, & Evans, 2019). Wikipedia allows us to study the inner workings of such large-scale collaboration. It also provides a glimpse into alternative work arrangements built on online intermediaries (Katz & Krueger, 2019) where, in the extreme, collaboration takes place exclusively in a virtual space without face-to-face contact among organizational members.

At the same time, Wikipedia also bears some interesting similarities with more traditional forms of organization: it has endogenously evolved hierarchical structures, formal rules and processes (e.g., for onboarding and mentoring newcomers), and it allows a separation of work in terms of content (with groups such as a "Statistics Department" and so-called WikiProjects that organize work around specific topics) and functions (e.g., content production, policy work, administrative maintenance). All of the production planning and quality management of the encyclopedia's articles takes place on "Wikipedia Talk pages", in the form of discussions among Wikipedia contributors. This allows us to study a rich trove of data, which covers a period of more than 15 years and contains 166,322 different discussion threads across 1,236 articles/topics on Wikipedia Talk pages (Prabhakaran & Rambow, 2016). Importantly, we have information on contributors' gender as well as their roles: general editors versus so-called "administrators" with greater decision-making power (e.g., the right to block and unblock other editors' accounts, to restrict or allow editing of certain Wikipedia articles, or to judge the outcome of certain discussions).[3]

Large-scale natural language datasets obtained from the Internet have proven extremely useful for understanding human behavior, with important applications in many fields, such as management (George, Osinga, Lavie, & Scott, 2016), public health (Hawn, 2009), cognitive science (Griffiths, 2015), marketing (Humphreys & Wang, 2017), and psychology (Harlow & Oswald, 2016; Kosinski & Behrend, 2017). Our use of the Wikipedia conversations dataset, along with novel techniques from natural language processing and computational linguistics, allow us to analyze differences in the expression of emotions (valence, arousal, as well as overall degree of emotionality) and how they unfold across different levels of the organizational hierarchy (normal editors versus administrators). Since Wikipedia aims to cover the sum of all human knowledge (as opposed to technical and focused communities such as StackOverflow), and since people self-select into topics of their choosing (rather than being told what to work on by managers), we can, moreover, study differences in the gender stereotype of the domain and in the controversiality of articles that different editors choose to converse about. This allows us not only to analyze gender differences in *conversational styles* (the expression of emotions) among general editors and those in positions of power, but also in their *conversational domain choice* with respect to the topic's gender stereotype and controversiality.

2. Theory

2.1. Gender differences in emotionality

Previous research shows that men and women in the general population differ systematically in terms of their preferences (Croson & Gneezy, 2009) and negotiation (Kugler et al., 2018) and linguistic behaviors (Carli, 1990; Mulac, 1998). With regards to emotionality, women have been found to use references to emotion (e.g., "I am happy") more frequently than men (Palomares, 2004), and to make more emotional and positive contributions in asynchronous computer-mediated communications (Guiller & Durndell, 2007). Although a large number of studies have drawn their observations from university students, it is possible to predict a communicator's gender with high accuracy from observing their language use (see, e.g., Mulac (1998) and more recent advances such as Schwartz et al. (2013)). Popular accounts such as Tannen's (1990) *You Just Don't Understand: Women and Men in Conversation* (a *New York Times* bestseller) even argue that men and women belong to different linguistic communities with stark differences in their conversational styles. But again, most observations stem from observing women from the general population, where power may be a confounding factor.

Such differences in emotionality may at least in part be explained by society's gender role beliefs (Eagly & Wood, 2012), or gender stereotypes, which lead to expectations for women to be communal (i.e., warm, emotional, supportive and caring) as opposed to agentic and dominant (e.g., Amanatullah & Tinsley, 2013; Eagly, 1987; Eagly & Carli, 2003; Williams & Tiedens, 2016). Gender role beliefs impact

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand_Total for Wikipedia-specific statistics, https://en.wikipedia.org/wiki/List_of_most_popular_websites for the ranking of websites in terms of web traffic.

[3] Adminship is an official status that is granted as a result of a community discussion and election in which any registered editor can participate. See https://en.wikipedia.org/wiki/Wikipedia:Administrators#Becoming_an_administrator.

individuals' behavior through various mechanisms (Wood & Eagly, 2010). One important mechanism is social sanctions for counter-stereotypical behavior, also termed the backlash effect (Rudman, 1998; Rudman & Fairchild, 2004). Thus, women may refrain from displaying leadership behaviors and using the concomitant language in order to avoid negative evaluations due to the perceived gender-leadership role incongruity (Eagly & Karau, 2002). In many cultural contexts, gender-specific norms make it appropriate for women but not for men to express positive emotions (Brody, 2000).

Even absent others' knowledge of an individual's gender, such as in many online contexts, gender role beliefs can produce gender differences in behavior through internalization of a given gender identity (Wood & Eagly, 2015). It is therefore an interesting question what happens when we consider modern knowledge production contexts, where gender cues are much less salient because individuals work in large-scale online communities and are not co-located. Despite the reduced prominence of gender cues in these contexts, past empirical research (Kucuktunc, Cambazoglu, Weber, & Ferhatosmanoglu, 2012; Laniado, Kaltenbrunner, Castillo, & Morell, 2012) as well as the gender identity mechanism (Wood & Eagly, 2015) suggest that we can expect to find among the general population of editors gender differences in emotionality that are similar to those in offline contexts.

2.2. Gender differences in domain choice

A well-established research stream following Gneezy, Niederle, and Rustichini (2003) and Niederle and Vesterlund (2007) in economics shows that women shy away from competition and conflict (see, e.g., Bear, Weingart, & Todorova, 2014; Schneider, Holman, Diekman, & McAndrew, 2016 as well as earlier work by Stuhlmacher & Walters, 1999). Following this research, we would expect female editors to be less likely to engage in conversations about controversial topics. Another explanation for observing such a relationship between gender and controversiality would be that particular topics come to be regarded as controversial specifically because men gravitate towards discussing them. Research in personality psychology suggests that men on average are somewhat less 'agreeable' than women (e.g., McCrae, Terracciano, & 78 Members of the Personality Profiles of Cultures Project, 2005). Therefore, a high concentration of men might lead to a different tone in the discussion, such that the article becomes labeled as controversial. We will not be able to rule out such an interpretation, but consider the alternative more plausible: that women in the general population of editors are more reluctant to engage in controversial content discussions. Recent survey evidence on Wikipedia editors indeed suggests that female editors display greater avoidance of conflict than men (Bear & Collier, 2016).

Similarly, albeit focused on non-work contexts, it has been suggested that men and women in the general population differ in their choice of conversation topics (Bischoping, 1993). Moreover, gender-incongruent situations may lead to increased anxiety, role conflict, backlash, and avoidance (Bem & Lenney, 1976; Luhaorg & Zivian, 1995; Rudman, 1998). We therefore expect to find a gender specific separation of labor, whereby female editors from the general population are more likely to converse on female-typed content, while male editors are more likely to converse on male-typed content. Such domain-specific sorting by gender should be reinforced by differences in previously accumulated expertise (e.g., somebody with expertise in arts will be more likely to contribute to articles related to the arts). If this is the case, we may expect the same domain-specific gender difference to persist as we consider editors in positions of power. This would contrast with the previously discussed gender differences in emotionality and article controversiality, as further discussed below.

2.3. The gender gap across the organizational hierarchy

Understanding whether systematic differences between men and women persist as we look up the organizational hierarchy is important because it speaks to whether gender differences found in the general population are absolute, or whether they may have been partly confounded with related differences in status and power (Johnson & Helgeson, 2002; Watson, 1994). Moreover, from a practical perspective, analyzing gender differences at the top of organizational hierarchies advances our understanding of the implications of increased female participation in organizational leadership (Adams & Funk, 2012). Differences in the expression of emotions and in the domain choices made by men and women in power have implications for the broader organizational culture (e.g., through the expression of emotions) and functioning (e.g., if female leaders were to avoid controversy).

2.3.1. Emotionality

Prior research suggests that the differences in male and female leaders' styles are merely "mild shading" and that general similarities in style prevail (see Gipson, Pfaff, Mendelsohn, Catenacci, & Burke, 2017 for a recent survey of the literature). Moreover, there appear to be no significant differences between female and male leaders' demonstrations of emotional intelligence competencies (Hopkins & Bilimoria, 2008). Elevated power has been found to be associated with increased freedom and more socially disinhibited behavior (Keltner, Gruenfeld, & Anderson, 2003). Thus, women in positions of authority may be less bound by the female gender role. We therefore expect to find smaller differences in the expression of emotions (valence, arousal) by women and men in positions of power, compared to the differences in the general population of editors.

Extending the analysis to the expression of mental states beyond emotions and, specifically, considering the extent to which reference is made to cognitive as opposed to affective processes, we do not have a prediction about whether female leaders would behave differently from women in the general population of editors. But in a context of online knowledge work, where at baseline comments can be expected to be significantly more cognitively-loaded and hence less affect-based, we do expect a pattern whereby comments from women are generally more affective than comments from men. Whether women's reference to cognitive processes changes as they come into positions of power is an empirical question that we will analyze.

2.3.2. Domain choice

Using a survey of directors, Adams and Funk (2012) show that several of the well-established gender differences found in the general population no longer hold or are even reversed when looking at female and male directors. Notably, female directors in their sample are more risk tolerant and less security- and tradition-oriented than their male counterparts. Translated to our context, this suggests that women in positions of power may be more likely than women in the general population to engage in conversations about controversial content, such that the gender gap may disappear. However, to the extent that women have greater knowledge of stereotypically female content, the gender gap in topic choice (male- vs. female-typed) may remain.

Hence, overall, we expect to find smaller or no gender differences in the expression of emotions (valence, arousal) and the choice of engaging in controversial content discussions. We conjecture that this will be driven by women converging to the behavior of their male counterparts as they come to occupy positions of authority. An intriguing question also for future research is what accounts for any potential closing of the gender gaps.

2.3.3. Mechanisms

There are three non-exclusive mechanisms why gender differences may disappear when considering men and women in power: first, a

treatment effect of the position of authority on behavior and possibly preferences (see Magee and Galinsky (2008) for a review of the effects of power on individuals' psychological states and behavior). This would suggest that the position of authority mutes gender differences. Putting women in positions of authority allows or compels them to express less positive emotions and to engage more in controversial content discussions. As the experience of power makes individuals more goal-directed and more likely to take action (Galinsky, Gruenfeld, & Magee, 2003), they may devote less attention to other dimensions, such as conforming with their gender role. Power has been found to make individuals less likely to consider others' perspectives (Galinsky, Magee, Inesi, & Gruenfeld, 2006), which may also reduce women's awareness of (or concern about) social expectancies related to their gender role.

Besides this explanation of a treatment effect of the position of authority, we consider two forms of sorting, whereby female editors who display a stereotypically male emotional tone or choice of domain are more likely to seek or find themselves in positions of power. Hence, the second mechanism is self-selection (in line with occupational sorting à la Polachek (1981)). This is a supply-side factor or, as referred to by psychologists, an intrapersonal effect (Gino, Wilmuth, & Brooks, 2015) in that it takes place within ("intra") the individual and reflects the person's own decisions. For instance, recent research shows that women see professional advancement as less desirable (Gino et al., 2015), and that they seem to be less status-seeking than men (Huberman, Loch, and Önçüler (2004)).[4] In short, it is possible that women who seek to advance to positions of authority systematically differ from women in the general population.

The third mechanism is social selection by the majority-male population of editors.[5] This would correspond to demand-side factors, or interpersonal effects, which may or may not be conscious. In contrast to intrapersonal effects, interpersonal effects take place at the intersection between ("inter") individuals and are taken to refer to others' decisions. This explanation is in line with the argument that women must act like men to climb organizational hierarchies and be successful (Branson, 2006). In virtual collaboration contexts such as Wikipedia, gender cues are less salient than in processes where physical characteristics are apparent (e.g., Brooks, Huang, Kearney, & Murray, 2014; Goldin & Rouse, 2000). Nevertheless, research has shown that gender differences often persist in computer-mediated contexts with gender anonymity (Guiller & Durndell, 2007; Herring, 2000; Lee, 2007), and that there continues to be greater conformity to ostensible male interaction partners even where linguistic features are used as bases for gender inferences (Lee, 2007). Thus, social selection may also occur based on behavioral differences, such that women who act more like men are more likely to be accorded higher-status positions.

A recent analysis by Fernandez-Mateo and Fernandez (2016) discusses the intricacies of distinguishing demand- and supply-side factors (social- and self-selection, respectively), including how anticipatory effects can make demand-side factors (e.g., a discriminatory environment) look like supply-side preferences on the part of women who select out to preempt being discriminated against. The authors propose an original approach for untangling the mechanisms in the context of executive search. Distinguishing these two sorting mechanisms from a treatment effect adds an additional layer of complication and is beyond the scope of the present paper. Yet investigating the past behavior of female editors who eventually rise to positions of authority will yield some insight as to the relevance of the different mechanisms. If the two forms of sorting are sufficient to explain a possible closing of the gender gap among editors in power, we would expect to see that women who rise to the top already differed from the general population before their ascent. We will present analyses in the Results section.

3. Methods

3.1. Dataset

Our dataset involves Wikipedia Talk page discussions collected by Prabhakaran and Rambow (2016) and made available at https://www.cs.stanford.edu/~vinod/publication.html. These discussions contain 906,671 comments made by 104,982 unique Wikipedia editors in 166,322 threads spanning 1,236 articles from 2001 to 2015. There are an average (mean) of 5.45 comments per thread and an average of 2.84 editors per thread, and comments have an average of 85.25 words. As shown in Prabhakaran and Rambow (2016), these comments are not distributed uniformly over time. The number of comments in a given year rises from 2001 to 2006 and drops from 2007 to 2015, with 2005 to 2008 being the peak of editor interactions in the corpus. Note that about 5% of the comments do not have an assigned date (as these are from a period when Wikipedia did not enforce formats for editor signatures in comments).

Crucially, this dataset contains editor metadata obtained through the MediaWiki API, which includes whether the editor is registered, whether the editor is an administrator at the time of the post, as well as editor gender and number of edits made. 57% of editors in this dataset are registered, out of which 12% reveal their gender in their user accounts. 92% and 8% of the gender-identifiable editors are male and female, respectively, and the bulk of our analysis pertains to the comments made by these editors (in the discussion section we consider the limitations of this restriction). The participation of male and female editors is relatively stable over time (i.e., there is no statistical relationship between the date at which a comment is posted and the gender of the poster). Additionally, around 1% of the editors are administrators, and the average number of prior edits at the time of the comment (a measure of experience) of all editors for whom we have edit data is 4,428. We summarize these and other variables relevant to our analysis in the *Variables* section below. Additional details regarding the dataset are presented in Prabhakaran and Rambow (2016).

3.2. Measuring emotion

3.2.1. Emotion ratings

We examine the emotionality of editors' comments by using automated text analysis. For this, we rely on valence and arousal norms collected by Warriner, Kuperman, and Brysbaert (2013), in which valence corresponds to the overall positive or negative qualities of the word, and arousal corresponds to the degree to which the word connotes excitement, intensity, and activation. Warriner et al. collected these norms using surveys among individuals who self-identified as being current residents of the US, aged between 16 and 87 years, about 60% of whom were female. Participants were asked to rate the valence and arousal of words on a scale from 1 to 9 (with higher ratings for higher valence or higher arousal). The highest valence words in this dataset are *vacation* and *happiness* (average ratings of 8.53 and 8.48, respectively), the lowest valence words are *pedophile* and *rapist* (average ratings of 1.26 and 1.30, respectively). The highest arousal words are *insanity* and *gun* (average ratings of 7.79 and 7.74, respectively), and the lowest arousal words are *grain* and *dull* (average ratings of 1.60 and 1.67, respectively). The split-half reliabilities for valence and arousal ratings are 0.91 and 0.69, respectively. This lexicon is balanced for the valence ratings: 55% of words are rated at or above 5 (positive) and 45% of words are rated below 5 (negative). It is not balanced for the arousal ratings: only 18% of words are rated at or above 5 (high arousal), whereas 82% are rated below 5 (low arousal). This likely

[4] Although there are also studies suggesting that the desire for status is universal (Anderson, Hildreth, and Howland (2015)).

[5] The most recent survey conducted by the Wikimedia Foundation puts the fraction of female editors on Wikimedia projects at 9% (Wikimedia, 2018), which corresponds closely to earlier surveys (Glott, Schmidt, & Ghosh, 2010). Readership rates, however, seem to be equal across genders (Zickuhr & Rainie, 2011).

reflects psycholinguistic features of the English language (most words are non-arousing) rather than a bias in Warriner et al.'s lexicon.

There are many measures of emotion. We limit our analyses to valence and arousal as these two dimensions capture the majority of the variance in the structure of emotional experience (Russell, 1980). We complement this with an examination of the degree to which comments by men and women are affective vs. cognitive, which extends our analysis to other mental states beyond emotions (see the *Affective* vs. *cognitive content* subsection below). Although there are other datasets that could be used to obtain valence and arousal ratings for words (Bradley & Lang, 1999), the Warriner et al. dataset is the largest lexicon currently in existence, and it contains participant-generated valence and arousal ratings for 13,915 different words. Importantly, this lexicon has been compiled by psychologists and is widely used in psychological research on emotion, language, memory, and decision making.

3.2.2. Word-frequency averaging method

We use two different methods for analyzing emotions in the Wikipedia comments. Our first method, the word-frequency averaging (WFA) method, measures the valence or arousal of a comment based simply on the aggregate valence or arousal of its component words. Specifically, this method first tokenizes the comment by lower-casing it, removing all punctuation, and splitting up the comment by white space. This step transforms the natural language sentence or paragraph that makes up the comment into a "bag-of-words" representation, i.e., a set of component words and their corresponding frequencies in the comment. After this step, the WFA method queries the Warriner et al. lexicon for the valence and arousal ratings of each word. Finally, it averages the valence or arousal ratings for all words in the comment that are also contained in the Warriner et al. lexicon, to obtain an aggregate measure of the valence or arousal of the comment (see, e.g., Humphreys and Wang (2017) for an overview of this approach).

More formally, the average valence or arousal rating of comment i using this method is:

$R_i = \sum_{j=1}^{N} f_{ij} \bullet r_j / \sum_{j=1}^{N} f_{ij}$. Here $j = 1, 2, \ldots N$ indexes the words in the Warriner et al. lexicon, f_{ij} is the frequency of occurrence of word j in comment i, r_j is the valence or arousal rating of word j in the Warriner et al. lexicon, and $N = 13{,}915$ is the total number of words in the lexicon. We have $f_{ij} = 0$ if word j is not present in comment i.

3.2.3. Embeddings method

One limitation of the WFA method is that not every comment contains words in the Warriner et al. lexicon, and there are many words not in the Warriner et al. lexicon that are commonly mentioned in the comments. To avoid these data sparsity issues and to ensure the robustness of our results, we also analyze the valence and arousal of comments using a second approach: the word embeddings method.

Word embeddings are popular tools in computational linguistics that quantify the meanings of words by describing them as high-dimensional vectors. Word vectors are derived from the structure of word co-occurrence in natural language, and are useful for a variety of text analysis applications (see Bhatia, Richie, and Zou (2019) or Lenci (2018) for overviews of word embeddings and a discussion of their relevance for psychological research). Here, we use word embeddings to extrapolate the valance and arousal ratings collected by Warriner et al. to other words mentioned in the Wikipedia comments that are not in the lexicon. Our analysis is based on the Google News word2vec embeddings model, a powerful pretrained model that possesses 300 dimensional representations for over 3 million words and phrases (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). We use the word2vec model to map each word in the Warriner et al. lexicon to a 300-dimensional embedding space, and then use the valence and arousal ratings for the words in the Warriner et al. lexicon to train an algorithm capable of predicting the valence or arousal of the 3 million words and phrases that exist in the word2vec model's space. Through this approach, we can measure the valence and arousal of a large number of words in the Wikipedia comments, including words used in the comments not present in the Warriner et al. lexicon (see Hollis, Westbury, and Lefsrud (2017) and Sedoc, Preoţiuc-Pietro, and Ungar (2017) for overviews of this method).

More specifically, each word $j = 1, 2 \ldots N$ in the Warriner et al. lexicon can be described as a vector $\boldsymbol{w_j}$ in the word2vec embedding space. Based on the valence and arousal ratings r_j associated with each word, we can learn a function g that best maps the vector $\boldsymbol{w_j}$ onto r_j. In our analysis we assume this function is linear, and we train the weights of this function using a ridge regression implemented in the sci-kit learn Python machine learning library. We optimize the regularization hyperparameter through cross-validation and find that the best performing ridge regression model achieves an out-of-sample correlation rate of 0.79 for predicting word valence and 0.61 for predicting word arousal in a ten-fold cross-validation exercise on the Warriner et al. data.

With the best-fit function g (now trained on the entire data), we can take an arbitrary vector $\boldsymbol{w}$ to make a valence or arousal rating prediction $R = g(\boldsymbol{w})$ for the word or comment corresponding to the vector. Thus, to obtain valence and arousal ratings for a given comment i in the Wikipedia corpus, we first tokenize the comment (by lower-casing and splitting by white space, as above), then vectorize the comment (by averaging the word vectors for each word in the comment), and then pass the comment vector through our trained function g. The predicted rating R_i is subsequently given by $R_i = g(\boldsymbol{w_i})$. Here, $\boldsymbol{w_i} = \sum_{k=1}^{M} f_{ik} \bullet \boldsymbol{w_k} / \sum_{j=1}^{N} f_{ik}$, and $k = 1, 2 \ldots M$ indexes the words in the word2vec vocabulary, f_{ik} is the frequency of occurrence of word k in comment i, $\boldsymbol{w_k}$ is the vector representation of word k in the word2vec model, and $M = 3{,}000{,}000$ is the total number of words in the word2vec model. We have $f_{ik} = 0$ if word k is not present in comment i. Note that we always normalize our vectors before passing them through the function g. Thus, we have $||\boldsymbol{w}|| = 1$ for all vectors used in the training and extrapolation parts of our analysis. Also note that this approach first aggregates the vectors of the words in a comment into an overall comment vector, before mapping the vector onto the valence or arousal rating scale. This is in contrast to using the vector for each word in the comment to obtain a predicted valence or arousal rating for the word, and then averaging the predicted ratings for all words in the comment as in the WFA method. The two approaches are nearly identical as our function g is linear (g applied to an average of a set of vectors is the same as the average of g applied to the individual vectors).

3.2.4. Affective vs. Cognitive content

In addition to the emotion ratings obtained using the WFA and embeddings methods described above, we also analyze the degree to which each comment expresses affective vs. cognitive content. This is done to account for the fact that not all mental states are emotions, and that they also comprise expressions of thinking, planning, and decision-making. Affective content corresponds to the use of emotion-related words and concepts (e.g., "I feel"), whereas cognitive content involves the use of belief-related words and concepts (e.g., "I think"). Measuring the relative amount of affective vs. cognitive content allows us to assess the overall degree of emotionality – or instead, emphasis on cognitive processes – of a comment.

Our analysis of affective vs. cognitive content relies on the distributed dictionary method (Garten et al., 2018) applied to the set of cognitive and affective process words in the linguistic inquiry and word count (LIWC) lexicon (Pennebaker, Boyd, Jordan, & Blackburn, 2015). As with the embeddings method for measuring valence and arousal, the distributed dictionary method uses a word embeddings representation of a comment to measure the semantic distance (i.e., dissimilarity in meaning) between the comment and a given construct characterized by a set of words. Since word embeddings quantify word meaning, comments that are closely related to the construct being analyzed will have vectors that are closer to the vectors of the words describing the

construct, whereas comments that are unrelated to the construct being analyzed will have vectors that are further away from the vectors of the words describing the construct. In our case, the constructs being analyzed are cognition and affect, and the words describing these constructs are words typically used to express the outcomes of thought and emotion processes. There are a total of 791 words in the LIWC cognitive processes word set and 1,391 words in the LIWC affective processes word set. The LIWC lexicon is a well-validated dataset of words reflecting a variety of psychological constructs, and it has been used in numerous text analysis applications in psychology (Pennebaker et al., 2015 for a summary of reliability statistics and an overview of applications).

Formally, we calculate the degree of affective vs. cognitive content in each comment by first tokenizing and vectorizing the comment with the word2vec embeddings model (as in the embeddings method described above). This gives us, for each comment i, a 300-dimensional vector $\boldsymbol{w_i}$. We also obtain vector representations for affective and cognitive processes by vectorizing and averaging each of the words in the LIWC lexicon. Specifically, we use the word2vec model to obtain word vectors $\boldsymbol{w_k}$ for each of the LIWC affective and cognitive processes words, and then average the vectors for the affective processes word set and the vectors for the cognitive processes word set to obtain a single vector representation $\boldsymbol{a}$ for affective processes and a single vector representation $\boldsymbol{c}$ for cognitive processes. Finally, for each comment i we calculate the relative semantic distance between its vector $\boldsymbol{w_i}$ and the affective and cognitive vectors $\boldsymbol{a}$ and $\boldsymbol{c}$. This is done using cosine distance, so that our measurement of the affective vs. cognitive content of vector i is given by $s_i = (1 - \boldsymbol{w_i} \cdot \boldsymbol{a} \,/\, ||\boldsymbol{w_i}|| \cdot ||\boldsymbol{a}||) - (1 - \boldsymbol{w_i} \cdot \boldsymbol{c} \,/\, ||\boldsymbol{w_i}|| \cdot ||\boldsymbol{c}||)$. As the measure of cosine distance used to compute the affective and cognitive content of the comment lies in the range $[-1,1]$, the difference in affective vs. cognitive content lies in the range $[-2,2]$. Comments with higher values of s_i have words that are more semantically related to cognitive process words than affective process words. The opposite is true for comments with lower values of s_i.

3.3. Variables

The richness or the Wikipedia dataset and the range of text analysis methods introduced above allow us to analyze the relationships between a large number of variables. Most of our analyses study the emotional characteristic of editor comments, and thus this section will summarize the statistics of our variables on the comment level. Detailed statistics for the variables discussed here are provided in Table 1.

3.3.1. Comment-level variables

As discussed in the *Dataset* section above, the Wikipedia discussions corpus contains a total of 906,671 comments. Out of these, 98.05% contain words present in the Warriner et al. lexicon. For these comments, we use the word-frequency averaging (WFA) method to calculate valence and arousal. The average WFA valence in this set of comments is 5.65 and the average WFA arousal is 3.91. Now, the main benefit of our embeddings method is that it does not suffer from WFA's data sparsity issues. Specifically, we can use the word2vec embeddings model to obtain representations for 99.41% of comments, and subsequently use our embeddings method to calculate the embeddings-based valence and arousal of these comments. The average embeddings valence in this set of comments is 5.39 and the average embeddings arousal is 4.15. The WFA and embeddings measures are positively correlated, with a correlation of 0.61 for valence and 0.44 for arousal across the comments in our dataset for which we have both measures ($p < 0.001$). Finally, we analyze the comments with embeddings representations for affective vs. cognitive content. Using the approach outlined above, we find that the average distance to the affective vs. cognitive processes words is 0.12. Thus, comments are on average more semantically related to cognitive process words than to affective process words. This is in line with Wikipedia's focus on being a knowledge creation platform, where claiming reason and objectivity would be expected to be a prevalent discussion tactic.

Other comment-level variables include the length of the comment, the order in which the comment is placed in the thread, and the date of the comment's posting. Comment length is measured as the number of words in the comment. This is not normally distributed (with most comments having few words and some comments having many words). For our analyses we will therefore use the log of the comment length, whose mean is 3.90. Comment order is simply the rank of the comment in the thread (with rank 1 for the first comment): The average comment order in our data is 9. Finally, comment date is measured as the number of days between the posting of the comment and 01/01/2000. We transform the date into a "days" variable for ease of analyzing time trends. The average time in days since 01/01/2000 for our comments is 3,269 (indicating a date in December 2008).

3.3.2. User-level variables

The dataset we use is unique in that it contains information about a subset of the editors' genders, as revealed by the editors on their user accounts. Out of the comments in the dataset, 17.97% are written by male editors, whereas 1.44% are written by female editors. The gender for 80.59% of comments cannot be determined as they are written either by non-registered editors (editors without user accounts) or registered editors who have decided not to reveal their gender. We analyze a number of other editor-level variables, such as whether or not the editor is an administrator at the time of the post and the editor's number of prior edits, which is a measure of editor experience. In our dataset, 9.09% of comments are made by administrators, and 90.91% are made by non-administrators. As mentioned in the dataset section, the average number of prior edits of all editors for whom we have edit data is 4,428. We use a log-transformation of this variable in all subsequent analyses as the number of edits is highly skewed (with most users making very few edits, and some users making a lot of edits). Over all the comments in our dataset the mean of the log number of prior edits of the commenter at the time of the comment is 8.32.

3.3.3. Article-level variables

We also consider various variables pertaining to the article being discussed. Two important variables here are whether or not the article is tagged as "controversial" on Wikipedia (31% of all articles in the dataset are controversial, though 61% of all comments are made on threads pertaining to controversial articles), as well as its gender-typedness. We compute the latter using both a WFA method and an embeddings method. The WFA gender-typedness measure of an article is obtained by calculating the ratio of the sum of male pronouns ("him", "he", "himself", "his") to the sum of male and female pronouns ("her", "she", "herself", "hers") in the article. There are a total of 1,144 unique articles that we could access, which mention at least one male or female pronoun, and these articles have an average proportion of male pronouns of 80.20% (the remaining articles either did not mention any English pronouns, or were not available on Wikipedia at the time of our analysis). There are 305 articles with exclusively male pronouns (including "God", "Walmart", "Communism", "BBC", and "American Civil War") and 12 articles with exclusively female pronouns (mostly pertaining to women's health, childbirth, and sexuality). The average gender-typedness (ratio of male pronouns to all pronouns) of articles associated with each comment is 0.84, indicating that most comments pertain to articles that are primarily male-typed.

The embeddings-based method for calculating article gender-typedness involves the type of distributed dictionary analysis (Garten et al., 2018) discussed in the *Affective* vs. *cognitive content* section above. Specifically, we first tokenize and vectorize the article using the word2vec embedding space, and then calculate the relative semantic distance to a set of 20 male words relative to a set of 20 female words. The male and female words are gender pronouns, as well as gendered relationship words (e.g. father, mother, nephew, niece), and other

Table 1
Descriptive statistics for key variables. WFA refers to variables measured using the word-frequency analysis method, whereas EMB refers to variables measured using the embeddings method. COV indicates the coverage of the variable on the data, that is the percentage of the comments over which the variable could be calculated. Variables labeled with * are typically used as dependent variables in our analysis, whereas the remainder are typically independent variables and controls.

	%	Mean	SD	COV
Comment-level variables				
Valence (WFA)*: valence of comment determined by word-frequency averaging method	–	5.65	0.47	98.05
Arousal (WFA)*: arousal of comment determined by word-frequency averaging method	–	3.91	0.31	98.05
Valence (EMB)*: valence of comment determined by embeddings method	–	5.39	0.36	99.41
Arousal (EMB)*: arousal of comment determined by embeddings method	–	4.15	0.18	99.41
Affective vs. cognitive content*: semantic cosine distance of comment to affective vs. cognitive words	–	0.12	0.03	99.41
Length*: log number of words in the comment	–	3.90	1.10	100
Order: order of comment in thread (first, second etc.)	–	9.00	11.97	100
Date: date of comment measured as days since 01/01/2000	–	3,269	1,042	95.28
User-level variables				
Male: % of comments made by users identified as male	17.97	–	–	100
Female: % of comments made by users identified as female	1.44	–	–	100
No gender: % of comments made by users without gender information	80.59	–	–	100
Administrator: % of comments made by users that are administrators	9.09	–	–	100
Not administrator: % of comments made by users that are not administrators	90.91	–	–	100
Prior edits: log number of prior edits of user at time of comment	–	8.32	2.41	85.57
Article-level variables				
Controversial: % of comments made on articles tagged as controversial	61.35	–	–	100
Non-controversial: % of comments made on articles not tagged as controversial	38.64	–	–	100
Gender-typedness (WFA): ratio of male to female pronoun counts in articles associated with comments	–	0.84	0.24	99.26
Gender-typedness (EMB): semantic cosine distance to male vs. female words of articles associated with comments	–	−0.05	0.02	99.26
Valence (WFA): valence of article associated with comments, based on word-frequency averaging method	–	5.50	0.22	99.26
Arousal (WFA): arousal of article associated with comments, based on word-frequency averaging method	–	4.11	0.13	99.26
Valence (EMB): valence of article associated with comments, based on embeddings method	–	5.12	0.30	99.26
Arousal (EMB): arousal of article associated with comments, based on embeddings method	–	4.23	0.12	99.26
Thread-level variables				
Number of comments: total number of comments in thread, averaged over comments	–	16.99	18.92	100
Number of users: number of unique editors in each thread, averaged over comments	–	5.31	3.90	100
Time difference: log days between first and last comment in thread, averaged over comments	–	1.86	1.87	72.72

Note: The statistics are aggregated on the comment-level. Thus, for example, the table presents the average gender-typedness of the articles associated with each of the comments rather than the average gender-typedness of the unique articles.

words describing men and women (e.g. male, female, man, woman, boy, girl). As this approach measures semantic distance to male vs. female words, higher values of the gender-typedness variable correspond to a greater female-typedness of the article.

The embeddings-based distributed dictionary approach can be applied to 1,154 articles discussed in the Wikipedia corpus (the remaining articles either did not mention any English words, or were not available on Wikipedia at the time of our analysis). The average gender-typedness score for these articles using the embeddings method is −0.04, indicating that the average article is more semantically distant to female words than male words. The average embeddings-based gender-typedness of articles associated with each comment is −0.05. The embeddings and WFA methods for calculating gender-typedness are quite similar, with a correlation of −0.62 on the comment level ($p < 0.001$) (this correlation is negative as articles with more male gender-typedness have positive values on the WFA measure and negative values on the embeddings measure). Similar to the WFA measure, articles with high levels of male content pertain to war, religion, and business, and articles with high levels of female content pertain to women's health, childbirth, and sexuality.

Importantly, we control for the valence and arousal of the article being discussed, as high or low valence and arousal articles are likely to have comments that are high or low in valence and arousal. We again do so by using both the WFA and the embeddings methods. There are a total of 1,154 unique articles for which we are able to compute WFA valence and arousal measures, with an average valence of 5.50 and an average arousal of 4.11. To illustrate, the articles with the highest and lowest valence scores in our dataset are "Ruth Westheimer" (an American sex therapist, media personality, and author) and "Crime in the United States", respectively. Other high valence articles include articles for popular celebrities, e.g. "Whoopi Goldberg", and articles for cultural products and phenomena such as "Smooth Jazz" and "Buddhism". Other low valence articles include ones for diseases, e.g. "Hodgkin lymphoma", social phenomena, e.g. "Hate group", and wars, e.g. "Korean War". In contrast, the articles with the highest and lowest arousal scores in our dataset are "Sexual abuse" and "Mesoamerican Long Count Calendar", respectively. Other high arousal articles include political movements and topics such as "Fascism" and "Nuclear war". Other low arousal articles include various uncontroversial topics, such as "Scientific method". The average WFA valence and arousal values of articles associated with each comment are 5.50 and 4.11, respectively.

The embeddings method also allows us to analyze the valence and arousal of 1,154 articles in our data. We find that the mean embeddings-based valence of these articles is 5.16 whereas the mean embeddings-based arousal of these articles is 4.22. The mean embeddings-based valence and arousal values of articles associated with each comment are 5.12 and 4.23, respectively. The embeddings and WFA methods for calculating valence and arousal are highly correlated, with a correlation of 0.90 and 0.81 on the comment level ($p < 0.001$). As is reflected in these strong correlations, the articles considered high or low in valence and high or low in arousal by the two methods are nearly identical.

3.3.4. Thread-level variables

A final set of controls involves thread-level variables. These are the number of unique editors commenting on the thread, the total number of comments on the thread, and the time difference (in number of days) between the first and the last comment on the thread. These have mean values of 5.31 unique editors, 16.99 comments, and 59.13 days on the comment level, respectively. Since the number of days between the first and last comment on a thread is highly skewed, with 36.76% of threads resolved within the same day, and 94.69% of threads resolved within

two weeks, we log-transform this variable in all subsequent analyses. The mean log-transformed value of thread time difference in days is 1.86.

4. Results

4.1. Overview

The code and analysis for this paper are available at https://osf.io/s8hef/. Below we analyze the relationship between gender, power, and a number of variables that, most notably, capture the emotional content of the comments in the Wikipedia discussions. As our goal is to understand these variables in the context of conversations within organizations, we exclude comments on threads in which only one editor makes a comment (i.e., there is no conversation). We also run regressions with numerous control variables, and thus exclude data for which these variables are not defined. Finally, we typically run two sets of regressions: one with variables obtained using our word-frequency averaging (WFA) method and one with variables obtained using our embeddings method. This means, for example, that our analysis of a comment's valance, as measured by our WFA method, will involve controlling for the valence, arousal, and gender-typedness of the article that is the topic of the thread using the WFA method. Conversely, our analysis of a comment's valence, as measured by our embeddings method, will involve controlling for the valence, arousal, and gender-typedness of the article that is the topic of the thread using the embeddings method. For expositional simplicity we do not include embeddings controls for WFA dependent variables, or vice versa.

4.2. Gender differences in domain choice

Before analyzing the emotionality of the conversations in our dataset, we examine whether there are systematic differences in the topics that women and men choose to converse on. We use a multiple logistic regression in which each observation corresponds to a comment, the dependent variable is whether or not the comment is written by a female editor, and the independent variables are various article-level characteristics, such as the article's valence, arousal and gender-typedness. We run two of these regressions, one with the article valence, arousal and gender-typedness variables obtained using our word-frequency averaging (WFA) method, and one with article valence, arousal and gender-typedness variables obtained using our embeddings method. For both these regressions we permit random effects in intercepts on the thread level. These random effects group (or nest) comments based on the thread they are in, in order to accommodate variability in gender across threads. In this sense our regression involves a hierarchical analysis.

As can be seen in Table 2, a comment is significantly more likely to be written by a woman if the article it pertains to is more female-typed (has fewer male pronouns than female pronouns, as with the WFA method, or is more semantically distant to male words relative to female words, as with the embeddings method). Using the WFA method we also find that women are less likely to comment on controversial articles when controlling for article characteristics like valence, arousal and gender-typedness, though this pattern is weaker and becomes non-significant when these characteristics are measured using the embeddings method. Finally, we find a positive relationship with article arousal and a non-systematic relationship with article valence. In the subsequent analyses, we control for these article-level variables when analyzing the relationship between the gender of the communicator and the emotionality of the comment.

Table 2
Word-frequency averaging and embeddings-based logistic regressions predicting whether the originator of the comment is female, as a function of various article-level variables.

	Coef.	S.E.	z	P > |z|	95%-L	95%-H
Word-frequency averaging						
Controversial	−0.13	0.06	−2.40	0.02	−0.24	−0.02
Gender-typedness (WFA)	−2.59	0.10	−26.13	0.00	−2.78	−2.39
Valence (WFA)	−0.50	0.13	−3.89	0.00	−0.75	−0.25
Arousal (WFA)	0.59	0.22	2.64	0.01	0.15	1.02
Embeddings method						
Controversial	−0.11	0.06	−1.90	0.06	−0.22	0.00
Gender-typedness (EMB)	41.00	1.32	30.98	0.00	38.41	43.59
Valence (EMB)	0.08	0.10	0.85	0.39	−0.11	0.27
Arousal (EMB)	3.09	0.23	13.28	0.00	2.64	3.55

Note: Random effects on thread level.

4.3. Gender differences in emotionality

We now examine whether there is a systematic gender difference in the expression of emotions among the general population of editors, where we first focus on valence and subsequently on arousal. We therefore regress comment valence and arousal on gender and also include the other editor-level, article-level, and thread-level variables discussed above. As before, we consider editor gender (=1 if female) as the main coefficient of interest, and we control for admin-status (=1 if the editor is an administrator) and experience (log number of prior edits). At the article-level, we control for valence, arousal, controversiality, and gender-typedness of the content. Thread-level controls are the number of comments, the number of unique editors, and the length of time between the first and last comments in the thread (in log days). To gain further insight about the structure of conversations, we also explore the role of comment order for emotionality by including a discrete variable indexing the comment's position in the thread. This variable takes on a value of 1 if the comment is the first in the thread, 2 if it is the second, and so on. We control for the date of the comment's posting (measured in days since 01/01/2000).

We use intercept random effects in our regressions to control for thread- and user-level heterogeneity not captured by our control variables. These nest comments made by each user in a thread in a single group. Thus, for example, we allow the valence of a comment to depend not only on the article-, thread-, and user-variables that are of central concern to our analysis, but also on an additive effect of the specific user in the thread. In this way comments by a given user in a given thread are grouped together, capturing user- and thread-level heterogeneity.

Lastly, as there are multiple variables being tested in each regression, we apply a Bonferroni correction for multiple comparisons. This yields a significance cutoff of $p = 0.0042$. We apply this regression in two ways: One using our WFA variables and one with our embeddings variables.

The results are shown in Table 3 for valence and Table 4 for arousal. As can be seen in Table 3, gender is a strong and significant predictor for comment valence for both the WFA and embeddings methods. The sign is positive, meaning that comments made by female editors are significantly higher in valence than comments from male editors. There are no other significant editor-level determinants of comment valence. There are, however, other article- and thread-level determinants (which we control for in the main analyses on gender and power dynamics). While these are not the focus of this study, we briefly present the patterns that emerge. Table 3 shows that comments have a significantly more positive valence ($p < 0.001$) if they are in threads about positively valenced articles, with fewer comments, fewer unique editors, and a shorter time between the first and last comments. Additionally, comments occurring later on in a conversation have a significantly higher valence than comments occurring towards the beginning. Comments occurring more recently in time also have higher valence. These patterns emerge with both the WFA and embeddings method. In addition to these, we also find that less arousing articles and articles

Table 3
Word-frequency averaging and embeddings-based regressions predicting comment valence from various user-, article-, thread-, and comment-level variables.

	Coef.	S.E.	z	P>|z|	95%-L	95%-H
Word-frequency averaging						
User female	0.054	0.006	8.560	0.000	0.042	0.067
User administrator	−0.005	0.005	−1.050	0.292	−0.015	0.004
User prior edits	0.002	0.001	2.130	0.033	0.000	0.004
Article valence (WFA)	0.329	0.008	39.870	0.000	0.312	0.345
Article arousal (WFA)	−0.033	0.014	−2.260	0.024	−0.061	−0.004
Article controversial	−0.002	0.004	−0.540	0.587	−0.009	0.005
Article gender-typedness (WFA)	−0.018	0.007	−2.460	0.014	−0.032	−0.004
Thread number of comments	−0.002	2.0×10^{-4}	−7.570	0.000	−0.002	−0.001
Thread number of users	−0.004	0.001	−4.680	0.000	−0.006	−0.002
Thread time difference	−0.004	0.001	−4.200	0.000	−0.006	−0.002
Comment order	0.001	2.1×10^{-4}	5.030	0.000	0.001	0.001
Comment date	1.6×10^{-5}	1.6×10^{-6}	9.960	0.000	1.3×10^{-5}	1.9×10^{-5}
Embeddings method						
User female	0.050	0.005	10.800	0.000	0.041	0.059
User administrator	0.002	0.003	0.570	0.569	−0.005	0.009
User prior edits	0.002	0.001	2.700	0.007	0.001	0.003
Article valence (EMB)	0.263	0.005	56.580	0.000	0.254	0.272
Article arousal (EMB)	−0.065	0.011	−5.830	0.000	−0.087	−0.043
Article controversial	−0.006	0.003	−2.440	0.015	−0.012	−0.001
Article gender-typedness (EMB)	0.519	0.063	8.240	0.000	0.396	0.643
Thread number of comments	−0.001	1.6×10^{-4}	−4.710	0.000	−0.001	-4.5×10^{-4}
Thread number of users	−0.004	0.001	−6.760	0.000	−0.005	−0.003
Thread time difference	−0.004	0.001	−5.510	0.000	−0.005	−0.003
Comment order	0.002	1.7×10^{-4}	10.460	0.000	0.001	0.002
Comment date	6.4×10^{-6}	1.2×10^{-5}	5.330	0.000	4.1×10^{-6}	8.7×10^{-6}

Note: Random effects on thread- and user-level. Bonferroni correction for multiple comparisons yields a significance cutoff of $p = 0.0042$ for each regression.

that are more female-typed have higher valence, as measured by the embeddings method.

Table 4 shows that comment arousal (which is a measure of the excitement or intensity of the comment) is not significantly predicted by editor-gender. It does, however, depend on the editor's prior experience, with editors with more prior edits writing relatively lower-arousal comments (in line with, e.g., Kucuktunc et al., 2012). Comments also have significantly higher arousal if they belong to conversations about lower-valence and higher-arousal articles, and if they involve a larger number of editors and unfold over a longer time span. More recently made comments have lower arousal. These differences appear with both the WFA and embeddings methods. Our regression with the embeddings method also suggests that threads with more comments have higher arousal.

We also examine the determinants of a comment's affective vs. cognitive content. Recall that this variable encodes the semantic distance between the comment and affective-process words relative to cognitive-process words (with positive values indicating a stronger

Table 4
Word-frequency averaging and embeddings-based regressions predicting comment arousal from various user-, article-, thread-, and comment-level variables.

	Coef.	S.E.	z	P > |z|	95%-L	95%-H
Word-frequency averaging						
User female	−0.006	0.004	−1.430	0.154	−0.014	0.002
User administrator	−0.002	0.003	−0.790	0.431	−0.008	0.004
User prior edits	−0.004	0.001	−6.330	0.000	−0.005	−0.003
Article valence (WFA)	−0.058	0.005	−11.070	0.000	−0.068	−0.047
Article arousal (WFA)	0.293	0.009	32.070	0.000	0.275	0.311
Article controversial	−0.001	0.002	−0.500	0.619	−0.006	0.003
Article gender-typedness (WFA)	0.013	0.005	2.960	0.003	0.004	0.022
Thread number of comments	-1.0×10^{-5}	1.4×10^{-4}	−0.700	0.482	-3.7×10^{-4}	1.7×10^{-4}
Thread number of users	0.003	0.001	5.500	0.000	0.002	0.004
Thread time difference	0.003	0.001	5.260	0.000	0.002	0.004
Comment order	3.0×10^{-5}	1.3×10^{-4}	0.250	0.806	-2.0×10^{-4}	3.0×10^{-4}
Comment date	-7.8×10^{-6}	1.5×10^{-6}	−7.510	0.000	-9.9×10^{-6}	-5.8×10^{-6}
Embeddings method						
User female	0.004	0.002	1.800	0.072	-2.1×10^{-6}	0.009
User administrator	−0.006	0.002	−3.370	0.001	−0.009	−0.003
User prior edits	−0.004	3.7×10^{-4}	−10.370	0.000	−0.005	−0.003
Article valence (EMB)	−0.019	0.002	−7.940	0.000	−0.024	−0.014
Article arousal (EMB)	0.273	0.006	47.960	0.000	0.262	0.284
Article controversial	0.002	0.001	1.110	0.267	−0.001	0.004
Article gender-typedness (EMB)	−0.063	0.032	−1.960	0.050	−0.126	4.7×10^{-6}
Thread number of comments	0.001	8.5×10^{-6}	6.390	0.000	3.7×10^{-4}	0.001
Thread number of users	0.002	3.2×10^{-4}	5.450	0.000	0.001	0.002
Thread time difference	0.005	3.6×10^{-4}	12.850	0.000	0.004	0.005
Comment order	1.4×10^{-4}	7.2×10^{-5}	1.930	0.053	-2.1×10^{-6}	2.8×10^{-4}
Comment date	-1.5×10^{-5}	6.1×10^{-7}	−24.400	0.000	-1.6×10^{-5}	-1.4×10^{-5}

Note: Random effects on thread- and user-level. Bonferroni correction for multiple comparisons yields a significance cutoff of $p = 0.0042$ for each regression.

Table 5
Embeddings-based regressions predicting cognitive vs. affective content of comment from various user-, article-, thread-, and comment-level variables.

	Coef.	S.E.	z	P>|z|	95%-L	95%-H
User female	−0.001	3.5×10^{-4}	−3.83	0.000	−0.002	−0.001
User administrator	1.4×10^{-4}	2.6×10^{-4}	0.52	0.603	-3.8×10^{-4}	0.001
User prior edits	3.7×10^{-4}	6.1×10^{-5}	6.64	0.000	2.6×10^{-4}	4.8×10^{-4}
Article valence (EMB)	−0.005	3.5×10^{-4}	−12.94	0.000	−0.005	−0.004
Article arousal (EMB)	−0.021	0.001	−25.52	0.000	−0.023	−0.02
Article controversial	0.002	2.0×10^{-4}	8.98	0.000	0.001	0.002
Article gender-typedness (EMB)	0.065	0.005	13.65	0.000	0.056	0.074
Thread number of comments	2.2×10^{-4}	1.2×10^{-5}	17.71	0.000	2.0×10^{-4}	2.4×10^{-4}
Thread number of users	-3.2×10^{-4}	5.0×10^{-5}	−6.78	0.000	-4.1×10^{-4}	-2.3×10^{-4}
Thread time difference	-1.2×10^{-4}	5.3×10^{-5}	−2.28	0.023	-2.3×10^{-4}	-2.3×10^{-5}
Comment order	-5.1×10^{-5}	1.1×10^{-5}	−4.96	0.000	-8.1×10^{-5}	-3.1×10^{-5}
Comment date	1.1×10^{-6}	9.0×10^{-6}	11.7	0.000	8.8×10^{-7}	1.2×10^{-6}

Note: Random effects on thread- and user-level. Bonferroni correction for multiple comparisons yields a significance cutoff of $p = 0.0042$ for each regression.

cognitive component). As measuring this variable involves using comment embeddings, we use only one set of regressions (with controls given by our embeddings approach for measuring article valence, arousal, and gender-typedness). The results of this regression are shown in Table 5, which indicates that comments made by female editors have more affective content than comments made by male editors, as expected. This table also shows that editors with fewer prior edits write comments with more affective rather than cognitive content. Additionally, there is more affective content in comments written about uncontroversial and male-typed articles as well as articles with higher valence and arousal. Table 5 also shows further significant thread- and comment-level predictors, which we do not discuss here.

While this regression focused on comments for which we could identify the editor's gender, the article- and thread-level patterns persist even when we examine all comments, including comments with non-identifiable editor-gender and prior edit count (see Tables A1–A3 in the Online Appendix). In the Online Appendix we also consider various user-, article-, and thread-level predictors for comment length. Table A4 shows that female editors write significantly longer comments than male editors, suggesting gender differences in commenting style that extend beyond emotionality.

4.4. Moderators of the gender-valence relationship

In the previous section we observed a strong main effect of editor gender on comment emotionality, with comments from female editors displaying a significantly more positive valence and more affective content than comments from male editors. In this section our goal is to understand the moderators of this tendency. While our main interest is to analyze the interaction between gender and power (admin-status, captured by the variable user administrator), we have also explored the interactions between gender and the ten other variables used in our analyses (user prior edits, article valence, article arousal, article controversiality, article gender-typedness, number of comments in thread, number of unique editors in thread, length of time of thread, position in thread, and date of comment). We report the results for completeness. We separate regressions with the emotionality of the comment (valence, arousal, or affective vs. cognitive content) as the dependent variable, the variables examined in the prior section as independent variables, and an interaction term between gender and one of these eleven variables. As above, our regressions include random effects on the user- and thread-level, and are performed with both the WFA and embeddings variables.

The outputs of the interaction effects for the regressions for comment valence are shown in Table 6. As can be seen, the only significant interaction with editor gender, for both the WFA and embeddings regressions, is admin-status (i.e., the position of authority). The negative value of this interaction shows that there is a drop in comment valence for female administrators relative to female non-administrators. Thus, it seems that the only variable that reduces the difference in comment valence across men and women is admin-status – i.e., the position of authority.

Table 7 shows a similar set of interactions for comment arousal. Here we see that there is no variable that crosses the threshold for significance when using a Bonferroni correction for multiple comparisons for both the WFA and embeddings methods. Thus, not only are there no gender differences in comment arousal, but gender also does not systematically interact with other variables to influence comment arousal.

Table 8 shows the results of these regressions for affective vs. cognitive content of comments. As we measure affective vs. cognitive content using embeddings, the interacting variables here include only embeddings variables. Interestingly, the previously observed tendency that female editors' comments have a higher affective vs. cognitive load is not mitigated by power, unlike our valence results. That is, female administrators also use more affective and less cognitive language than male administrators. We discuss possible interpretations of this finding in the last section of the paper. Finally, Table 8 also shows a significant interaction for article valence, suggesting that comments made by female editors have more affective content in threads involving higher-valenced articles.

4.5. The gender gap across the organizational hierarchy

In this final section, our goal is to examine the interaction between gender and power (as proxied by admin-status) in more detail.

4.5.1. Domain choice

Our analysis in Table 2 has shown that there are differences between men and women in terms of the articles they choose to converse on, with women more frequently commenting on female-typed articles, which are higher in arousal. In this analysis we did not find systematic effects of article controversiality and valence that persisted with both the WFA and embeddings methods. However, this analysis pooled administrators and non-administrators, and thus examined aggregate effects for gender, irrespective of power. Here we attempt similar tests separately for individuals at varying levels of the organizational hierarchy. We use a random-effects logistic regression on the comment level to predict whether a given comment is written by a man or a woman, using various article-level characteristics.

Table 9 shows that there are important reversals in gender differences for administrators vs. non-administrators in terms of their domain choice. Female non-administrators are significantly more likely than male non-administrators to comment on female-typed content and articles that are uncontroversial. In contrast, although female administrators still disproportionately comment on female-typed articles, gender differences in article controversiality reverse, with female administrators being slightly more likely than male administrators to

Table 6
Interaction effects between user gender and other possible predictors of comment valence, from eleven separate regressions for the word-frequency averaging and embeddings methods.

	Coef.	S.E.	z	P > |z|	95%-L	95%-H
Word-frequency averaging						
User administrator	−0.044	0.014	−3.080	0.002	−0.073	−0.016
User prior edits	−0.003	0.003	−1.030	0.302	−0.010	0.003
Article valence (WFA)	−0.037	0.029	−1.290	0.196	−0.094	0.019
Article arousal (WFA)	−0.080	0.049	−1.640	0.101	−0.176	0.016
Article controversial	−0.015	0.011	−1.390	0.165	−0.037	0.006
Article gender-typedness (WFA)	0.022	0.022	1.000	0.320	−0.021	0.064
Thread number of comments	3.6×10^{-4}	0.001	0.060	0.949	−0.001	0.001
Thread number of users	−0.001	0.002	−0.580	0.563	−0.005	0.003
Thread time difference	−0.002	0.003	−0.480	0.629	−0.009	0.005
Comment order	0.001	0.001	0.970	0.332	−0.001	0.002
Comment date	-3.1×10^{-6}	6.3×10^{-6}	−0.490	0.624	-1.5×10^{-7}	9.3×10^{-6}
Embeddings method						
User administrator	−0.044	0.010	−4.170	0.000	−0.064	−0.023
User prior edits	−0.005	0.002	−1.990	0.046	−0.009	8.1×10^{-6}
Article valence (EMB)	−0.006	0.015	−0.430	0.668	−0.035	0.022
Article arousal (EMB)	−0.046	0.037	−1.260	0.208	−0.118	0.026
Article controversial	−0.012	0.008	−1.530	0.127	−0.028	0.004
Article gender-typedness (EMB)	0.200	0.191	1.040	0.296	−0.175	0.574
Thread number of comments	8.6×10^{-6}	3.7×10^{-5}	0.230	0.818	−0.001	0.001
Thread number of users	−0.001	0.002	−0.910	0.364	−0.004	0.002
Thread time difference	0.001	0.003	0.550	0.581	−0.004	0.006
Comment order	0.001	4.8×10^{-4}	1.380	0.169	-2.8×10^{-4}	0.002
Comment date	3.0×10^{-7}	4.6×10^{-6}	0.070	0.948	-8.9×10^{-6}	8.2×10^{-6}

Note: Each of the eleven regressions includes our standard set of controls as well as random effects on the user- and thread-level. Bonferroni correction for multiple comparisons yields a significance cutoff of $p = 0.0045$ for each set of regressions.

comment on controversial articles. These results emerge for both our WFA and embeddings methods, and suggest that the non-systematic effects of article controversiality documented in our prior analysis (Table 2) were a product of gender differences across levels of power. In contrast, we do not find systematic and consistent differences across the two methods in terms of the valence and arousal of articles that women and men at different levels of the hierarchy are commenting on. (We do, however, consistently find that female non-administrators comment more on arousing articles than male non-administrators).

4.5.2. Emotionality

The above analysis finds that gender interacts with power to influence comment valence. In contrast, there are no systematic interactions between gender and power for comment arousal or a comment's affective vs. cognitive content. To develop an intuition of how the gender difference in valence changes as we consider individuals in positions of authority, we perform a simple aggregate analysis of comment valence across the four groups of male administrators, female administrators, male non-administrators, and female non-administrators. The basic

Table 7
Interaction effects between user-gender and other possible predictors of comment arousal, from eleven separate regressions for the word-frequency averaging and embeddings methods.

	Coef.	S.E.	z	P>|z|	95%-L	95%-H
Word-frequency averaging						
User administrator	0.009	0.009	0.970	0.331	−0.009	0.027
User prior edits	−0.002	0.002	−0.760	0.444	−0.006	0.003
Article valence (WFA)	0.039	0.018	2.140	0.032	0.003	0.075
Article arousal (WFA)	0.043	0.031	1.400	0.161	−0.017	0.104
Article controversial	0.009	0.007	1.210	0.225	−0.005	0.022
Article gender-typedness (WFA)	−0.038	0.014	−2.740	0.006	−0.064	−0.011
Thread number of comments	-3.6×10^{-4}	3.2×10^{-4}	−1.130	0.258	−0.001	2.7×10^{-4}
Thread number of users	−0.001	0.001	−1.030	0.304	−0.004	0.001
Thread time difference	4.4×10^{-4}	0.002	0.200	0.840	−0.004	0.005
Comment order	−0.001	4.3×10^{-4}	−2.150	0.031	−0.002	-8.0×10^{-6}
Comment date	-3.7×10^{-6}	4.0×10^{-6}	−0.920	0.359	-1.1×10^{-7}	4.6×10^{-6}
Embeddings method						
User administrator	−0.014	0.005	−2.550	0.011	−0.024	−0.003
User prior edits	−0.003	0.001	−2.390	0.017	−0.005	−0.001
Article valence (EMB)	0.006	0.007	0.860	0.392	−0.008	0.021
Article arousal (EMB)	−0.017	0.019	−0.920	0.356	−0.054	0.019
Article controversial	−0.004	0.004	−1.020	0.306	−0.012	0.004
Article gender-typedness (EMB)	0.312	0.098	3.190	0.001	0.120	0.503
Thread number of comments	1.4×10^{-4}	1.8×10^{-4}	−0.770	0.443	−0.001	2.2×10^{-4}
Thread number of users	−0.002	0.001	−2.040	0.041	−0.003	-6.0×10^{-5}
Thread time difference	0.002	0.001	1.660	0.096	-3.8×10^{-4}	0.005
Comment order	-4.5×10^{-4}	2.4×10^{-4}	−1.860	0.063	−0.001	2.4×10^{-4}
Comment date	1.4×10^{-6}	2.3×10^{-6}	0.580	0.561	-3.2×10^{-6}	6.0×10^{-6}

Note: Each of the eleven regressions includes our standard set of controls as well as random effects on the user- and thread-level. Bonferroni correction for multiple comparisons yields a significance cutoff of $p = 0.0045$ for each set of regressions.

Table 8
Interaction effects between user gender and other possible predictors of comment cognitive vs. affective content, from eleven separate regressions for the embeddings method.

	Coef.($\times 10^{-3}$)	S.E.($\times 10^{-3}$)	z	P > |z|	95%-L($\times 10^{-3}$)	95%-H($\times 10^{-3}$)
User administrator	−4.732	7.883	−0.6	0.548	−20.182	10.717
User prior edits	4.714	1.802	2.62	0.009	1.182	8.245
Article valence (EMB)	−36.512	11.039	−3.31	0.001	−58.148	−14.876
Article arousal (EMB)	74.473	27.507	2.71	0.007	20.559	128.386
Article controversial	−13.532	6.054	−2.24	0.025	−25.397	−1.667
Article gender-typedness (EMB)	−285.597	143.923	−1.98	0.047	−567.680	−3.513
Thread number of comments	−0.252	0.283	−0.89	0.373	−0.807	0.303
Thread number of users	−1.362	1.128	−1.21	0.227	−3.574	0.849
Thread time difference	−2.975	1.898	−1.57	0.117	−6.694	0.745
Comment order	−0.568	0.363	−1.56	0.118	−1.279	0.144
Comment date	0.002	0.003	0.54	0.592	−0.005	0.009

Note: Each of the coefficient-, standard error-, and confidence interval values have been multiplied by 10^3 to aid exposition. To obtain the actual values multiply each number by 10^{-3}. Each of the eleven regressions includes our standard set of controls as well as random effects on the user- and thread-level. Bonferroni correction for multiple comparisons yields a significance cutoff of $p = 0.0045$ for each set of regressions.

Table 9
Word-frequency averaging and embeddings-based logistic regressions predicting whether the originator of the comment is female, using various article-level variables, for non-administrators and administrators, respectively.

	Coef.	S.E.	z	P > |z|	95%-L	95%-H
Word-frequency averaging						
Not administrator						
Controversial	−0.235	0.062	−3.800	0.000	−0.356	−0.114
Gender-typedness (WFA)	−2.305	0.111	−20.760	0.000	−2.523	−2.088
Valence (WFA)	−0.198	0.141	−1.410	0.159	−0.474	0.078
Arousal (WFA)	0.884	0.243	3.640	0.000	0.408	1.360
Administrator						
Controversial	0.383	0.151	2.540	0.011	0.087	0.679
Gender-typedness (WFA)	−3.875	0.269	−14.390	0.000	−4.403	−3.347
Valence (WFA)	−1.355	0.346	−3.920	0.000	−2.033	−0.677
Arousal (WFA)	−0.548	0.613	−0.890	0.372	−1.749	0.654
Embeddings method						
Not administrator						
Controversial	−0.331	0.062	−5.310	0.000	−0.453	−0.209
Gender-typedness (EMB)	42.023	1.481	28.370	0.000	39.120	44.926
Valence (EMB)	0.094	0.108	0.870	0.384	−0.118	0.307
Arousal (EMB)	3.033	0.256	11.830	0.000	2.531	3.536
Administrator						
Controversial	0.525	0.153	3.440	0.001	0.226	0.824
Gender-typedness (EMB)	48.101	3.309	14.530	0.000	41.615	54.587
Valence (EMB)	0.148	0.225	0.660	0.509	−0.292	0.589
Arousal (EMB)	3.697	0.610	6.060	0.000	2.502	4.892

Note: Random effects on article-level. Bonferroni correction for multiple comparisons yields a significance cutoff of $p = 0.0125$ for each regression.

analysis regresses comment valence on gender (1 if female, 0 otherwise), admin-status (1 if administrator, 0 otherwise), and their interaction, and does not control for the other editor-, article-, or thread-level variables (Fig. 1A for WFA method and 1B for embeddings method). It nonetheless shows a robust interaction of gender and admin-status ($\beta = -0.05$, $z = -4.39$, $p < 0.001$, 95%CI $= [-0.07, -0.03]$ for WFA and $\beta = -0.05$, $z = -6.24$, $p < 0.001$, 95%CI $= [-0.06, -0.03]$ for embeddings). The comments written by male administrators, female administrators, and male non-administrators are not statistically distinguishable in terms of their valence, but the comments of female non-administrators are. These comments are much more positive than all other comments in the dataset. This suggests that, while the expected gender differences in valence emerge for the general population, the valence of the comments of women in positions of power are indistinguishable from the valence

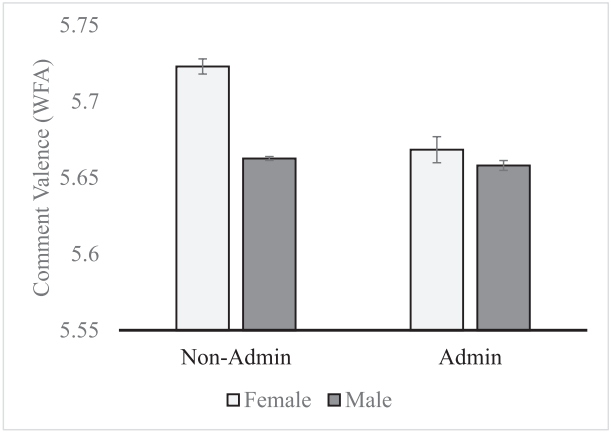

Fig. 1A. Simple aggregate comment-valence scores across user groups (Word-frequency averaging method). Error bars indicate +/- 1 SE.

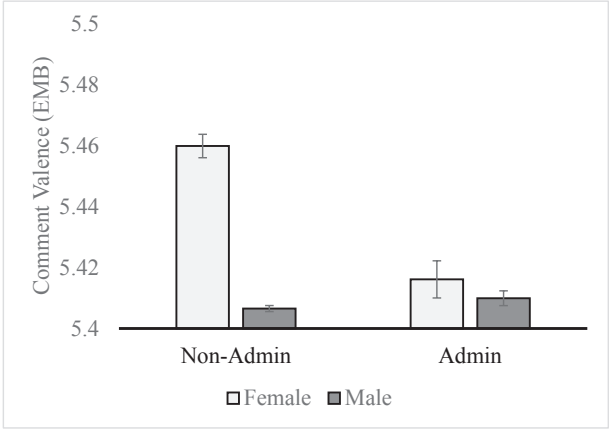

Fig. 1B. Simple aggregate comment-valence scores across user groups (Embeddings method). Error bars indicate +/- 1 SE.

of the comments of their male counterparts.

To gain a deeper understanding of the interaction effects, we also perform an analysis of the valence of the words with the highest relative probabilities of being used by either of the four groups. The analysis only considers Warriner et al. words that occur more than 1,000 times in the dataset. This is done to ensure that the results are not driven by rare words, which have low probabilities of occurrence and are subsequently very hard to predict (Taleb, 2007). Including such rare words would yield spurious, highly-skewed probabilities that would bias our

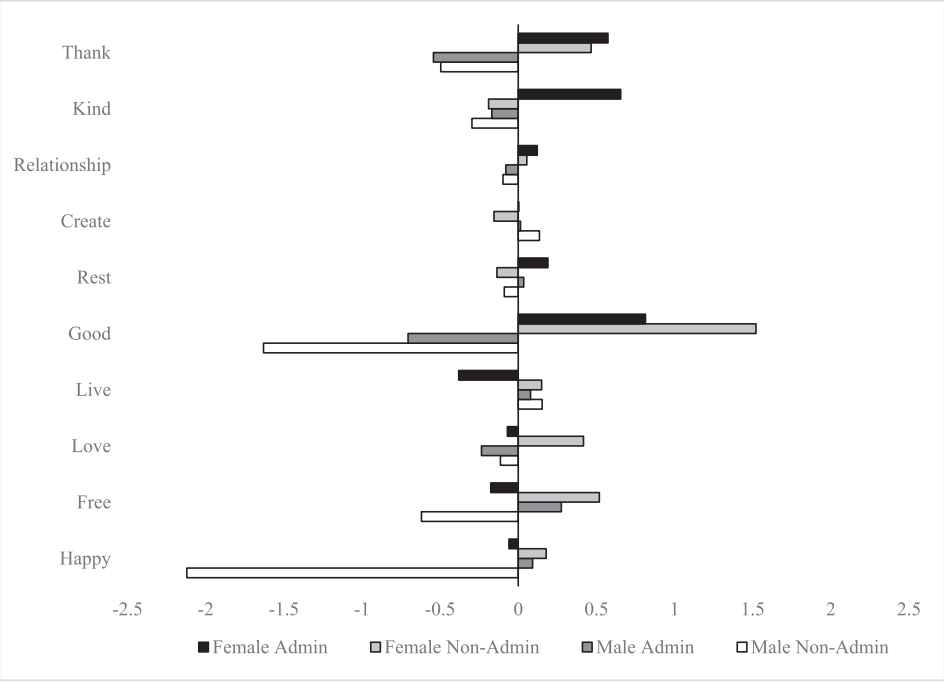

Fig. 2. Relative probabilities of occurrence for the ten highest-valence words.

results.

There are 12,338 words that occur more than 1,000 times in the dataset. To measure the relative probabilities of these words being used by the four groups, we first calculate how many times each of the words occurs in comments made by male administrators, female administrators, male non-administrators, and female non-administrators. We then divide each word's frequency by the total number of words written by the four groups of editors, to get each word's probability of occurrence in comments made by each of the four groups. We write these probabilities for word i as p_i^{MA} (male admin), p_i^{FA} (female admin), p_i^{MnA} (male non-admin), and p_i^{FnA} (female non-admin). Finally, we compute the relative probabilities of occurrence for each word in each group by subtracting the average of these four probabilities, p_{ave} = Average $\{p_i^{MA}, p_i^{FA}, p_i^{MnA}, p_i^{FnA}\}$. We denote these relative probabilities for word i as r_i^{MA}, r_i^{FA}, r_i^{MnA}, and r_i^{FnA}, with $r_i^{MA} = p_i^{MA} - p_{ave}$, $r_i^{FA} = p_i^{FA} - p_{ave}$, and so on.

Fig. 2 shows the relative probabilities of occurrence for the ten highest-valence words that occur at least 1,000 times in our dataset. Here, we see that female non-administrators have the highest relative probabilities for four out of these ten words ("happy", "free", "love", and "good"), and the second-highest relative probabilities for another three of these words ("live", "relationship", and "thank").

We use a logistic regression to test for this relationship between the valence of each of the 12,338 words that occur more than 1,000 times in the dataset (our independent variable) and whether or not the word has the highest relative probability of occurrence in the comments made by female non-administrators (our dependent variable, which assumes the value zero for the three remaining user groups). This analysis reveals a significant positive relationship ($\beta = 0.20$, $z = 2.43$, $p = 0.015$, 95%CI = [0.04, 0.36]), showing that higher-valenced words are indeed statistically significantly more likely to be coming from female non-administrators (compared to male administrators, male non-administrators, and female administrators).

In Fig. 3 we divide these 12,338 words into four quartiles based on their valence (1st and 4th quartiles corresponding to the lowest- and highest-valence words, respectively), and show the proportion of words in each of the four quartile groups with the highest relative probability of occurrence in the comments made by female non-administrators.

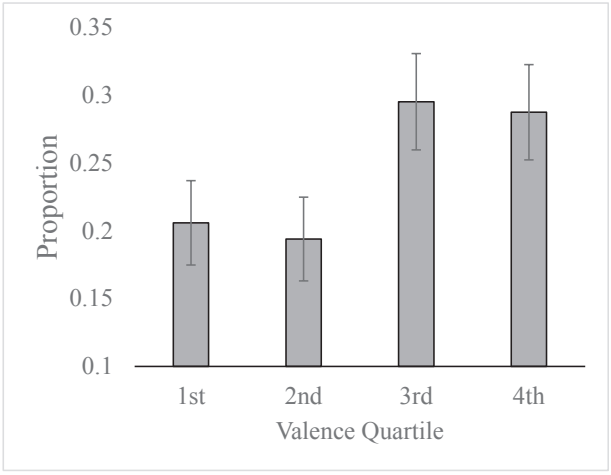

Fig. 3. Proportion of words with highest relative probability of usage by female non-admins. Error bars indicate +/- 1 SE

Here we can see that lower-valenced words (1st and 2nd quartiles) typically do not have the highest relative probability of occurrence in the comments made by female non-administrators, whereas higher-valenced words (3rd and 4th quartiles) do. This again shows that female non-administrators are relatively more likely to use higher-valenced words, relative to the other three groups.

4.5.3. *Exploratory analysis of mechanisms*

As discussed in the theory section, the main mechanisms behind the convergence we observe may be a treatment effect, or sorting in the form of social- and self-selection. A comprehensive comparison of these mechanisms would require novel, ideally experimental data involving the random assignment of users to positions of power, which is beyond the scope of the current paper. A weaker analysis involves comparing the emotional styles of users who eventually become administrators with those of users who do not come to occupy administrator positions, or alternatively comparing the emotional styles of users before and after they become administrators. Although our dataset is extensive, the

gender imbalances in Wikipedia editor and administrator roles mean that there are only twenty-six women for whom we observe comments made in both non-administrator and administrator positions. Thus, our ability to test for underlying mechanisms is restricted. Nonetheless, we include some exploratory tests, which indicate that a treatment effect of the position of authority may be involved (again, the three mechanisms are not mutually exclusive).

First, we analyze whether comments made by female editors who *later* rise to a position of authority differ from those of their female peers from the general population who do not become administrators later on. We do this by regressing comment valence on a binary variable indicating whether or not the user would eventually become an administrator. We run this regression only for comments made by female non-administrators, and include our standard set of controls (user log-edit count, article valence, arousal, controversiality, and gender-typedness, comment order, comment date, thread number of comments, users, and time difference) as well as random effects on the user- and thread-level. We do not find a significant difference between the comment valence of female non-administrators who eventually become administrators and the comment valence of female non-administrators who do not become administrators when running this regression with the WFA method ($\beta = 0.03$, $SE = 0.04$, $z = 0.76$, $p = 0.49$, *95% CI* $= [-0.05,0.10]$) or with the embeddings method ($\beta = 0.04$, $SE = 0.03$, $z = 1.45$, $p = 0.15$, *95% CI* $= [-0.01,0.10]$). This suggests that female administrators do not differ in their emotionality (valence) from other women before they come to occupy the position of authority.

Second, we tentatively explore whether there may be a treatment effect of the position of authority on women's subsequent behavior by analyzing the data on women for whom we have observations on both the time before and during their adminship. We test whether there is a change in comment valence as they become administrators. Again, this involves a regression of comment valence on a binary variable indicating whether or not the user is an administrator at the time of posting (using only the comments generated by women for whom we have data from before and after they become administrators). We run this regression with the controls discussed above and include random effects on the user- and thread-level. We observe a directional drop in comment valence as women come to occupy the position of authority using the WFA method, though this drop is not significant ($\beta = -0.04$, $SE = 0.04$, $z = -1.06$, $p = 0.29$, *95% CI* $= [-0.12,0.04]$). However, we do find a significant drop using the embeddings method ($\beta = -0.07$, $SE = 0.03$, $z = -2.39$, $p = 0.02$, *95% CI* $= [-0.13, -0.01]$). While this analysis is limited as there are only twenty-six editors for whom we have the requisite data, meaning that we remain cautious about robustness, these results would be consistent with an interpretation that holding powerful office may have an influence on behavior – possibly legitimizing or compelling women to reduce the valence in their communications. Replicating these results and analyzing these mechanisms in more detail is an important and promising avenue for future research.

5. Discussion and conclusion

Our analysis yields several implications for research on gender differences, leadership behavior, and conversational phenomena within modern forms of knowledge production. In many of these novel organizational forms (e.g., Wikipedia, open source software production), selection into different work domains is voluntary and neither mandated nor predominantly motivated by pecuniary incentives. Millions of people around the world coordinate their efforts in virtual space, often without much personal interaction (e.g., without private communication in small or two-person teams, discussion of matters not related to work, or face-to-face meetings). Analysis of communication in such forms of production provides interesting insights for the future of work given the increasing predominance of large teams and the rise in alternative, often platform-based work arrangements. This makes understanding the linguistic coordination of people working on Wikipedia important for organizational scholars. Wikipedia has attracted much interest based on what is considered a relatively anti-authoritarian and decentralized structure. It is therefore surprising to see the role played by authoritative positions even in such an environment where workers are possibly less influenced by authority.

With regards to research on gender in organizations, we show that there are significant gender differences in people's *conversational styles* (specifically, in their emotionality and emphasis on affect vs. cognition) and *domain choices* (controversiality and gender-typedness). Importantly, once we look up the organizational hierarchy to individuals in positions of power, these differences diminish or even disappear: female and male authorities are just as (un)emotional in terms of valence in their language use, and they are just as likely to engage in conversations about controversial content. As our analyses also show, this change is driven by women who converge to the behavior of their male counterparts as they assume positions of power. The two notable exceptions are that the gender-specific separation of labor – sorting into conversational topics based on their gender stereotype – seems to increase. This may be explained by differences in accumulated knowledge and expertise that editors can leverage once they become administrators. Moreover, female administrators continue to use fewer cognitive process words and more affective process words than male administrators. This is an interesting result that future research should explore further. It might be an indication of female leaders' intent to navigate a competence-warmth trade-off (Fiske, Cuddy, Glick, & Xu, 2002), whereby they counterbalance their position of power by renouncing the use of overly cognitive words. (The average comment in our dataset is more semantically related to cognitive rather than affective words, which is expected given Wikipedia's focus on being a knowledge creation and not a social media platform.)

Our finding of the disappearance of important gender gaps in emotionality and domain choice among people in positions of power is in line with other work in the gender literature (Croson & Gneezy, 2009). Previous work shows, for instance, that the well-established gender difference in risk preferences does not extend from the general population to managers. Croson and Gneezy (2009) conclude, that "the evidence suggests that managers and professional business persons present an important exception to the rule that women are more risk averse than men" (p. 454). These findings were obtained for trained managers, which opens the possibility (also discussed by Croson and Gneezy) that the training may have affected women's behavior (see, e.g., Johnson and Powell (1994), who compare trained and untrained subpopulations, as well as Masters and Meier (1988) and Birley (1987) who focus on entrepreneurs). We find such convergence even in a population of untrained individuals, as Wikipedia administrators presumably did not undergo formal management education.

We find suggestive evidence that the position of authority may have an effect on the disappearance of the gender gaps. In line with our findings, a recent study that looks at laughter occurrences documents a similar pattern where women in positions of power converge to the behavior of men and exhibit less inauthentic laughter – even when power is exogenously assigned (Bitterly, Brooks, Aaker, & Schweitzer, 2020).

Other possible mechanisms behind the smaller and even disappearing gender gaps in our data are self- and social selection (i.e., supply- and demand-side factors). Analyzing these mechanisms, including how they interact, is an important avenue for research (Fernandez-Mateo & Kaplan, 2018). Such future work could also consider further measures of power, for instance by using social network analyses to build centrality measures. Replicating our results with such measures would be useful, and it would also open other intriguing questions, such as about the extent to which formal (adminship) and informal (social network-based) measures of power overlap in the context of Wikipedia and beyond.

To draw implications for interventions, it will be important to replicate our findings, ideally by experimentally assigning power in order to identify its causal effects. To illustrate, if lower valence is a cause of power, then organizations may want to actively counterbalance its weight in promotion processes (assuming it is not related to ability and leader effectiveness). Conversely, however, if power causes lower-valenced communication, the implications would be different. For instance, to the extent that low valence creates a less enjoyable organizational culture, policies that foster more positive interactions would be conducive.

More generally, understanding differences in the behaviors of men and women across different hierarchy levels of organizations is a necessary first step to understanding how to remedy gender differences in organizational outcomes. Our analysis takes this step, and it sets the basis for a more rigorous and naturalistic examination of power-gender dynamics. Another practical benefit is that our methods can be deployed at scale and in real time. Using automated text analysis, organizations can thus monitor possible gender differences. This will help them better understand organizational dynamics and, if need be, control for these when making promotion decisions.

Follow-up work should further investigate the role played by the mode of collaboration, comparing conversations in more traditional, small-scale, and co-located team production settings (Leavitt, 1989) to novel conversational phenomena that arise from large-scale collaborations among self-governing "peers". Our study focuses on the latter. It is interesting that even in such a context, where gender cues are reduced and work takes place in virtual space, we observe notable gender differences among the general population of editors. This provides further evidence, from a non-student sample, that gendered power differentials may persist in online contexts (Guiller & Durndell, 2007). It is likely that different conversational dynamics unfold where gender cues are more salient and where voice and nonverbal behaviors may be used by women in an attempt to mitigate adverse, possibly gender-specific, consequences from leader-like behaviors (Carli, LaFleur, & Loeber, 1995; Eagly & Karau, 2002; Hall, Coats, & LeBeau, 2005; Schroeder, Kardas, & Epley, 2017). Such analyses in offline contexts would have the benefit that they do not rely on people's decision to reveal their gender and/or other characteristics (e.g., with our dataset we cannot study or control for ethnicity). It remains an empirical question whether people who share their gender differ systematically from those who do not, and on what dimensions. Our results pertain to the population of editors whose gender is revealed, which is only a subgroup of Wikipedia editors.

By considering emotions as a window into the psychology of knowledge production, we hope to provide a basis for further research into the motivations driving the production of global public goods such as Wikipedia. It would be interesting to study the extent to which expressions of emotions in this virtual knowledge production context reflect actually experienced feelings as opposed to possible attempts to conform to gender- and leadership-role specific display rules (Brody, 2000; Simpson & Stroh, 2004). More generally, future work could use automated text analysis to examine a variety of psychological variables and constructs in naturally occurring conversations (see Humphreys & Wang, 2017), with important implications for our understanding of gender, power, and other key social variables in organizations and in everyday life. By using automated text analysis applied to a large dataset of Wikipedia editor conversations, our paper is intended to help lay the groundwork for such an analysis.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.obhdp.2020.02.003.

References

Bitterly, T. B., Brooks, A. W., Aaker, J. L., & Schweitzer, M. E. (2020). Why women laugh more than men. Unpublished manuscript.

Adams, R. B., & Funk, P. (2012). Beyond the glass ceiling: Does gender matter? *Management Science, 58*(2), 219–235.

Amanatullah, E. T., ... Tinsley, C. H. (2013). Punishing female negotiators for asserting too much…or not enough: Exploring why advocacy moderates backlash against assertive female negotiators. *Organizational Behavior and Human Decision Processes, 120*(1), 110–122. https://doi.org/10.1016/j.obhdp.2012.03.006.

Anderson, C., Hildreth, J. A. D., & Howland, L. (2015). Is the desire for status a fundamental human motive? A review of the empirical literature. *Psychological Bulletin, 141*(3), 574.

Bear, J. B., & Collier, B. (2016). Where are the women in Wikipedia? Understanding the different psychological experiences of men and women in Wikipedia. *Sex Roles, 74*(5), 254–265.

Bear, J. B., Weingart, L. R., & Todorova, G. (2014). Gender and the emotional experience of relationship conflict: The differential effectiveness of avoidant conflict management. *Negotiation and Conflict Management Research, 7*(4), 213–231. https://doi.org/10.1111/ncmr.12039.

Bem, S. L., & Lenney, E. (1976). Sex typing and the avoidance of cross-sex behavior. *Journal of Personality and Social Psychology, 33*(1), 48.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom.* New Haven, CT: Yale University Press.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences, 29*, 31–36.

Birley, S. (1987). Female entrepreneurs: Are they really any different? *Journal of Small Business Management, 27*(1), 32–37.

Bischoping, K. (1993). Gender differences in conversation topics, 1922–1990. *Sex Roles, 28*(1–2), 1–18.

Bohnet, I. (2016). *What works: Gender equality by design.* Cambridge, MA: Harvard University Press.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report C-1, The center for research in psychophysiology*. University of Florida.

Branson, D. M. (2006). *No seat at the table: How corporate governance and law keep women out of the boardroom.* New York City, NY: NYU Press.

Brody, L. R. (2000). The socialization of gender differences in emotional expression: Display rules, infant temperament, and differentiation. In A. Fischer (Ed.). *Gender and emotion: Social psychological perspectives* (pp. 24–47). Cambridge, UK: Cambridge University Press.

Brooks, A. W., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences, 111*(12), 4427–4431. https://doi.org/10.1073/pnas.1321202111.

Carli, L. L. (1990). Gender, language, and influence. *Journal of Personality and Social Psychology, 59*(5), 941.

Carli, L. L., LaFleur, S. J., & Loeber, C. C. (1995). Nonverbal behavior, gender, and influence. *Journal of Personality and Social Psychology, 68*(6), 1030–1041.

Costa, P. T., Jr, Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*(2), 322–331.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature, 47*(2), 448–474.

Cross, S. E., & Madson, L. (1997). Models of the self: Self-construals and gender. *Psychological Bulletin, 122*(1), 5.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation.* Hillsdale, NJ: Erlbaum.

Eagly, A. H., & Carli, L. L. (2003). The female leadership advantage: An evaluation of the evidence. *Leadership Quarterly, 14*(6), 807–834.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573–598.

Eagly, A. H., & Wood, W. (2012). Social role theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.). *Handbook of theories in social psychology*. London: Sage.

Faraj, S., Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge collaboration in online communities. *Organization Science, 22*(5), 1224–1239.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116*(3), 429.

Fernandez-Mateo, I., & Fernandez, R. M. (2016). Bending the pipeline? Executive search and gender inequality in hiring for top management jobs. *Management Science, 62*(12), 3636–3655. https://doi.org/10.1287/mnsc.2015.2315.

Organization Science29(6), 1229–1236. https://doi.org/10.1287/orsc.2018.1249.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902.

Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From power to action. *Journal of Personality and Social Psychology, 85*(3), 453–466.

Galinsky, A. D., Magee, J. C., Inesi, M. E., & Gruenfeld, D. H. (2006). Power and perspectives not taken. *Psychological Science, 17*(12), 1068–1074. https://doi.org/10.1111/j.1467-9280.2006.01824.x.

Gallus, J. (2017). Fostering public good contributions with symbolic awards: A large-scale natural field experiment at Wikipedia. *Management Science, 63*(12), 3999–4015.

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods, 50*(1), 344–361.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. *Academy of Management Journal, 59*(5), 1493–1507.
Gino, F., Wilmuth, C. A., & Brooks, A. W. (2015). Compared to men, women view professional advancement as equally attainable, but less desirable. *Proceedings of the National Academy of Sciences, 112*(40), 12354–12359.
Gipson, A. N., Pfaff, D. L., Mendelsohn, D. B., Catenacci, L. T., & Burke, W. W. (2017). Women and leadership: Selection, development, leadership style, and performance. *Journal of Applied Behavioral Science, 53*(1), 32–65. https://doi.org/10.1177/0021886316687247.
Glott, R., Schmidt, P., & Ghosh, R. (2010). Wikipedia survey – overview of results. UNU-MERIT, http://www.ris.org/uploadi/editor/1305050082Wikipedia_Overview_1305050015March1305052010-FINAL.pdf.
Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics, 118*(3), 1049–1074.
Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review, 90*(4), 715–741.
Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition, 135*, 21–23.
Guiller, J., & Durndell, A. (2007). Students' linguistic behaviour in online discussion groups: Does gender matter? *Computers in Human Behavior, 23*(5), 2240–2255.
Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin, 131*(6), 898–924.
Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods, 21*(4), 447.
Hawn, C. (2009). Take two aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care. *Health Affairs, 28*(2), 361–368.
Herring, S. C. (2000). Gender differences in CMC: Findings and implications. *Computer Professionals for Social Responsibility Journal, 18*(1).
Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology, 70*(8), 1603–1619.
Hopkins, M. M., & Bilimoria, D. (2008). Social and emotional competencies predicting success for male and female executives. *Journal of Management Development, 27*(1), 13–35.
Huang, L., Gino, F., & Galinsky, A. D. (2015). The highest form of intelligence: Sarcasm increases creativity for both expressers and recipients. *Organizational Behavior and Human Decision Processes*. https://doi.org/10.1016/j.obhdp.2015.07.001.
Huberman, B. A., Loch, C. H., & Önçüler, A. (2004). Status as a valued resource. *Social Psychology Quarterly, 67*(1), 103–114.
Humphreys, A., & Wang, R. J.-H. (2017). Automated text analysis for consumer research. *Journal of Consumer Research, 44*(6), 1274–1306.
Johnson, J. E., & Powell, P. L. (1994). Decision making, risk and gender: Are managers different? *British Journal of Management, 5*(2), 123–138.
Johnson, M., & Helgeson, V. S. (2002). Sex differences in response to evaluative feedback: A field study. *Psychology of Women Quarterly, 26*(3), 242–251. https://doi.org/10.1111/1471-6402.00063.
Katz, L. F., & Krueger, A. B. (2019). The rise and nature of alternative work arrangements in the United States, 1995–2015. *ILR Review, 72*(2), 382–416.
Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review, 110*(2), 265–284.
Klapper, H., & Reitzig, M. (2018). On the effects of authority on peer motivation: Learning from Wikipedia. *Strategic Management Journal, 39*(8), 2178–2203.
Kosinski, M., & Behrend, T. (2017). Editorial overview: Big data in the behavioral sciences. *Current Opinion in Behavioral Sciences, 18*, iv–vi. https://doi.org/10.1016/j.cobeha.2017.11.007.
Kucuktunc, O., Cambazoglu, B. B., Weber, I., & Ferhatosmanoglu, H. (2012). *A large-scale sentiment analysis for Yahoo! answers. Paper presented at the proceedings of the fifth ACM international conference on web search and data mining*.
Kugler, K. G., Reif, J. A., Kaschner, T., & Brodbeck, F. C. (2018). Gender differences in the initiation of negotiations: A meta-analysis. *Psychological Bulletin, 144*(2), 198–222.
Lakhani, K. R., & Von Hippel, E. (2003). How open source software works: "Free" user-to-user assistance. *Research Policy, 32*(6), 923–943.
Laniado, D., Kaltenbrunner, A., Castillo, C., & Morell, M. F. (2012). Emotions and dialogue in a peer-production community: The case of Wikipedia. In: Paper presented at the proceedings of the eighth annual international symposium on wikis and open collaboration.
Lazear, E. P., & Shaw, K. L. (2007). Personnel economics: The economist's view of human resources. *Journal of Economic Perspectives, 21*(4), 91–114.
Leavitt, H. J. (1989). Suppose we took groups seriously. In H. J. Leavitt, L. R. Pondy, & D. M. Boje (Eds.). *Readings in managerial psychology*(4th ed.). Chicago and London: University of Chicago Press.
Lee, E.-J. (2007). Wired for gender: Experientiality and gender-stereotyping in computer-mediated communication. *Media Psychology, 10*(2), 182–210.
Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics, 4*, 151–171.
Lerner, J., Pathak, P. A., & Tirole, J. (2006). The dynamics of open-source contributors. *American Economic Review, 96*(2), 114–118.
Levine, S. S., & Prietula, M. J. (2014). Open collaboration for innovation: Principles and performance. *Organization Science, 25*(5), 1287–1571.
Lih, A. (2009). *The Wikipedia revolution: How a bunch of nobodies created the world's greatest encyclopedia.* New York City, NY: Hachette Books.
Luhaorg, H., & Zivian, M. T. (1995). Gender role conflict: The interaction of gender, gender role, and occupation. *Sex Roles, 33*(9–10), 607–620.
Magee, J. C., & Galinsky, A. D. (2008). Social hierarchy: The self-reinforcing nature of power and status. *Academy of Management Annals, 2*(1), 351–398. https://doi.org/10.5465/19416520802211628.
Masters, R., & Meier, R. (1988). Sex differences and risk-taking propensity of entrepreneurs. *Journal of Small Business Management, 26*(1), 31.
McCrae, R. R., Terracciano, A., & 78 Members of the Personality Profiles of Cultures Project (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology, 88*(3), 547–561.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality. Paper presented at the advances in neural information processing systems*.
Mulac, A. (1998). The gender-linked language effect: Do language differences really make a difference? In D. J. D. K. Canary (Ed.). *Sex differences and similarities in communication: Critical essays and empirical investigations of sex and gender in interaction* (pp. 127–155). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics, 122*(3), 1067–1101.
Palomares, N. A. (2004). Gender schematicity, gender identity salience, and gender-linked language use. *Human Communication Research, 30*(4), 556–588.
Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Retrieved from http://hdl.handle.net/2152/31333.
Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*(1), 547–577.
Polachek, S. W. (1981). Occupational self-selection: A human capital approach to sex differences in occupational structure. *Review of Economics and Statistics, 63*(1), 60–69.
Prabhakaran, V., & Rambow, O. (2016). A corpus of Wikipedia discussions: Over the years, with topic, power and gender labels. Paper presented at the LREC.
Prinsen, F. R., Volman, M. L. L., & Terwel, J. (2007). Gender-related differences in computer-mediated communication and computer-supported collaborative learning. *Journal of Computer Assisted Learning, 23*(5), 393–409.
Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology, 74*(3), 629–645.
Rudman, L. A., & Fairchild, K. (2004). Reactions to counterstereotypic behavior: The role of backlash in cultural stereotype maintenance. *Journal of Personality and Social Psychology, 87*(2), 157–176.
Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161.
Schein, E. H. (2004). *Organizational culture and leadership* (3rd ed.). San Francisco, CA: Jossey-Bass.
Schneider, M. C., Holman, M. R., Diekman, A. B., & McAndrew, T. (2016). Power, conflict, and community: How gendered views of political power influence women's political ambition. *Political Psychology, 37*(4), 515–531. https://doi.org/10.1111/pops.12268.
Schroeder, J., Kardas, M., & Epley, N. (2017). The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. *Psychological Science, 28*(12), 1745–1762.
Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... Seligman, M. E. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE, 8*(9) e73791.
Sedoc, J., Preoţiuc-Pietro, D., & Ungar, L. (2017). *Predicting emotional word ratings using distributional representations and signed clustering. Paper presented at the Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, Short Papers*.
Simpson, P. A., & Stroh, L. K. (2004). Gender differences: Emotional expression and feelings of personal inauthenticity. *Journal of Applied Psychology, 89*(4), 715–721.
Stuhlmacher, A. F., & Walters, A. E. (1999). Gender differences in negotiation outcome: A meta-analysis. *Personnel Psychology, 52*(3), 653–677.
Taleb, N. N. (2007). Black swans and the domains of statistics. *American Statistician, 61*(3), 198–200.
Tannen, D. (1990). *You just don't understand: Men and women in conversation.* New York: Morrow.
von Hippel, E. (2017). *Free innovation.* Cambridge, MA; London, England: MIT Press.
Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207.
Watson, C. (1994). Gender versus power as a predictor of negotiation behavior and outcomes. *Negotiation Journal, 10*(2), 117–127.
Wikimedia (2018). Community engagement insights: 2018 Report.
Williams, M. J., & Tiedens, L. Z. (2016). The subtle suspension of backlash: A meta-analysis of penalties for women's implicit and explicit dominance behavior. *Psychological Bulletin, 142*(2), 165–197.
Wood, W., & Eagly, A. H. (2010). Gender. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.). *Handbook of social psychology* (pp. 629–667). Hoboken, NJ: Wiley.
Wood, W., & Eagly, A. H. (2015). Two traditions of research on gender identity | SpringerLink. *Sex Roles, 73*(11–12), 461–473. https://doi.org/10.1007/s11199-015-0480-2.
Wu, L., Waber, B. N., Aral, S., Brynjolfsson, E., & Pentland, A. (2008). Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an IT configuration task. Available at SSRN 1130251.
Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744), 378–382.
Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036–1039.
Zhang, X. M., & Zhu, F. (2011). Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review, 101*(4), 1601–1615. https://doi.org/10.1257/aer.101.4.1601.
Zickuhr, K., & Rainie, L. (2011). Wikipedia, past and present: A snapshot of current Wikipedia users. Pew Internet & American Life Project.