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

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

[a] University of Pennsylvania, Department of Psychology, Solomon Labs, 3720 Walnut Street, Philadelphia, PA 19104-6018, Email: bhatiasu@sas.upenn.edu.

[b] UCLA, Anderson School of Management, 110 Westwood Plaza, Los Angeles, CA 90095, Email: jana.gallus@anderson.ucla.edu. *Corresponding author.

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 and in domains that are less controversial and more female-typed. Importantly, these differences disappear among people in positions of power: female authorities converge to the behavior of their male counterparts, such that the gender gaps in emotionality and willingness to converse on controversial content disappear. Sorting into topics according to their gender stereotypicality increases. We discuss mechanisms and implications for research on gender differences, leadership behavior, and conversational phenomena arising from such large-scale forms of knowledge production.

Keywords: Conversations, Gender, Power, Emotionality, Wikipedia

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 (e.g., 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 meted out? 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 be explained by underlying confounding power differentials (see, e.g., M. Johnson & Helgeson, 2002). They matter from an organizational perspective because they allow us to shed light on how conversations and verbal interactions influence emotional experience and motivation in organizations, but also on the influence of an important and growing subgroup of workers: women who advance to the core of collaborative knowledge production processes. Finally, answers to these questions can also be used to develop 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 the most comprehensive encyclopedia and a prime example of peer production systems where millions of voluntary contributors establish and curate a global public knowledge good (Benkler, 2006; Gallus, 2017; Lih, 2009; Zhang & Zhu, 2011). All of the production planning and quality management of the encyclopedia's articles takes place on "Wikipedia Talk pages", in the form of discussions between Wikipedia contributors. The data we analyze cover a period of more than 15 years and contain 166,322 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).

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) 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 on. 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.

Theory

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). Although a large number of studies have drawn their observations from

university students, it is possible to predict the 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 (1991) *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 given how stark the differences are 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 individuals' behavior through various mechanisms (Wood & Eagly, 2010). One important mechanism is social sanctions for counterstereotypical 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 are not co-located and work in large-scale online communities,

such as Wikipedia. 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 similar gender differences in emotionality to emerge among the general population of editors.

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 (Bear, Weingart, & Todorova, 2014; Schneider, Holman, Diekman, & McAndrew, 2016; Stuhlmacher & Walters, 1999; Tannen, 1990). Following this research, we would expect female editors to be less likely to engage in conversations about controversial topics. This is indeed in line with a recent analysis of self-reported survey responses by Wikipedia editors by Bear and Collier (2016), which suggests that, at least among occasional contributors, women are more afraid of facing conflict than men.

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). Since gender-incongruent situations may lead to increased anxiety, role conflict, backlash and avoidance (Bem & Lenney, 1976; Luhaorg & Zivian, 1995; Rudman, 1998), we expect to find a gender specific separation of labor, whereby female (male) editors from the general population are more likely to converse on female-typed (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 observe the same domain-specific gender difference to persist as we

consider editors in positions of power. This would stand in contrast to the previously discussed gender differences in emotionality and article controversiality, as further discussed below.

Evolution of 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 (M. 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 and 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).

Emotionality

Prior research suggests that the differences in male and female leaders' styles are merely "mild shading" (Eagly and Carli 2007: 127) 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.

Domain choice

Using a survey of directors, Adams and Funk (2012) find that several of the well-established gender differences that had been found for the general population no longer hold or are even reversed when looking at female and male directors. Notably, female directors in their broad sample are more risk tolerant and less security and tradition oriented than their male counterparts. Other differences previously found in the general population (e.g., women's being more benevolent and universally concerned) remain. Translated to our context, this suggests that women in positions of power may be as likely as men to engage in conversations about controversial content, where the outcome is more uncertain. 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 onto 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.

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 metes out 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 (Adam D Galinsky, Gruenfeld, & Magee, 2003), they may spend 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 (Adam D. 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 whose emotional tone or choice of domain are more like those of men get into 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, & Wood Brooks, 2015). 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); although see also Anderson, Hildreth, and Howland (2015), who suggest that the desire for status is universal). The type of woman who seeks to advance to the position of authority may thus on average be different from women in the general population.

The third mechanism is social selection by the majority-male population of editors (the most recent survey conducted by the Wikimedia Foundation (Wikimedia, 2018) puts the fraction of female editors on Wikimedia projects at 9%, which corresponds closely to earlier surveys (Glott, Schmidt, & Ghosh, 2010) – although readership rates are equal across genders (Zickuhr & Rainie, 2011)). This explanation points to demand-side factors (or interpersonal effects),

which may or may not be conscious. It is supported by evidence showing that female professionals who display anger are being conferred lower status than angry male professionals, both by male and female evaluators (Brescoll & Uhlmann, 2008). This 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 revealing physical characteristics, which can often produce biased evaluations (e.g., Brooks, Huang, Kearney, & Murray, 2014; Goldin & Rouse, 2000). Nevertheless, some editors choose pseudonyms that convey the person's gender. Even where this is not the case, social selection may occur based on behavioral differences, screening for women who act more like men.

A recent analysis by Fernandez-Mateo and Fernandez (2016) discusses the intricacies of distinguishing demand- and supply-side factors (including anticipatory effects), and proposes an original approach for doing so in the context of executive search. Disentangling 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.

Methods

Observing natural conversations, and doing so across different work domains and group constellations, is a difficult endeavor if the researcher's goal is to remain unobtrusive. Much

progress has been made recently by using new technologies, such as sociometric badges (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010; Wu et al., 2008). In this paper, we make use of observational data from a large corpus of Wikipedia editor conversations on Wikipedia Talk pages, which is where editors discuss their work.

Measuring emotionality

We examine the emotionality of editors' comments by using automated text analysis. Particularly, we apply lexicon-based methods, which measure the emotionality of a text based on the aggregate emotionality of its component words (see e.g. Humphreys & Wang (2017) for a detailed overview of this approach; and the special issues Harrow & Oswald (2016) and Kosinski & Behrend (2017) for a collection of representative papers using such methods). The specific lexicon we use involves 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. Although there are many other measures of emotionality that we could obtain from our text, we limit our analysis to valence and arousal as these two dimensions capture the majority of the variance in the structure of emotional experience (Russell, 1980). Furthermore, although there are other datasets that could be used to obtain valence and arousal ratings for words (Bradley & Lang, 1999), the Warriner dataset is the largest lexicon currently in existence, and contains participant-generated valence and arousal ratings for over 13,000 words. Importantly, this lexicon has been compiled by psychologists and is widely used in psychological research on emotion, language, memory, and decision making.

As outlined below, we apply automated text analysis to both individual comments and the articles that the comments pertain to. In both cases, we first lower-case the text (either the

comment or the article) and remove all punctuation. We then split the text into its component words by whitespace, and query the Warriner et al. lexicon for the valence and arousal ratings of each component word. Finally, we average the valence and arousal ratings for all words in the text that are also contained in the Warriner et al. lexicon, to obtain an aggregate measure of the valence and arousal of the text.

Dataset and variables

The dataset we employ contains 906,671 comments in 166,322 threads by 104,982 unique editors, spanning a wide range of topics over a period of more than 15 years (see Prabhakaran and Rambow 2016 for detailed information on the data source). We exclude comments for which we are unable to calculate valence and arousal (i.e., comments whose component words are not in the Warriner et al. lexicon), as well as threads in which there is only one participant. This yields 824,277 comments in 112,852 threads by 89,169 unique editors. The average valence in this set of comments is 5.64 (SD = 0.47, min = 1.26, max = 8.47) and the average arousal is 3.91 (SD = 0.31, min = 1.67, max = 7.74).

Our dataset 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, 151,210 (18.34%) are written by male editors, whereas 12,108 (1.47%) are written by female editors. The gender for 660,959 comments (80.19%) 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. There are a total of 6,033 unique male editors and 527 unique female editors in this set (92% and 8% of the gender-identifiable editors, respectively). We extract a number of other editor-level variables, in addition to gender: 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. Prior research (Kucuktunc et al., 2012) suggests that emotionality may decrease as community members gain experience, and we therefore want to control for it. In our dataset, 9.26% of comments are made by administrators, and there are a total of 1,349 unique administrators in this dataset. The average number of prior edits of all editors for whom we have edit data is 4,428 (SD = 28,340, min = 0, max = 3,734,324). 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).

We also consider various article-level variables. The two variables we focus on are whether or not the article is tagged as "controversial" (31% of all articles are controversial, though 63% of all comments are made on threads pertaining to controversial articles), as well as its gender-typedness. We compute the latter by calculating the relative number of male pronouns ("him", "he", "himself", "his") vs. female pronouns ("her", "she", "herself", "hers") in the article. There are a total of 1,144 unique articles which mention at least one male or female pronoun, and these articles have an average proportion of male and female pronouns of 80.20% and 19.80%, respectively. 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).

Importantly, we consistently 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. Overall, there are a total of 1,164 unique articles for which we are able to compute valence and arousal measures, with an average valence of 5.52 (SD = 0.24, min = 4.66, max = 6.17) and an average arousal of 4.11 (SD = 0.15, min = 3.49, max = 4.73). 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's 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 outcomes such as "fascism" and "nuclear war". Other low arousal articles include various uncontroversial topics, such as "scientific method".

A final set of controls involves thread-level variables. These are the number of unique editors commenting on the thread (mean = 3.62, SD = 2.24, min = 2, max = 54), the total number of comments on the thread (mean = 7.30, SD = 8.77, min = 2, max = 245), and the number of days between the first and the last comment on the thread (mean = 68, SD = 244, min = 0, max = 5,443). There are 21,361 threads (roughly 19% of our data) for which we cannot get a measure of thread time. We exclude these threads in most subsequent analyses. 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. Additionally, as the topic of each thread influences the valence and arousal of the component comments, we cluster comments by threads in the mixed-effects models used below. Table 1 provides descriptive statistics for all variables aggregated on the comment-level. Thus, for example, this table presents the average gender-typedness of the articles associated with each of the 824,277 comments (which is 0.84) rather than the average gender-typedness of the 1,144 unique articles (which is 0.80).

[Insert Table 1 here]

Results

Gender differences in domain choice

Before analyzing the emotionality of the conversations in our dataset, we examine whether there are systematic gender 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 (we also permit random effects on the thread-level to accommodate variability across threads).

Our results suggest that female editors from the general population indeed shy away from conversations about controversial content, and that there is sorting into conversational topics based on their gender stereotypicality. 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 more female pronouns), is lower in valence and higher in arousal, and if it is not tagged as being controversial. This suggests that women disproportionately comment on gender-congruent articles and ones that are about non-controversial topics, though they are more likely to comment on negative and arousing topics (we return to this later on in the manuscript). 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.

[Insert Table 2 here]

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. At the editor-level, we consider 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 use random effects in our regressions to control for thread- and user-level heterogeneity not captured by our control variables. As there are multiple variables being tested in each regression, we apply a Bonferroni correction for multiple comparisons. This yields a significance cutoff of $p = 0.0045$.

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. 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. 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.

Finally, comments occurring later on in a conversation have a significantly higher valence than comments occurring towards the beginning. There is a weak relationship ($p < 0.05$) with article controversiality, with comments on controversial articles being more negatively valenced, although this does not pass our Bonferroni correction for multiple comparisons.

[Insert Table 3 here]

Table 4 shows that comment arousal (which is a measure of the excitement or intensity of the comment) is not significantly influenced by editor-gender. It does, however, depend on the editor's prior experience, with editors with many prior edits writing relatively low-arousal comments (in line with, e.g., Kucuktunc et al. 2012). Comments also have significantly higher arousal if they belong to conversations about low valence and high arousal articles, and if they involve a large number of editors and unfold over a longer time span.

[Insert Table 4 here]

Finally, while this regression focused on comments for which we could identify the editor's gender, the article- and thread-level effects persist even when we examine all comments, including comments with non-identifiable editor-gender and prior edit count (see Tables A1 and A2 in the Online Appendix).

Moderators of the gender-valence relationship

In the previous section we observed a strong main effect of editor gender on comment valence, with comments from female editors displaying a significantly more positive valence than comments from male editors. In this section our goal is to understand the moderators of this tendency by considering interactions between gender and the ten other editor-level (admin-status, experience), article-level (valence, arousal, controversiality, gender-typedness), and thread-level variables (number of comments, number of unique editors, length of time, position in thread).

While our main interest was to analyze the interaction between gender and power, we also explored the nine other interactions and report the results for completeness. We therefore run ten separate regressions with the valence of the comment as the dependent variable, the variables examined in the prior section as independent variables, and an interaction term between gender and one of these ten variables. As above, our regressions include random effects on the user- and thread-level.

The outputs of the regressions are shown in Table 5. As can be seen, the only significant interaction with editor-gender is admin-status (i.e., power). 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 power.

[Insert Table 5 here]

Table 6 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. Thus, not only are there no gender differences in comment arousal, but gender also does not interact with other variables to influence comment arousal.

[Insert Table 6 here]

Evolution of 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.

Domain choice

Our analysis so far 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, uncontroversial articles, which are low in valence and high in arousal. Here we test how these relationships change as a function of the individual's power. As above, 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 7 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. However, there are no significant differences between these groups in the valence of the articles they comment on (although women do comment on more arousing articles). 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 comment on controversial articles. Additionally, female administrators are significantly more likely than male administrators to comment on low valenced articles.

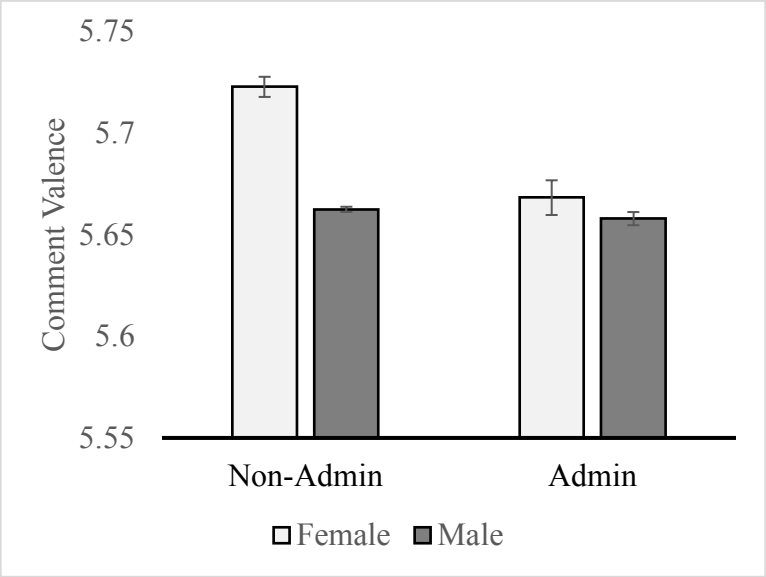
[Insert Table 7 here]

Emotionality

To develop an intuition of how the gender difference in emotionality changes as we consider individuals in positions authority, we first 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 analysis regresses comment valence on gender (1 if female), admin-status (1 if administrator, 0 otherwise), and their interaction, and does not control for the other editor-, article- and thread-level variables (Figure 1A). It nonetheless shows a robust interaction of gender and admin-status (β = -0.05, z = -4.39, p

< 0.001, 95%CI = [-0.07,-0.03]). 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, women in positions of power are indistinguishable from their male counterparts.

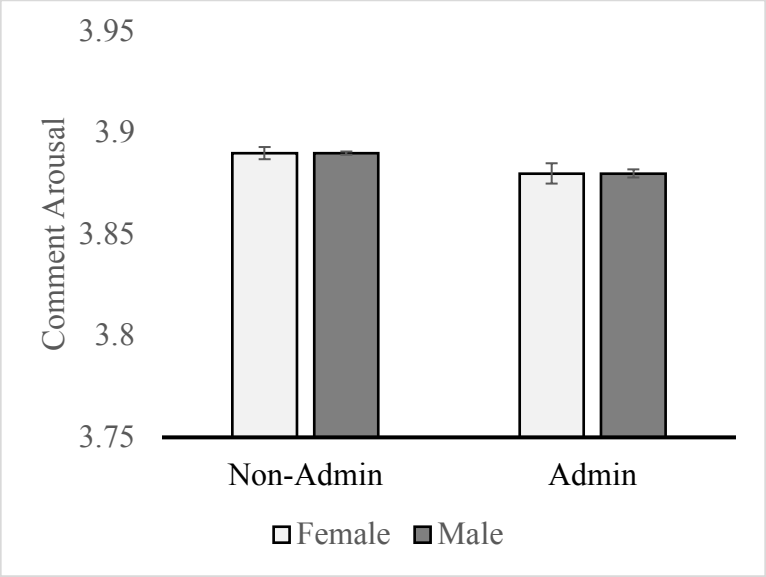
Figure 1A. Comment valence as predicted by gender and power



Error bars indicate 95%CIs.

We perform a similar analysis for comment arousal (Figure 1B). Unlike comment valence, we do not find a significant interaction effect (β = -0.0004, z = -0.07, p = 0.95, 95%CI = [-0.01, 0.01]), consistent with the findings from the prior section. Thus, it seems that it is only comment valence, and not arousal, that changes as a function of gender and power.

Figure 1B. Comment arousal as predicted by gender and power



Error bars indicate 95%CIs.

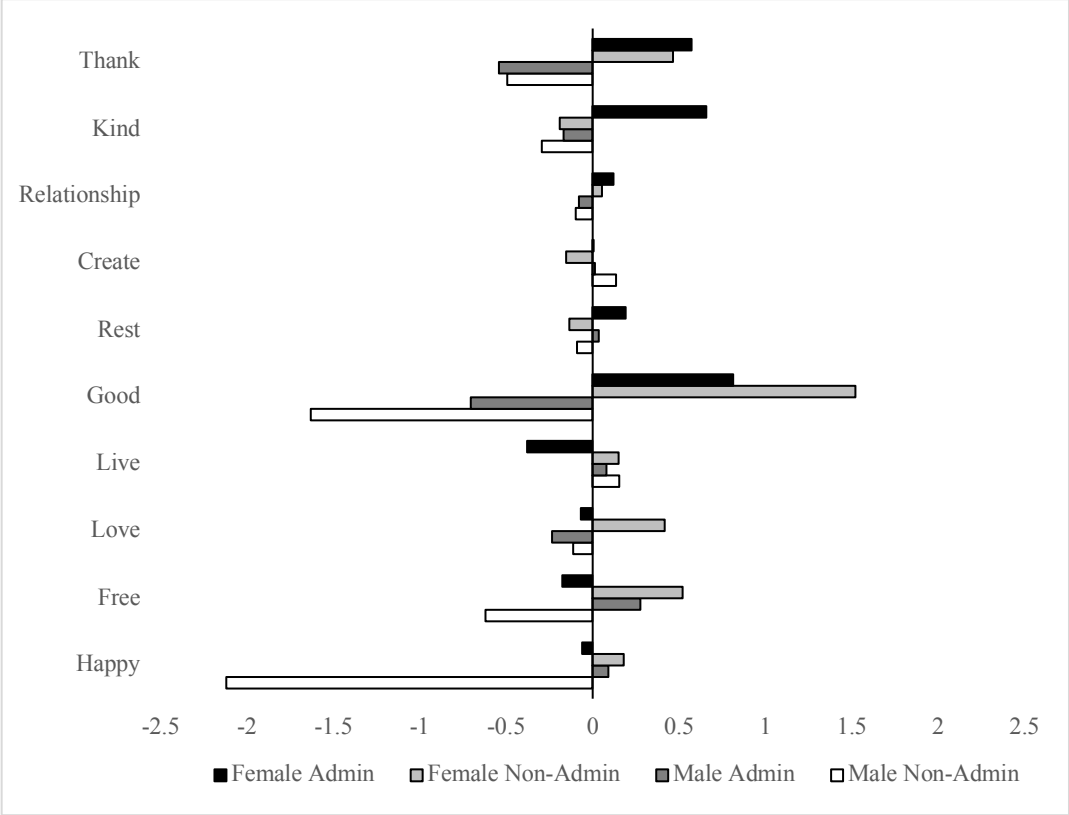
To gain a deeper understanding of the interaction effects observed for comment valence, we also perform an analysis of the valence of the words with the highest relative probabilities of being used by either of the four groups. The analysis only considers words that occur more than 1,000 times in the 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 yields spurious, highly-skewed probabilities that would bias our 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.

Figure 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 four of these words ("live", "relationship", "kind", "thank").

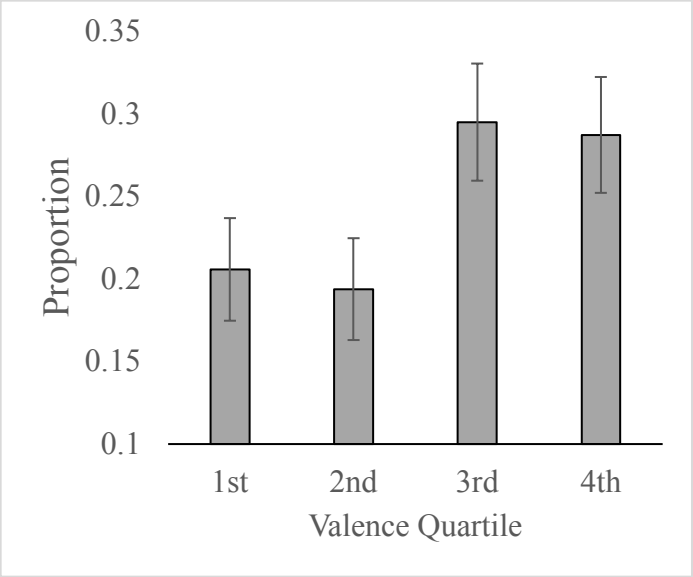
Figure 2. Relative probabilities of word usage for high valenced words



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). 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 Figure 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. Here we can see that low 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 high valenced words (3rd and 4th quartiles) do. This again shows that female non-administrators are relatively more likely to use highly valenced words, relative to the other three groups.

Figure 3: Relative word proportions in comments made by female non-administrators, as a function of valence



Error bars indicate 95% CIs.

Finally, we examine the words that are the most likely to be used in comments made by female non-administrators, irrespective of their valence. These are words with the highest relative probabilities of occurrence. Figure 4 displays the resulting word cloud. It suggests that female non-administrators are typically politer, with words such as “agree”, “thanks”, “like”, and “good” being especially likely to be used by these editors relative to male non-administrators, male administrators, and female administrators.

Figure 4: Words most likely to be used by female non-administrators

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 administrative 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 imply that there are only twenty-five 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 (article valence, arousal, controversiality and gender-typedness; user log-edit count; comment order; thread number of comments, users and log-thread length in days) as well as random effects on the user- and thread-level. We find that there is absolutely no difference between the comment valence of female non-administrators who eventually become administrators and the comment valence of female non-administrators who do not ($p = 0.99$). This suggests that female administrators do not differ in their conversational style (emotionality) 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 became administrators). We run this regression with the controls discussed above and include random effects on the user- and thread-level. While this analysis is limited as there are only twenty-five editors for whom we can obtain this data, we do find a directional drop in comment valence as women come to occupy the position of authority ($\beta = -0.05$, $SE = 0.04$, $z = -1.29$, $p = 0.19$, $95\%\ CI = [-0.12,0.03]$). Both of these analyses are thus 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. Analyzing these mechanisms in more detail is an important and promising avenue for future research.

Discussion and conclusion

Our analysis yields several implications for research on gender differences, leadership behavior, and conversational phenomena arising from modern forms of knowledge production, where selection into different work domains is voluntary and not mandated or motivated by pecuniary incentives.

We show that there are significant gender differences in people's *conversational styles* (specifically, in their emotionality) and *domain choices* (controversiality and gender-typedness). Importantly, once we look up the organizational hierarchy to individuals in positions of power, these differences disappear: female and male authorities are just as emotional 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 one notable exception is that the gender specific separation of labor, or sorting into conversational topics based on their gender stereotype, seems to increase. This may be explained by differences in previously accumulated knowledge and expertise, which editors can leverage once they become administrators.

Our finding of the disappearance of the gender gap among people in positions of power is in line with previous work in the gender literature (Croson and Gneezy 2009), which 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., J. E. Johnson and Powell (1994), who compare trained and untrained subpopulations, as well as Masters and Meier (1988) and Birley (1987) who focus on entrepreneurs, a population that is more akin to ours). We find such convergence even in a population of untrained individuals (i.e., Wikipedia administrators presumably did not undergo formal management education). We do, however, also find suggestive evidence that the position of authority may have had an effect on the disappearance of the gender gap. Other possible mechanisms are self- or social selection (i.e., supply- and demand-side factors). Analyzing the mechanisms behind the convergence, including how they interact, is an important avenue for future research (Fernandez-Mateo & Kaplan, 2018).

Similarly, follow-up work should further investigate the role played by the mode of collaboration, comparing the more traditional, small-scale, and co-located team production settings (Leavitt, 1989) to the novel conversational phenomena that arise from large-scale collaborations among self-governing "peers" who rarely if ever interact with one another in face-to-face contexts. Our study focuses on the latter. It is likely that different conversational dynamics unfold where gender cues are more salient and where nonverbal behaviors may be used by women in an attempt to mitigate adverse consequences from leaderlike behaviors (Carli, LaFleur, & Loeber, 1995; Eagly & Karau, 2002; Hall, Coats, & LeBeau, 2005).

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. This research could usefully advance our understanding of how the expression of emotions and attempts to conform to gender- and leadership-role specific

display rules in such virtual contexts correlates with actually experienced feelings (e.g., of authenticity or belonging) and well-being (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 has helped lay the groundwork for such an analysis.

References

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. doi: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. doi:doi: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.

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.

Brescoll, V. L., & Uhlmann, E. L. (2008). Can an angry woman get ahead? Status conferral, gender, and expression of emotion in the workplace. *Psychological Science, 19*(3), 268-275.

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* (Vol. 2, 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. doi: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 Jr., P. T., 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* (Vol. 2). London: Sage.

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. doi:10.1287/mnsc.2015.2315

Fernandez-Mateo, I., & Kaplan, S. (2018). Gender and Organization Science: Introduction to a Virtual Special Issue. *Organization Science, 29*(6), 1229-1236. doi:10.1287/orsc.2018.1249

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. doi: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.
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., & Wood Brooks, A. (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. doi: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.
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.
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*. doi:http://dx.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. doi:doi:10.1111/1471-6402.00063
Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review, 110*(2), 265-284.
Kosinski, M., & Behrend, T. (2017). Editorial overview: Big data in the behavioral sciences. *Current Opinion in Behavioral Sciences, 18*, iv-vi. doi: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.

Laniado, D., Kaltenbrunner, A., Castillo, C., & Morell, M. F. (2012). *Emotions and dialogue in a peer-production community: the case of Wikipedia.* Paper presented at the Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration.

Leavitt, H. J. (1989). Suppose we took groups seriously. In H. J. Leavitt, L. R. Pondy, & D. M. Boje (Eds.), *Readings in managerial psychology* (4 ed.). Chicago and London: University of Chicago Press.

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. doi:10.5465/19416520802211628

Masters, R., & Meier, R. (1988). Sex differences and risk-taking propensity of entrepreneurs. *Journal of small business management, 26*(1), 31.

Mulac, A. (1998). The gender-linked language effect: Do language differences really make a difference? In D. J. D. Canary, K. (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.

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.

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* (3 ed. Vol. 356). 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. doi:10.1111/pops.12268

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.

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*.

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* (Vol. 5, 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. doi:10.1007/s11199-015-0480-2

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686-688.

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*.

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. doi: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*.

Table 1: Descriptive statistics for key variables

	Mean	SD	Min	Max
Comment-level variables				
Comment valence	5.64	0.47	1.26	8.47
Comment arousal	3.91	0.31	1.67	7.74
User-level variables				
Male gender	18.34			
Female gender	1.47			
Gender not-determined	80.19			
Administrator	9.26			
Not administrator	90.74			
# Prior edits (log transformed)	8.27	2.54	0.00	15.13
Article-level variables				
Controversial	63.07			
Non-controversial	36.93			
Gender-typedness	0.84	0.24	0.00	1.00
Article valence	5.50	0.21	4.67	6.17
Article arousal	4.11	0.12	3.49	4.73
Thread-level variables				
# Unique editors	5.55	3.76	2.00	54.00
Time difference (log transformed)	1.99	1.86	0.00	8.60

Notes: The statistics are aggregated on the comment-level. Thus, for example, the table presents the average gender-typedness (proportion of male relative to female pronouns) of the articles associated with each of the 824,277 comments (which is 0.84) rather than the average gender-typedness of the 1,144 unique articles (which is 0.80 – see main text).

Table 2: Logistic regression 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
Controversial	-0.13	0.06	-2.40	0.02	-0.24	-0.02
Gender-typedness	-2.59	0.10	-26.13	0.00	-2.78	-2.39
Valence	-0.50	0.13	-3.89	0.00	-0.75	-0.25
Arousal	0.59	0.22	2.64	0.01	0.15	1.02

Note: Gender-typedness is the relative proportion of male to female pronouns in the article.

Table 3: Regression predicting comment valence from various user-, article-, thread-, and comment-level variables

	Coef.	S.E.	z	P>|z|	95%-L	95%-H
User female	0.054	0.006	8.56	0.000	0.041	0.066
User admin at post	-0.012	0.005	-2.52	0.012	-0.021	-0.003
User log edit count	0.002	0.001	2.43	0.015	0.000	0.004
Article valence	0.324	0.008	39.39	0.000	0.308	0.341
Article arousal	-0.033	0.015	-2.27	0.023	-0.062	-0.005
Article controversial	-0.008	0.004	-2.26	0.024	-0.015	-0.001
Article gender-typedness	-0.016	0.007	-2.30	0.022	-0.030	-0.002
Thread number of comments	-0.002	0.000	-7.51	0.000	-0.002	-0.001
Thread number of users	-0.004	0.001	-4.77	0.000	-0.006	-0.002
Thread log time difference	-0.004	0.001	-4.44	0.000	-0.006	-0.002
Comment ID	0.001	0.000	5.87	0.000	0.001	0.002

Note: Random effects on thread- and user-level. Bonferroni correction for multiple comparisons yields a significance cutoff of $p = 0.0045$.

Table 4: Regression predicting comment arousal from various user-, article-, thread-, and comment-level variables

	Coef.	S.E.	z	P>|z|	95%-L	95%-H
User female	-0.006	0.004	-1.46	0.144	-0.014	0.002
User admin at post	0.000	0.003	0.07	0.944	-0.006	0.006
User log edit count	-0.004	0.001	-6.80	0.000	-0.006	-0.003
Article valence	-0.055	0.005	-10.66	0.000	-0.065	-0.045
Article arousal	0.295	0.009	32.05	0.000	0.277	0.313
Article controversial	0.003	0.002	1.31	0.191	-0.001	0.007
Article gender-typedness	0.011	0.004	2.51	0.012	0.002	0.020
Thread number of comments	0.000	0.000	-1.10	0.273	0.000	0.000
Thread number of users	0.003	0.001	5.29	0.000	0.002	0.004
Thread log time difference	0.003	0.001	5.56	0.000	0.002	0.005
Comment ID	0.000	0.000	0.32	0.751	0.000	0.000

Note: Random effects on thread- and user-level. Bonferroni correction for multiple comparisons yields a significance cutoff of $p = 0.0045$.

Table 5: Interaction effects between user-gender and other possible predictors of comment valence, from ten separate regressions

	Coef.	S.E.	z	P>|z|	95%-L	95%-H
User admin at post	-0.04	0.01	-2.92	0.00	-0.07	-0.01
User log edit count	0.00	0.00	-1.36	0.17	-0.01	0.00
Article valence	-0.03	0.03	-1.02	0.31	-0.09	0.03
Article arousal	-0.07	0.05	-1.39	0.16	-0.17	0.03
Article controversial	-0.01	0.01	-0.95	0.34	-0.03	0.01
Article gender-typedness	0.02	0.02	0.83	0.41	-0.02	0.06
Thread number of comments	0.00	0.00	-0.14	0.89	0.00	0.00
Thread number of users	0.00	0.00	-0.84	0.40	-0.01	0.00
Thread log time difference	0.00	0.00	-0.41	0.68	-0.01	0.01
Comment ID	0.00	0.00	0.88	0.38	0.00	0.00

Notes: Gender is coded such that 1 indicates a female user (0 if male). Each of the ten regressions includes our standard set of controls as well as random effects on the user- and thread-level.

Table 6: Interaction effects between user-gender and other possible predictors of comment arousal, from ten separate regressions

	Coef.	S.E.	z	P>|z|	95%-L	95%-H
User admin at post	0.01	0.01	0.96	0.34	-0.01	0.03
User log edit count	0.00	0.00	-0.62	0.54	-0.01	0.00
Article valence	0.04	0.02	2.03	0.04	0.00	0.07
Article arousal	0.06	0.03	1.93	0.05	0.00	0.12
Article controversial	0.01	0.01	0.68	0.49	-0.01	0.02
Article gender-typedness	-0.03	0.01	-2.26	0.02	-0.06	0.00
Thread number of comments	0.00	0.00	-1.06	0.29	0.00	0.00
Thread number of users	0.00	0.00	-1.00	0.32	0.00	0.00
Thread log time difference	0.00	0.00	-0.06	0.95	0.00	0.00
Comment ID	0.00	0.00	-2.21	0.03	0.00	0.00

Notes: Gender is coded such that 1 indicates a female user (0 if male). Each of the ten regressions includes our standard set of controls as well as random effects on the user- and thread-level.

Table 7: Logistic regression predicting whether the originator of the comment is female, using various article-level variables, for non-administrators and administrators, respectively

Not administrator	**Coef.**	**S.E.**	**z**	**P>|z|**	**95%-L**	**95%-H**
Controversial	-0.36	0.06	-5.86	0.00	-0.49	-0.24
Gender-typedness	-2.25	0.11	-20.15	0.00	-2.47	-2.04
Valence	-0.24	0.14	-1.65	0.10	-0.51	0.04
Arousal	0.92	0.25	3.74	0.00	0.44	1.41
Administrator	**Coef.**	**S.E.**	**z**	**P>|z|**	**95%-L**	**95%-H**
Controversial	0.38	0.15	2.53	0.01	0.09	0.68
Gender-typedness	-4.26	0.27	-15.78	0.00	-4.79	-3.73
Valence	-1.43	0.34	-4.18	0.00	-2.10	-0.76
Arousal	-0.65	0.61	-1.06	0.29	-1.84	0.55

Note: Gender-typedness is the relative proportion of male to female pronouns in the article.

Online Appendix

Table A1, Robustness check: Regressions predicting comment valence using various user-, article-, thread-, and comment-level variables (not restricted to users for whom we have gender information)

	Coef.	S.E.	z	P>|z|	95%-L	95%-H
User admin at post	0.012	0.003	4.63	0.000	0.007	0.017
User log edit count	0.002	0.000	6.53	0.000	0.001	0.003
Article valence	0.354	0.004	86.89	0.000	0.346	0.362
Article arousal	-0.074	0.007	-10.44	0.000	-0.087	-0.060
Article controversial	-0.005	0.002	-3.00	0.003	-0.009	-0.002
Article gender-typedness	-0.018	0.003	-5.20	0.000	-0.025	-0.011
Thread number of comments	-0.002	0.000	-14.63	0.000	-0.002	-0.001
Thread number of users	-0.003	0.000	-6.18	0.000	-0.004	-0.002
Thread log time difference	-0.005	0.000	-10.85	0.000	-0.006	-0.004
Comment ID	0.001	0.000	10.02	0.000	0.001	0.001

Table A2, Robustness check: Regressions predicting comment arousal using various user-, article-, thread-, and comment-level variables (not restricted to users for whom we have gender information)

	Coef.	S.E.	z	P>|z|	95%-L	95%-H
User admin at post	-0.008	0.002	-4.73	0.000	-0.011	-0.005
User log edit count	-0.004	0.000	-18.71	0.000	-0.004	-0.003
Article valence	-0.045	0.003	-17.37	0.000	-0.051	-0.040
Article arousal	0.320	0.005	70.55	0.000	0.311	0.329
Article controversial	0.006	0.001	4.98	0.000	0.003	0.008
Article gender-typedness	0.010	0.002	4.30	0.000	0.005	0.014
Thread number of comments	0.000	0.000	-2.89	0.004	0.000	0.000
Thread number of users	0.003	0.000	11.22	0.000	0.003	0.004
Thread log time difference	0.004	0.000	12.44	0.000	0.003	0.004
Comment ID	0.000	0.000	1.56	0.118	0.000	0.000