No Longer Conforming to Stereotypes? Gender, Political Style, and Parliamentary Debate in the UK *

Lotte Hargrave *University College London*
Jack Blumenau *University College London*

Research on political style suggests that where women make arguments that are more emotional, empathetic, and positive, men use language that is more analytical, aggressive, and complex. However, existing work does not consider how gendered patterns of style vary over time. Focusing on the UK, we argue that pressures for female politicians to conform to stereotypically 'feminine' styles have diminished in recent years. To test this argument, we describe novel quantitative text analysis approaches for measuring a diverse set of styles at scale in political speech data. Analysing UK parliamentary debates between 1997 and 2019, we show that female MPs' debating styles have changed substantially over time, as women in parliament have increasingly adopted stylistic traits that are typically associated with 'masculine' stereotypes of communication. Our findings imply that prominent gender-based stereotypes of politicians' behaviour are significantly worse descriptors of empirical reality now than they were in the past.

Keywords: gender; legislative politics; debate; style; stereotypes; text-as-data

Word count: 9974

*__This version__: June 18, 2021. We thank Lucy Barnes, Jennifer Hudson, Markus Kollberg, Tone Langengen, Ben Lauderdale, Rebecca McKee, Alice Moore, Tom O'Grady, Meg Russell, Jess Smith, and Sigrid Weber for helpful conversations and insightful comments. We are also grateful to participants at working group seminars at University College London, PSA Polmeth 2020, and the EGEN Summer Working Group 2020. Finally, we thank Agnes Magyar and Alicia O'Malley for excellent research assistance.

Introduction

Have incentives for politicians to conform to gender stereotypes diminished over time? In addition to the fact that female and male politicians speak about systematically different sets of political issues (Bäck and Debus, 2019; Catalano, 2009), another dimension on which gendered differences are said to arise is regarding argumentation style. Gendered communication styles are thought to be rooted in stereotypes that create social expectations for women to act "like women" and men "like men" (Eagly and Wood, 2012). If politicians internalise these expectations before entering politics, or if voters punish them for contravening gender stereotypes (Bauer, 2015*a*; Boussalis et al., 2021; Cassese and Holman, 2018), legislators are likely to engage in gender-role consistent behaviour, and we should expect systematic differences in the political styles that female and male politicians adopt. Empirical evidence supports this view: compared to their male colleagues, female politicians' speeches are more emotional (Dietrich, Hayes and O'Brien, 2019), less complex and jargonistic (Coates, 2015), less repetitive (Childs, 2004), less aggressive (Kathlene, 1994), and use different types of evidence to support their arguments (Hargrave and Langengen, 2020).

We contribute to this literature by evaluating the degree to which gendered differences in political style vary over time. Consistent with work that views stereotypes as dynamic constructs (Eagly and Wood, 2012; Diekman and Eagly, 2000), we argue that over the past 25 years in the UK, where we situate our study, several factors are likely to have decreased the degree to which UK MPs (and especially women) will conform to stereotype-consistent behaviours in parliament. First, politicians are drawn from a broader population which has itself diverged from stereotypical communicative styles in recent years. Second, changes to the social roles played by women in public life, and in politics, have reduced the validity of gender stereotypes in the eyes of the pub-

lic. Consequently, we argue that voters are less likely to sanction female legislators for gender-incongruent behaviours now than in the past. Finally, the increased prominence of women in parliament and leadership roles is also likely to reduce the degree to which female politicians internalise expectations that they need to behave in "feminised" ways. Together, these arguments lead to a central behavioural prediction which we test empirically: that UK MPs will conform less to gender stereotypes now than in the past, and that women in particular will adopt styles that are further from feminine stereotypes over time.

To evaluate this expectation, we examine politicians' styles as they manifest in one prominent legislative activity: parliamentary debates. We conceive of *debating* style as a characteristic of speech which is distinct from its content. Intuitively, the style of a speech reflects the manner in which an argument is delivered. In social psychology, women's "communal" styles are thought to be marked by higher levels of emotionality, positivity, empathy, and warmth, while men's "agentic" styles are thought to be marked by higher levels of aggression, logic, and confidence (Eagly and Wood, 2012; Schneider and Bos, 2019). In political science, these concepts have been operationalised using a diverse set of indicators. We survey both literatures and identify eight styles that reflect the ideas of communality and agency, and are also – in principle – detectable in the speeches politicians deliver. The eight styles which we use as the basis of our empirical analysis are human narrative, affect, positive emotion, negative emotion, factual language, aggression, complex language, and repetition.

In addition to our substantive argument, our paper also provides a methodological contribution to the measurement of style in legislative settings (Boussalis et al., 2021; Dietrich, Hayes and O'Brien, 2019). Our goal is to construct measures that closely approximate the conceptual definitions of each of the styles that we highlight in our review of the literature. For some styles, we use existing quantitative text analysis measures

that have been extensively validated in other settings (e.g. Kincaid et al., 1975). For others, we develop new measures that combine traditional dictionary approaches with a word-embedding model. Our strategy overcomes limitations of standard dictionary approaches as it enables us to detect styles as they manifest *in the specific context of parliamentary debate*. Evidence from a human validation task shows that our measures significantly outperform standard measures that have been used extensively in previous research on political style.

We apply our measures to nearly half a million speeches delivered in the House of Commons between 1997 and 2019 and report three main findings. First, in the early parts of our study period, we document patterns of style which are broadly consistent with expectations from the literature on gender stereotypes. Male MPs' speeches are marked by substantially higher levels of aggression and complexity, while female MPs' speeches are considerably more emotional, positive, and make greater use of human narrative. Second, and crucially, we find that these differences have reduced dramatically in recent years. In six out of eight styles, MP gender explains less variation in style use between individuals at the end of our study period than at the beginning. For cases where we report *diverging* behaviour over time, these stylistic shifts run counter to gender-role expectations. Third, we show that the evolving variation in style use has primarily resulted from women's decreasing use of communal styles, and increasing use of agentic styles, over time.

Our work builds upon a large literature (cited above) on gender differences in politicians' communication styles, the vast majority of which considers whether gender stereotypes accurately capture real behavioural differences at fixed points in time. By contrast, we show that the descriptive validity of prominent stereotypes of how men and women communicate is considerably lower in the contemporary House of Commons than it was in the past.

Our findings also have important implications beyond scholarly accounts of legislative politics. Prescriptive stereotypes for how men and women *ought* to behave are ultimately rooted in our collective understanding of how we expect men and women *will* behave (Eagly and Wood, 2012). In politics, these prescriptions can form the basis of voter judgements about the behaviour of male and female politicians. Documenting when behavioural shifts run counter to gender-based stereotypes is important, then, because it potentially undermines these prescriptions, and could thereby diminish the degree to which women will be subject to penalties for failing to conform to stereotypical expectations (Cassese and Holman, 2018; Ditonto, 2017).

Additionally, our results cast doubt on the idea that the election of more women into office will automatically result in a less adversarial and more deliberative culture in Westminster.[1] At least in the context of parliamentary debate, our findings suggest that such effects are unlikely to materialise because they are based on an outdated assumption about the distinctiveness of female MPs' political styles. In addition, by placing hopes of cultural change on newly elected women, proponents of these views may also be setting female politicians up to fail if their increased presence does not result in a "better" political culture. To the extent that cultural change of this sort is a desideratum of modern parliamentary politics, our results suggest that hopes of affecting such change should not rely on the presumption that newly elected women will conform to anachronistic stereotypes, and that purposive reforms to parliamentary practices may instead be necessary.

Gender, stereotypes, and debating style

Why might men and women in parliament employ different debating styles? Gender role theory (Eagly and Karau, 2002) suggests that gender stereotypes concerning the typical

[1] See, for example, Designing a new parliament with women in mind, *Democratic Audit*, 29th July 2016.

conduct of men and women can affect behaviour via two main channels. First, repeated exposure to stereotypes from a young age may lead men and women to *internalise* expectations relevant to their genders, which then become self-imposed standards against which they regulate their own behaviour (Eagly and Wood, 2012). Second, descriptive stereotypes (e.g., the perceived tendency for women to be emotional) are often thought to lead to prescriptive stereotypes (e.g., the view that women *should* be emotional), the violation of which leads to the imposition of *social sanctions* by others which further incentivise conformity with gender-based norms (Brescoll and Uhlmann, 2008).

Female politicians are especially subject to pressures to conform to role-consistent behavioural standards, as voters punish women for displaying behaviour that counters feminine stereotypes. Voters form gender-biased impressions of candidates (Bauer, 2015*a*), and penalise women for appearing to be too ambitious (Okimoto and Brescoll, 2010) or negative (Cassese and Holman, 2018), while rewarding them for displays of happiness (Boussalis et al., 2021). These penalties are more acute when the campaign environment is characterised by "masculine" issues (Holman, Merolla and Zechmeister, 2016), and are more commonly applied by low-attention voters (Bauer, 2015*b*) and voters with sexist attitudes (Mo, 2015) or aggressive personalities (Bauer, Kalmoe and Russell, 2021). By contrast, voters are less sensitive to role-inconsistent behaviour by male politicians (Okimoto and Brescoll, 2010).

However, as political leaders, legislators are also expected to display behaviours consistent with *leadership* stereotypes. Men's historical occupation of leadership positions means that leadership stereotypes have been shaped by traditional "masculine" traits such as being assertive, competitive, and outgoing (Koenig et al., 2011). The congruence between leadership stereotypes and masculine stereotypes therefore poses little challenge for male politicians to conform to both sets of expectations. By contrast, if women seek to conform to *leadership* stereotypes, they may risk incurring penalties for violat-

ing *feminine* stereotypes (Bauer, 2017; Eagly and Karau, 2002). Women must therefore attempt to balance a complicated array of behavioural expectations in a way that men do not.

Political scientists have evaluated whether these incentives induce male and female politicians to adopt systematically different political styles. We focus on one aspect of legislative activity where differences are likely to become manifest: in political speech. On which dimensions of style should we expect gender differences? Of central concern in the social psychology literature is the distinction between *communal* characteristics of style, which are associated with women, and *agentic* characteristics which are associated with men (Schneider and Bos, 2019). These labels are heuristics for clusters of behavioural attributes, where communal characteristics are said to include being "affectionate, helpful, kind, sympathetic, interpersonally sensitive, nurturant, and gentle", while agentic characteristics include being "aggressive, ambitious, dominant, forceful, independent, self-sufficient, and self-confident" (Eagly and Karau, 2002, 574). By surveying the large empirical literature on gendered styles in political science, we identified eight styles that are representative of "communal" or "agentic" behaviour and which previous work had either shown to be associated with male or female use in politics, or that previous work had *expected* to be associated with gender differences. We use these styles as the basis of our empirical analysis below.

We identified three "communal" characteristics of style that are typically associated with women. First, women are said to make greater use of **human narrative** through reliance on personal experience, analogies, and anecdotes in their speeches (Blankenship and Robson, 1995). This idea is supported both by politicians' testimonies (Childs, 2004), and qualitative studies of political speech (Hargrave and Langengen, 2020). Second, women are also thought to make greater use of emotional language or **affect** (Huddy and Terkildsen, 1993), and there is clear evidence that women's language exhibits greater

overall emotionality than men's (Dietrich, Hayes and O'Brien, 2019; Jones, 2016). Third, and more specifically, women have been found to use more **positive emotion**, such as expressing happiness, in their political speeches than men (Boussalis et al., 2021; Yu, 2013).

We identified five "agentic" characteristics of style that are typically associated with men. First, men are thought to rely more on **fact-based** language, which is more "analytical, organised and impersonal" and relies more on statistical evidence (Jamieson, 1988, 76). In the UK, MPs suggest that male politicians pay greater attention to "scientific research" (Childs, 2004, 181), though in other settings there is evidence that women may use more factual language (Hargrave and Langengen, 2020).

Second, male politicians' speech is also thought to feature higher levels of linguistic **complexity**, marked by formalistic and jargonistic word use (Childs, 2004), while women are thought to be more accessible and clear (Coates, 2015). Third, men are also thought to be more **repetitive** (Dahlerup, 1988; Childs, 2004, 184), and, fourth, more **aggressive**, whereas women are said to avoid combative and aggressive styles (Brescoll and Uhlmann, 2008; Kathlene, 1994), and empirical work suggests women are significantly less adversarial than men in parliamentary debate (Grey, 2002; Hargrave and Langengen, 2020). Fifth, women are thought to avoid the use of excess **negative emotion** for fear of backlash (Cassese and Holman, 2018), while men are thought to make greater use of negativity (Brooks, 2011).

We summarise these eight styles in table 1. We provide a short definition and categorise each as either "communal" or "agentic" according to our discussion above. The expectations that derive from the literature on gendered stereotypes suggests that women will be more likely to use communal debating styles, and men more likely to use agentic debating styles.

Table 1: Political styles

Style	Type	Definition
Human Narrative	Communal	Use of personal examples or experiences; stories of other people; constituency stories; illustrative examples; referring to individual people, including other MPs.
Affect	Communal	Use of emotive language, which might be either positive or negative; such as expressing criticism, praise, disapproval, pride, empathy or fear.
Positive Emotion	Communal	Use of positive emotional language, which might include expressing empathy, praise, celebration or congratulations.
Fact	Agentic	Use of numbers, statistics, numerical quantifiers, figures and empirical evidence.
Complexity	Agentic	Use of jargonistic, complicated and elaborate language that is challenging to understand.
Repetition	Agentic	Repeated use of the same words or phrases.
Aggression	Agentic	Use of aggressive or combative language, which might include criticisms or insults; language that suggests forceful action; or declamatory or adversarial language.
Negative Emotion	Agentic	Use of negative emotional language, which might include expressing fear, anxiety, unpleasantness, sadness or disapproval.

Dynamic gender stereotypes

Despite this rich literature, few studies consider whether politicians' conformity with gender stereotypes has changed over time.[2] This is surprising, as gender role theorists emphasise that the content and strength of stereotypes are dynamic (Diekman and Eagly, 2000; Eagly and Wood, 2012). These accounts posit that gender stereotypes arise from men and women's historical occupation of different social roles which are associated with different characteristics. For instance, because women have traditionally occupied roles in which they provide care to others, "caring" as a characteristic became stereotypic

[2]Though see Jones (2016) for a case study of the evolution of Hillary Clinton's style, and Grey (2002) who demonstrates that female MPs in New Zealand are increasingly aggressive over time.

of women. However, as the distribution of men and women into different roles changes, so too will the characteristics associated with the stereotypes themselves such that the stereotype of women will be marked by "increasing masculinity and… decreasing femininity" (Diekman and Eagly, 2000, 1173). Building on this logic, we argue that recent changes in the roles played by women in both politics and the broader public are likely to have weakened traditional gender stereotypes in the UK, and we therefore expect a decline in the degree to which MPs, and especially women, will adopt styles that are congruent with the stereotypes described above.

First, politicians are selected from a broader population, which has itself diverged from gender-stereotypical behaviours over time. In most advanced economies in recent decades, women's traditional role as care-givers has declined, and women's educational attainment, participation in the workforce, and occupancy of senior management positions have increased (Sayer, Bianchi and Robinson, 2004; Diekman and Goodfriend, 2006, 370). As societal gender roles have changed, women in the public have come to demonstrate increasingly agentic behaviours across a wide set of contexts and countries (Twenge, 2001; Leaper and Ayres, 2007, 357). As politicians are likely to reflect the characteristics of the population from which they are drawn, if women in the UK are now more agentic on average, we should expect these changes to be reflected in the behaviour of politicians too.

Second, changing social roles have affected public perceptions of the validity of traditional gender stereotypes. Women in general are perceived as more agentic now than in the past (Eagly et al., 2020; Sendén et al., 2019), and while attributes associated with men have remained relatively stable, masculine characteristics are increasingly ascribed to women (Diekman and Eagly, 2000). Do changes in attitudes regarding the content and validity of stereotypes mean that voters are less likely to punish counter-stereotypic behaviours? As we reported above, several papers document voters' tendency to pun-

ish female candidates for displaying agentic traits. However, we are not aware of any existing empirical literature which tracks the extent to which politicians are punished by voters for contravening stereotypes *over time*. Several more recent studies suggest that voters do not *always* punish female politicians for violating feminine stereotypes (Brooks, 2011; de Geus et al., 2021; Saha and Weeks, 2020), but these papers again only provide evidence from a single point in time.

There is, however, evidence that the public have become less likely to endorse traditional gender stereotypes over time. As women's position in the labour market has improved, support for traditional gender norms and associated stereotypes has eroded both in the UK and further afield (Inglehart and Norris, 2003; Twenge, 1997; Seguino, 2007). In the UK, voters have come to hold substantially more gender-egalitarian attitudes between the mid-1980s and the present (Taylor and Scott, 2018). Further, between 1990 and 2010, voters in Western Europe, including in the UK, have become significantly less likely to agree with the traditional division of social roles performed by men and women (Shorrocks, 2018). This latter finding is particularly relevant given that it is the association of men and women with particular social roles that is at the heart of theories of gender stereotypes (Eagly and Karau, 2002). It therefore seems likely that as voters have become less willing to endorse gender stereotypes, they also will apply fewer sanctions to politicians who transgress such stereotypes. Consequently, as Mo (2015, 360) argues, "gender attitudes in the electoral process remain consequential, but have grown subtler". To the extent that politicians in the UK are sensitive to the expectations of voters, then, changing voter attitudes about stereotypes will likely have reduced pressures on female politicians to conform to gender-stereotypic behaviours over time.

Third, the dramatic shifts in the roles that women play in *political* life in recent decades might also reduce the degree to which female politicians conform to traditional gender stereotypes. In the House of Commons, women held just 18% of seats in 1997,

but this increased to 32% by 2019 (IPU, 2020). Moreover, female politicians in the UK now occupy more high-powered positions within the legislative hierarchy (Blumenau, 2021). As women enter politics at a higher rate, role theory predicts that female politicians will come to be seen as possessing more masculine characteristics (Diekman et al., 2005), and the increasing prevalence of women in leadership has been shown to reduce the degree to which communal qualities are ascribed to women (Dasgupta and Asgari, 2004). As Diekman et al. (2005, 212) argue, "women's increased representation as elected officials and government employees should foster the ascription to women of traditionally masculine qualities."

Accordingly, in addition to a general tendency for stereotypes of women to become more oriented towards agentic characteristics in recent years, female politicians *specifically* may have become associated with more masculine characteristics over time as they have become more numerous and more powerful in the (historically male) political domain. These shifts in the political sphere might help to further reduce voter sanctions against female politicians who adopt agentic styles, but they are also likely to reduce the degree to which women in parliament *internalise* expectations of feminine behaviour. That is, as female politicians witness more examples of women in politics adopting more agentic and less communal styles, this may weaken the self-imposed standards of femininity that are typically seen as the internal drivers of stereotype-consistent behaviour (Eagly and Wood, 2012). As female politicians become increasingly associated with forms of behaviour normally ascribed to men, the incentives for them to conform to more traditional feminine styles should be expected to decrease.

In our empirical analysis we do not attempt to disentangle which of the three mechanisms outlined here, or others, may be responsible for changes in parliamentary behaviour. Rather, we aim to test a central prediction that emerges from our discussion of dynamic gender stereotypes: that MPs will conform less to gender-stereotypic styles in

recent years than was true in the past, and that women in particular will be more likely to adopt agentic rather than communal styles over time.

Pressures to conform to institutional behavioural norms

In this section, we contrast the predictions of our argument with expectations generated by theories of feminist institutionalism. These perspectives hold that, as historically male-dominated institutions, legislatures are gendered spaces that maintain, favour and recreate traditional masculine behaviours (Hawkesworth, 2003; Krook and Mackay, 2011). While work in this literature does not always articulate clear predictions regarding the dynamics of gendered behaviour over time, implicit in these arguments is the idea that the pressure for women to conform to the prevailing (male) institutional style will be strongest when women are more marginalised in the legislature. When this is the case, as Franceschet (2011, 66) argues, "women may respond by disavowing distinctly feminine (and feminist) concerns, instead favouring the style and substantive issues of the dominant group." By contrast, as women gain higher levels of representation and more political power, the culture of parliament will change to be more "conducive to women acting in a feminized way" (Childs, 2004, 187).

The implication of this argument in our setting is that the pressure on women to conform to the dominant "masculine" institutional style will be strongest at the beginning of our study period (in the late 1990s) when women's representation in the Commons was at lower levels, but that it should weaken over time. In addition to the increasing number of women in parliament and in leadership positions, during the period we study (from 1997 to 2019) the House of Commons also introduced a series of reforms designed to strengthen the position of women within the legislature.[3] Therefore, while

[3]For example, this period includes the establishment of the Women and Equalities Committee, the introduction of the Speaker's Reference Group on Representation and Inclusion, as well as the introduction of initiatives such as proxy voting for MPs on baby leave.

the Commons remains majority male, institutionalist perspectives predict that changes in composition and working practices will mean that women will be better able to "perform their tasks as politicians the way they individually prefer" (Dahlerup, 2006, 519). Therefore, while our argument suggests that women are likely to respond to changing gender stereotypes by adopting more *agentic* styles over time, the institutionalism argument suggests that women are likely to adopt increasingly *communal* styles as they become more numerous and powerful in parliament. The empirical strategy we outline below allows us to adjudicate between these contrasting predictions.

Data and Methodology

We consider the words of politicians' speeches as the primary locus of debating style, and we use texts of political speeches delivered in parliament to infer the styles adopted by different speakers. Parliamentary speech is a useful source of information for measuring style as it provides long-running panel data at the individual level. In the UK, MPs are afforded a large degree of autonomy regarding the debates to which they contribute, and party leaders exert no control over who participates, nor over the content of speeches that MPs deliver.

We study House of Commons debates between May 1997 and March 2019. Our study period is motivated by the fact that prior to the 1997 election women accounted for less than 10% of MPs, and so analysis of earlier periods would likely be sensitive to the styles of only few specific women. We collapse our data such that all speeches made by an MP in a debate constitute a single speech-document, making our unit of analysis an individual MP in a debate. We remove all speech-documents shorter than 50 words, as well as contributions by the Speaker of the House, whose speeches are almost entirely procedural. We also exclude any debate that has fewer than five participants.[4] Our final

[4]Our model becomes computationally burdensome with very large numbers of debates. Small debates contribute little to our estimates given the random-effect structure of the model described below, and so

sample consists of 14,864 debates, 1,422 MPs (370 female, 1,052 male), and 418,147 MP-debate observations.

Measuring "style" with context-specific dictionaries

A common approach to measuring latent concepts, such as style, in text data is to assign each text a score based on a predefined dictionary that aims to capture the concept of interest. However, dictionary-based approaches are highly domain-specific, as the words used to capture a concept in one context – say, parliamentary speeches – are likely to be different to those used to express the same concept in another context. We propose an alternative approach that combines standard dictionaries with a locally-trained word-embedding model to construct domain-specific dictionaries that are better able to capture our style types *as they manifest in the context of parliamentary debate*. The key advantage of our approach is that it allows us to account for context-specific patterns of word use. That is, rather than simply using an off-the-shelf dictionary that may be poorly suited to capturing, for instance, aggression in the parliamentary setting, this approach allows us to automatically create a bespoke aggression dictionary which is firmly rooted in the way that vocabulary is used in parliamentary debate. We use this approach to measure six of our styles: aggression, affect, positive emotion, negative emotion, fact, and human narrative.

For each style, we follow three steps to construct the relevant score for each speech. First, we define a "seed" dictionary that represents our concept of interest. For four styles (affect, negative emotion, positive emotion, and fact), we use existing dictionaries and for two styles (aggression and human narrative) we create our own seed dictionaries based on a close reading of a sample of parliamentary texts.[5]

Second, we estimate a set of word-embeddings using the GloVe model described by

our results are very unlikely to be sensitive to this decision.

[5]We include a full description of the seed dictionaries in the appendix.

Pennington, Socher and Manning (2014). Word-embedding models rely on the idea that words which are used in similar contexts will have similar meanings, and the embedding model allows us to *learn* the semantic meaning of each word directly from how the word is used by MPs in debate. We train the embedding model on the full set of parliamentary speeches, and the main output of the model is the set of word-embeddings themselves. These are dense vectors that correspond to each unique word in the corpus, the dimensions of which capture the semantic "meanings" of the words. Crucially for our purposes, the distances *between* word-vectors have been shown to effectively capture important semantic similarities between different words (Mikolov et al., 2013).

By calculating the cosine similarity between every word in the corpus and the words in each of our seed dictionaries, we can therefore use this property to define the set of words that, *in the specific context of parliamentary debate*, are used in a semantically similar fashion to the seed words. We label this quantity as Sim_w^s, where w indexes words, and s indexes each style. Words closely related to the average semantic meaning of the seed words for a given dictionary will have a high similarity score (close to 1), and words that are less closely related will have a low similarity score (close to 0). The Sim_w^s scores therefore define a domain-specific dictionary for a given style type. They describe the degree to which each unique word in our corpus is used similarly to the ways in which the words in our seed dictionary are used, on average. In essence, the embeddings enable our seed dictionaries to automatically expand to incorporate words that are used in a similar manner to the words that they already include. We provide full details of our approach, and an extensive set of validation checks, in the appendix.

Third, we use the word-level scores, Sim_w^s, to score each *sentence* in the corpus on each style according to the words they contain. In particular, the score for a given sentence on a given style is:

$$Score_i^s = \frac{\sum_w^W Sim_w^s N_{wi}}{\sum_w^W N_{wi}} \quad (1)$$

where Sim_w^s is the similarity score defined above, and N_{wi} is the (weighted) number of times that word w appears in sentence i, where the weights are term-frequency inverse-document-frequency weights.[6] When words with high scores for a given style appear frequently in a given sentence, the sentence will be scored as highly relevant to the style. The score for each *document* is then the weighted average of the relevant sentence-level scores, where the weights are equal to the number of words in each sentence.

The approach we outline here is similar to that developed in Rice and Zorn (2021), who also use a word-embedding model to create context-specific dictionaries. We build on this work by addressing the question of whether word-embedding dictionary construction "[yields] valid dictionaries for widely-varying types of specialised vocabularies" (Rice and Zorn, 2021, 34). We extend the idea of word-embedding based dictionaries to a new setting – the UK House of Commons – and to six new specialised vocabularies. As our validation exercises in the appendix show, there is strong evidence that our approach significantly outperforms standard dictionary approaches across the set of styles we study.

Measuring "complexity" and "repetition"

For our final two styles – complexity and repetition – dictionary approaches (domain-specific or otherwise) are unsuitable, as these styles are not detectable from the occurrence of specific words. Instead, we adopt two different metrics to capture these concepts. For *complexity*, we use the Flesch-Kincaid Readability Score (Kincaid et al., 1975). The intuition behind this measure is that documents that have fewer words per sentences, and fewer syllables per word, are easier to understand (more "readable"). We

[6] TF-IDF weighting is used to down-weight very common words, and up-weight relatively rare words.

rescale the original formulation of the score such that higher numbers indicate higher levels of complexity. While Benoit, Munger and Spirling (2020, 501) show that domain-specific measures of textual complexity have some performance gains over the Flesch-Kincaid score, they also demonstrate that this metric correlates highly with more sophisticated measures. We opt for the simpler metric here because our own validation demonstrates that this measure performs well in comparisons with human judgements in our setting.

We consider MPs to be *repetitive* when they use the same language repeatedly during a debate. To measure repetition, we use a lossless text-compression algorithm introduced by Ziv and Lempel (1977), which underpins a variety of common computer applications. Compression algorithms work by finding repeated sequences of text and using those patterns to reduce the overall size of the input document. The efficiency of the compression of a text is directly related to the number and length of the repeated sections in that text. We apply the compression algorithm to every document in our corpus, measure the degree of compression, and treat that quantity as the measure of repetition for each MP in each debate. Simply put, the more compression that a speech receives, the more repetitive we deem it to be.[7]

Finally, to put our eight style measures on comparable scales, we normalise each measure across documents to have mean zero and standard deviation one. This means that average differences for each style between men and women can be interpreted in standard deviations of the outcome variable.

In the appendix, we provide extensive validation of our measures. In addition to a wide range of face validity checks, we provide results from a task which assesses whether

[7]An alternative measure of repetition for a given text, j, might be $\frac{\text{\# Words}_j}{\text{\# Unique Words}_j}$, which captures the intuition that texts with a smaller fraction of unique words are likely to be more repetitive. Although it correlates highly with our measure ($\rho = 0.71$), this metric is likely to underestimate the degree of repetitiveness in instances where long sequences of words are repeated, but where those sequences are themselves constituted of many unique words.

our measures mirror human codings of the same concepts. The results are very encouraging: across all styles, the correlation between our text-based scores and human judgements is always strongly-positive, indicating that we are able to reliably detect our styles of interest in parliamentary speech. In addition, our word-embedding measures clearly predict human codings more strongly than do measures based on standard dictionary approaches, which have been used in previous studies on gender and political style.[8]

Modelling political style

Our goal is to assess the degree to which style use varies by MP gender, and whether such differences change over time. To investigate these patterns, we adopt a Bayesian dynamic hierarchical model that allows us to account for a wide variety of both individual- and topic-level confounders (described below), while also flexibly estimating changing gender dynamics in style over time.

For each speech i, we have a continuous measurement of style s, which we denote as y_i^s. For speech i, by MP j, in debate d, and time period t, we model the data as a function of individual- and debate-level parameters:

$$y^s_{i(jdt)} \sim N(\alpha_{j,t} + \delta_d, \sigma_y) \tag{2}$$

where $\alpha_{j,t}$ is an MP-specific random effect which captures average differences in MP style use. The t subscript indicates that we fit one intercept for each MP in each time period that they appear in the data, thus allowing us to capture average style use at different points in time. We use parliamentary sessions as our unit of time, of which there were 20

[8] In particular, in appendix table S4 we show that the correlation between our measures and human codings is substantially higher than the correlation between human codings and a more standard dictionary measure which uses the proportion of words in each sentence that appear in each of our seed dictionaries.

between 1997 and 2019. We observe speeches from 635 MPs on average in each session, and each MP appears in 9 sessions on average. The δ_d parameters are random effects which capture average differences in style use in different debates.

Our primary interest is in describing variation in the $\alpha_{j,t}$ parameters. We model these random-effects at the second level of the model as a function of MP gender, while allowing the relationship between gender and style-use to vary over time:

$$\alpha_{j,t} \sim N(\mu_{0,t} + \mu_{1,t} Female_j, \sigma_\alpha) \tag{3}$$

Here, $\mu_{0,t}$ represents the average use of a style among male MPs in time period t, and $\mu_{1,t}$ describes the average difference in style use for women relative to men, again in time period t. The standard deviation σ_α, describes how much, on average, the MP-session intercepts vary around the mean style use for MPs of each gender.[9] Gender differences in one parliamentary session are not independent of those in previous sessions, and in order to reflect a more realistic evolution of these differences we model the $\mu_{0,t}$ and $\mu_{1,t}$ parameters as a first-order random-walk process:

$$\begin{aligned} \mu_{0,t} &\sim N(\mu_{0,t-1}, \sigma_{\mu_0}) \\ \mu_{1,t} &\sim N(\mu_{1,t-1}, \sigma_{\mu_1}) \end{aligned} \tag{4}$$

This specification assumes that the average use of a style by women and men will be similar in t and $t+1$, and that changes over time will therefore occur gradually. This encourages smooth coefficient changes over time, but still allows for large deviations from one period to the next if the information from the data is sufficiently strong.

The $\mu_{0,t}$ and $\mu_{1,t}$ parameters are our main quantities of interest. $\mu_{1,t}$ captures the difference in average style use between genders in each time period, and our review

[9] We use a common variance parameter for all time periods.

of the theoretical literature implies general expectations for the *sign* of $\mu_{1,t}$ for each style (see table 1). Consistent with our theoretical discussion of how the incentives for conforming with gender stereotypes have changed in recent years, we also expect the *magnitude* of $\mu_{1,t}$ for each style to decrease over time, and for those changes to be driven mostly by changes to the average behaviour of female MPs (which, for each year t, is captured by $\mu_{0,t} + \mu_{1,t}$). We report both quantities below.

Our model allows us to account for individual-level confounders by including a set of MP-specific covariates into the model. To do so, in some specifications we replace equation 3 with:

$$\alpha_{j,t} \sim N(\mu_{0,t} + \mu_{1,t} Female_j + \sum_{k=1}^{k} \lambda_k X_{j,t}^k, \sigma_\alpha) \tag{5}$$

where $X_{j,t}$ is a vector of individual-level covariates which can vary by session.[10]

We include several such controls. First, MPs in leadership positions may use systematically different styles than backbench MPs, and women have come to occupy a greater share of legislative leadership roles over our study period (Blumenau, 2021). We therefore control for whether the MP held a frontbench position for either the government or opposition in each session, and whether they were a committee chair.

Second, if MPs from different parties use styles at different rates, then any change we observe in gendered use of styles might be confounded by the fact that proportionally

[10]Debate intercepts are drawn from a mean-zero normal distribution, with estimated variance:

$$\delta_d \sim N(0, \sigma_\delta) \tag{6}$$

Our model is completed by normal prior distributions over the λ parameters:

$$\lambda^k \sim N(0, 2) \tag{7}$$

and half-normal prior distributions over the variance-parameters:

$$\sigma_\alpha, \sigma_{\mu_0}, \sigma_{\mu_1}, \sigma_\delta, \sigma_y \sim N(0, 2) \tag{8}$$

more Conservative Party female MPs have been elected to parliament in recent years. We therefore also include a set of party dummies.

Third, opposition MPs use significantly more negative language than government MPs (Proksch et al., 2019) and, because the Labour Party has proportionally more women than other parties, any increase in women's use of more agentic styles might be attributable to Labour's move into opposition in 2010. To address this possibility, we control for whether an MP is a member of a governing or an opposition party in each time period.

Fourth, we also add controls for MPs' occupational background and educational attainment. The professional and educational backgrounds of MPs have changed over time (Lamprinakou et al., 2017), and it is plausible that these characteristics will be associated with both speechmaking styles and gender.

Finally, MPs' local electoral environment might affect language use. For instance, MPs in more competitive seats might be more likely to use human narrative to emphasise constituents' concerns. If there have been changes in the relative competitiveness of seats won by men and women over time, this could confound the differences we observe in gendered-language use. We therefore control for the percentage point margin of victory of the MP in the previous election.

We are also able to use our model to account for confounding that relates to differential usage of styles across topics. Men and women systematically participate in debates devoted to different topics (Bäck and Debus, 2019; Catalano, 2009), and debate topic may correlate with style in ways which work to confound our inferences. For instance, if women participate more in debates on education which contain language related to human narrative, while men participate more in debates on the economy which include more factual language, then gender differences in topic usage will confound gender differences in style. However, the debate-level intercepts, δ_d, mean that it is only *within-debate* variation in style use that informs the estimates of our central quantities

of interest. In other words, $\mu_{1,t}$ will capture only the degree to which men and women use different styles when speaking about the same substantive topic.

We estimate our model separately for each style in Stan (Carpenter et al., 2017), where we use three chains of 500 iterations, after 250 iterations of burn-in.

Results

Figure 1 shows the values of $\mu_{1,t}$ – the average difference between men and women for each style type, in each parliamentary session. Positive values indicate that a style is used more by women, and negative values indicate higher use of the style by men. The green shading indicates the expected direction of the gender effects based on previous literature (see table 1).

For five of the debating styles we study, we find that – in the early years of our sample – male and female speechmaking behaviour broadly conforms to stereotypes. Female MPs are more likely to draw on examples that emphasise human narrative, and to use positive and emotional language than men. Similarly, men use more aggressive and complex language, at least before 2010. Interestingly, for three of our styles – fact, repetition, and negative emotion – we find that debating style in the Commons does not clearly conform to the expectations of the existing literature. For all three of these agentic styles, for much of the period we study, female MPs are *more* likely than male MPs to express these styles.

However, and in some sense more importantly, figure 1 also reveals that there is significant variation in the size of these gender differences over time. Women are more likely to use "communal" style types – affect, positive emotion, and human narrative – in the early period in our data, but gender differences become smaller over time. For positive emotion and affect, there is no consistent significant difference between men and women by the latest years in our data. Similarly, while men use significantly more

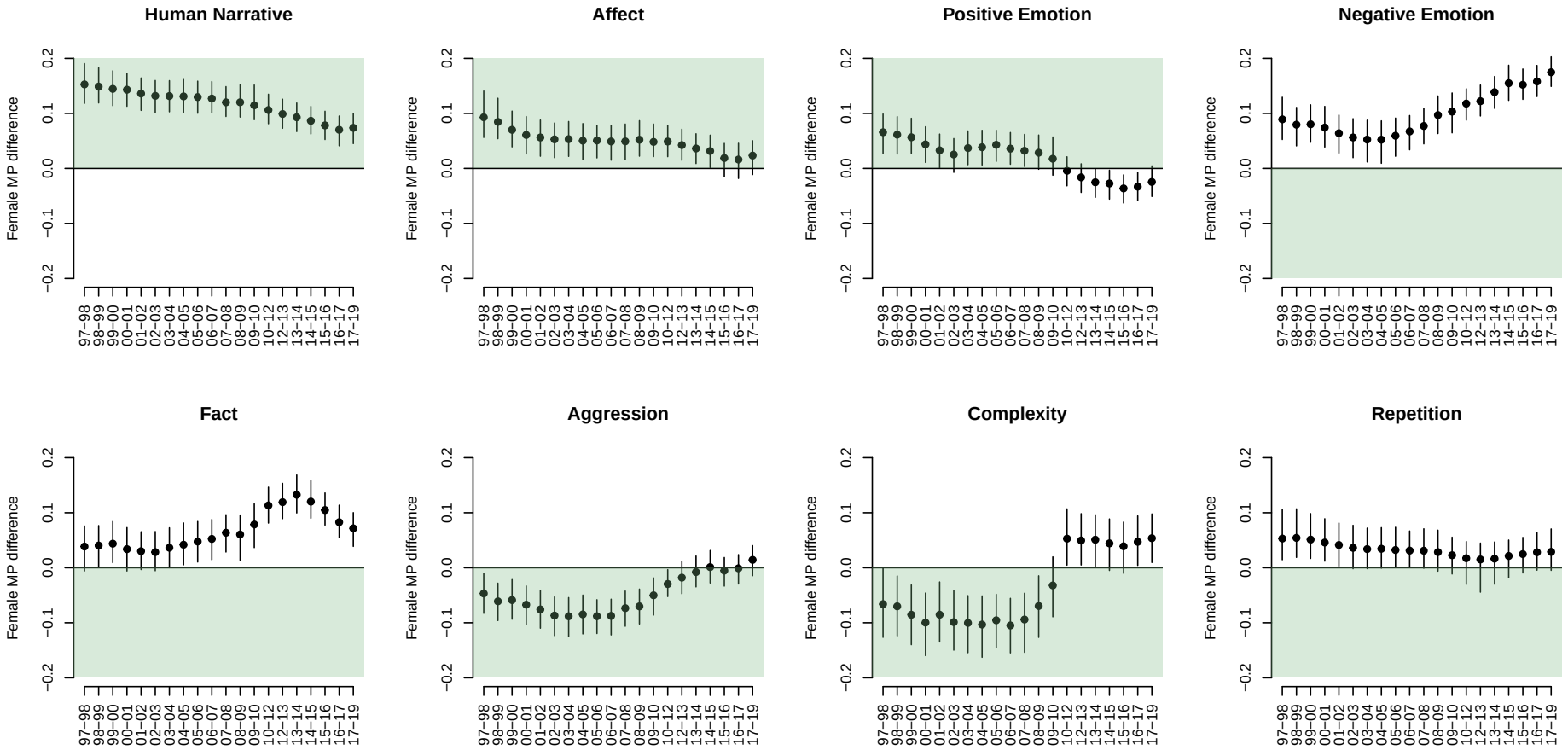

Figure 1: Gender differences in style over time

"agentic" styles – particularly aggressive and complex language – than women before 2010, this difference has also disappeared in recent years. These changes are non-trivial: for those styles where we see a convergence between men and women, the proportion of MP-level variation in style use explained by gender decreases by between 29% and 60%, depending on style, when comparing the periods before and after 2007.[11]

Further, for other styles we observe increasing gender differences over time, but the direction of these shifts also suggest that women are becoming increasingly agentic relative to men. For instance, though there are negligible gender differences in the earlier period, from 2007 onward, women use significantly more factual language. Similarly, although women use negative emotion in their speeches at higher rates throughout the time period, this gender difference has grown substantially larger over time. Between 1997 and 2007, gender explained just 0.6% and 2% of MP-level variation in factual language and negative emotion, respectively, but this increased to approximately 4% and 9% after 2007. Accordingly, even for these agentic styles which women adopt more than men throughout the study period, it remains the case that women become *more* likely to deploy this type of language in recent years than in the past. The only style for which we document relatively stable gender differences is repetition. While women appear to become somewhat less repetitious relative to men over time, the trend for this style is less pronounced.

Taken together, these findings are consistent with our argument that the pressures for women to conform to stereotypes have declined over time. In general, relative to men,

[11]To calculate these quantities we follow Gelman and Pardoe (2006) and describe the proportion of individual variation in style use explained by MP gender in each parliamentary session using an R^2-style metric:

$$R^2_{\alpha,t} = 1 - \frac{E(V_{j=1}^{J}\hat{\epsilon}_{j,t})}{E(V_{j=1}^{J}\hat{\alpha}_{j,t})}$$

where

$$\hat{\epsilon}_{j,t} = \hat{\alpha}_{j,t} - \hat{\mu}_{0,t} + \hat{\mu}_{1,t} Female_j$$

women demonstrate less communal (human narrative, affect, and positive emotion) and more agentic (negative emotion, aggression, fact, and complexity) styles in recent years than they did in the past.

Are these patterns the result of changes in the behaviour of male or female MPs? Our argument implied that these changes were likely to be rooted in the behaviour of women as they respond to the changing content and power of gender stereotypes. Figure 2, which depicts changes in style use separately for women and men, shows that across almost all the styles we study, the largest year-to-year shifts in speechmaking behaviour do indeed occur among women. The figure shows that women have used each of our communal style types – human narrative, affect, and positive emotion – to a decreasing extent over time. Similarly, for the "agentic" styles of negative emotion, fact, and aggression, while men's behaviour has remained relatively stable, women's use of these styles has increased over time. While both men and women have adopted more complex language over time, the increase has been somewhat larger for women than for men.

Threats to inference

We have argued that politicians' conformity to traditional stereotypes has diminished over time, but there are alternative explanations that could account for the behavioural patterns we document.

First, as we outlined above, the literature on feminist institutionalism suggests that women will face pressures to conform to institutionally-dominant, masculine behaviours that are favoured and recreated by the culture of the House. While institutional pressures of this sort are surely a feature of life in the contemporary House of Commons, for this perspective to explain our results, these pressures would need to have *strengthened* in recent years. However, during this period, women's presence increased in the

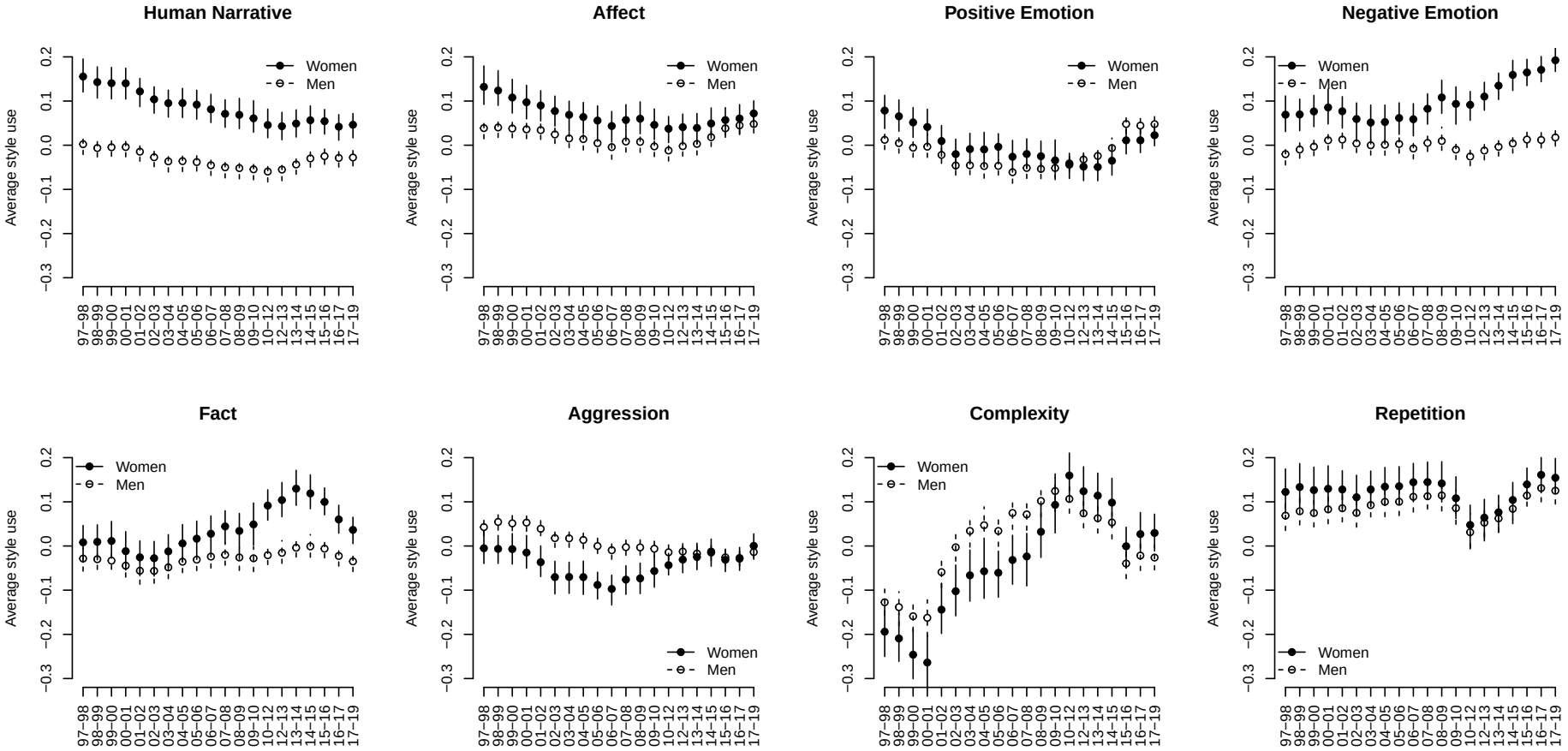

Figure 2: Average style use for men and women over time

Commons, female MPs came to hold more senior positions, and institutional reforms designed to strengthen women's institutional position were introduced. Consequently, if women face stronger incentives to conform to masculine styles when an institution is more male dominated, then we should observe women becoming *less* agentic over time. Our results show the opposite pattern, suggesting that institutionalist accounts are unlikely to explain the over-time dynamics in speechmaking that we document.

Second, figures S4 and S5 in the appendix report results from the model described in equation 5 in which we control for a host of MP-level covariates. If the changes we observe over time are driven by factors such as party, opposition status, and so on, we would expect to see large differences between these two sets of results. Although there is some attenuation of the over-time changes for aggression and complexity when controlling for covariates, we nonetheless still observe stereotype-consistent differences at the beginning of the sample period and clear evidence of women using more agentic styles over time.

Third, the aggregate patterns we document could reflect changes to the parliamentary agenda, rather than changes in gendered behaviour. If there are certain topics on which women are more likely to demonstrate agentic styles, and these topics feature more prominently on the parliamentary agenda in the later period, then our results might be explained by changing topical prevalence over time. In the appendix, we use statistical topic-models to measure the differences between male and female style use across a wide variety of topics, and then evaluate whether topics that are marked by large stylistic differences become more or less prevalent over time relative to topics marked by smaller differences. We find scant evidence of such topical confounding.

Finally, in the appendix, we also assess whether women with more agentic styles participate more, and women with more communal styles participate less, in parliamentary debate over time. We show that differential participation does not explain our results:

MPs' styles largely fail to predict debate participation throughout our study period. In addition, we also investigate whether the changes we document are due to changes in styles of men and women throughout their careers ("within-MP" effects), or because the men and women entering parliament over time are systematically different from those leaving ("replacement" effects). While there is some evidence that replacement is more important for explaining the changes in agentic styles and within-MP change is somewhat more important for explaining change in communal styles, overall we find that neither replacement nor within-MP change can alone explain the patterns that we document above.

Conclusion

Our central substantive contribution is to document the fact that gender stereotypes are worse descriptors for actual political behaviour in the UK now than was true in the past. In particular, in recent years, women in the House of Commons demonstrate less communal and more agentic styles, and the gender gap on most dimensions of style that we examine has decreased. We see these results as an important corrective to the scholarly literature on gender differences in legislative behaviour, which typically emphasises that male and female politicians argue in ways that are broadly consistent with stereotypes. Though this may be true in some settings, gender stereotypes of communication styles have become significantly less predictive of the reality of contemporary British political debate.

These findings do not, however, imply that gender-stereotypes play no role in UK politics. We show that recent parliamentary behaviour is poorly described by traditional stereotypes, but we do not provide empirical evidence regarding the mechanisms that led to these changes. For instance, previous work shows that the public are less likely to endorse traditional stereotypes now than in the past, but we lack data to as-

sess whether there has been a concomitant decline in the sanctions that voters apply for gender-role-inconsistent behaviour. Anecdotally, there continue to be examples of British female politicians being criticised for stereotype-incongruent behaviour[12] and stereotypes may well continue to condition voter responses. Though we think this type of sanctioning is likely to have declined because of general changes in voter attitudes regarding stereotypes, future work should focus on collecting over-time survey data on voters' attitudes towards non-stereotypical behaviour by politicians.

One optimistic view of our findings, however, is that there may be a virtuous circle in which female politicians diverge from stereotypical behaviours, and that this in turn changes perceptions of appropriate feminine behaviour, which thereby reduces the pressures women are under to conform to such stereotypes. Changes in the typical behaviours of male and female politicians are likely to translate only slowly into revised public expectations of the standards against which men and women MPs are judged. However, to the degree that the behavioural shifts that we document are noticed and internalised by the public, they might also help to reduce the social penalties applied to female politicians who display more agentic styles.

Our findings also have implications for wider debates about political culture in the UK. There is a strand of popular commentary that implies some of the more unattractive features of Westminster's adversarial culture would be ameliorated if only more women were to be elected to public office. Our results suggest, however, that simply increasing women's numbers in parliament is unlikely to make UK politics gentler or more deliberative. The pursuit of a "better" politics requires more than vaguely hoping, on the basis of a dogged adherence to outdated gender stereotypes, that the election of women will fundamentally change the ways that our representatives communicate.

Methodologically, our paper addresses a well-known problem for quantitative text

[12] See "The Making of the Maybot", *Spectator*, 2nd November 2017.

analysis based on dictionaries: the words that demonstrate a given concept in one context may be poorly suited to detecting the use of that concept in another context. We used a word-embedding model to capture how different political styles manifest in the specific setting of parliamentary debate. Results from our validation (in the appendix) show that this approach significantly outperforms existing methods, a finding we believe justifies adoption elsewhere. Our strategy is likely to be useful whenever researchers are interested in measuring a latent concept from a large corpus of texts, but where the domain of interest differs from the domain in which existing dictionaries were developed. This describes a large fraction of applications of dictionary methods, and so our approach has the potential to be applied widely elsewhere.

Finally, we focus only on style as expressed in legislative debates. Style may, of course, manifest in other forms of legislative behaviour, or, indeed, other forms of political speech. While we show that gender gaps in legislative speech have declined, it remains possible that gender stereotypes may still be powerful in other arenas, such as in campaign communication where politicians may be particularly sensitive to voter penalties. We hope that our findings motivate other scholars to explore how gender-based differences in political communication have evolved over time in other contexts.

References

Bäck, Hanna and Marc Debus. 2019. "When Do Women Speak? A Comparative Analysis of the Role of Gender in Legislative Debates." *Political Studies* 67(3):576–596.

Bauer, Nichole M. 2015*a*. "Emotional, Sensitive, and Unfit for Office? Gender Stereotype Activation and Support Female Candidates." *Political Psychology* 36(6):691–708.

Bauer, Nichole M. 2015*b*. "Who stereotypes female candidates? Identifying individual differences in feminine stereotype reliance." *Politics, Groups, and Identities* 3(1):94–110.

Bauer, Nichole M. 2017. "The Effects of Counterstereotypic Gender Strategies on Candidate Evaluations." *Political Psychology* 38(2):279–295.

Bauer, Nichole M., Nathan P. Kalmoe and Erica B. Russell. 2021. "Candidate Aggression and Gendered Voter Evaluations." *Political Psychology* pp. 1–21.

Benoit, Kenneth, Kevin Munger and Arthur Spirling. 2020. "Measuring and Explaining Political Sophistication Through Textual Complexity." *American Journal of Political Science* 63(2):491–508.

Blankenship, Jane and Deborah C. Robson. 1995. "A 'feminine style'; in women's political discourse: An exploratory essay." *Communication Quarterly* 43(3):353–366.

Blumenau, Jack. 2021. "The Effects of Female Leadership on Women's Voice in Political Debate." *British Journal of Political Science* 51(2):750–771.

Boussalis, Constantine, Travis G. Coan, Mirya R. Holman and Stefan Müller. 2021. "Gender, Candidate Emotional Expression, and Voter Reactions During Televised Debates." *American Political Science Review* pp. 1–39.

Brescoll, Victoria L. and Eric Luis Uhlmann. 2008. "Can an Angry Woman Get Ahead?" *Psychological Science* 19(3):268–275.

Brooks, Deborah Jordan. 2011. "Testing the Double Standard for Candidate Emotionality: Voter Reactions to the Tears and Anger of Male and Female Politicians." *The Journal of Politics* 73(2):597–615.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 2017. "Stan: A probabilistic programming language." *Journal of Statistical Software* 76(1):1–32.

Cassese, Erin C. and Mirya R. Holman. 2018. "Party and Gender Stereotypes in Campaign Attacks." *Political Behavior* 40:785–807.

Catalano, Ana. 2009. "Women Acting for Women? An Analysis of Gender and Debate Participation in the British House of Commons 2005-2007." *Politics & Gender* 5(1):45–68.

Childs, Sarah. 2004. *New Labour's Women MPs: Women Representing Women*. London, UK: Routledge.

Coates, Jennifer. 2015. *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. London: Routledge.

Dahlerup, Drude. 1988. "From a Small to a Large Minority: Women in Scandinavian Politics." *Scandinavian Political Studies* 11(4):275–298.

Dahlerup, Drude. 2006. "The Story of the Theory of Critical Mass." *Politics & Gender* 2(4):511–522.

Dasgupta, Nilanjana and Shaki Asgari. 2004. "Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping." *Journal of Experimental Social Psychology* 40(5):642–658.

de Geus, Roosmarijn A., John R. McAndrews, Peter John Loewen and Aaron Martin. 2021. "Do Voters Judge the Performance of Female and Male Politicians Differently? Experimental Evidence from the United States and Australia." *Political Research Quarterly* 74(2):302–316.

Diekman, Amanda B. and Alice H. Eagly. 2000. "Stereotypes as dynamic constructs: Women and men of the past, present, and future." *Personality and Social Psychology Bulletin* 26(10):1171–1188.

Diekman, Amanda B, Alice H. Eagly, Antonio Mladinic and Maria Cristina Ferreira. 2005. "Dynamic stereotypes about women and men in Latin America and the United States." *Journal of Cross-Cultural Psychology* 36(2):209–226.

Diekman, Amanda B. and Wind Goodfriend. 2006. "Rolling with the changes: A role congruity perspective on gender norms." *Psychology of Women Quarterly* 30(4):369–383.

Dietrich, Bryce J., Matthew Hayes and Diana Z. O'Brien. 2019. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech." *American Political Science Review* 113(4):941–962.

Ditonto, Tessa. 2017. "A High Bar or a Double Standard? Gender, Competence, and Information in Political Campaigns." *Political Behavior* 39:301–325.

Eagly, Alice H, Christa Nater, David I Miller, Michèle Kaufmann and Sabine Sczesny. 2020. "Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018." *American Psychologist* 75(3):301–315.

Eagly, Alice H. and Wendy Wood. 2012. "Biosocial Construction of Sex Differences and Similarities in Behavior." *Advances in Experimental Social Psychology* 46:55–123.

Eagly, Alice and Steven Karau. 2002. "Role Congruity Theory of Prejudice Toward Female Leaders." *Psychological Review* 109(3):573–598.

Franceschet, Susan. 2011. Gendered Institutions and Women's Substantive Representation: Female Legislators in Argentina and Chile. In *Gender, Politics and Institutions: Towards a Feminist Institutionalism*, ed. Mona Lena Krook and Fiona Mackay. Palgrave Macmillan pp. 58–78.

Gelman, Andrew and Iain Pardoe. 2006. "Bayesian measures of explained variance and pooling in multilevel (hierarchical) models." *Technometrics* 48(2):241–251.

Grey, Sandra. 2002. "Does Size Matter? Critical Mass and New Zealand's Women MPs." *Parliamentary Affairs* 55:19–29.

Hargrave, Lotte and Tone Langengen. 2020. "The Gendered Debate: Do Men and Women Communicate Differently in the House of Commons?" *Politics & Gender* pp. 1–27.

Hawkesworth, Mary. 2003. "Congressional Enactments of Race-Gender: Toward a Theory of Raced-Gendered Institutions." *American Political Science Review* 97(4):529–550.

Holman, Mirya R., Jennifer L. Merolla and Elizabeth J. Zechmeister. 2016. "Terrorist Threat, Male Stereotypes, and Candidate Evaluations." *Political Research Quarterly* 69(1):134–147.

Huddy, Leonie and Nayda Terkildsen. 1993. "Gender Stereotypes and the Perception of Male and Female Candidates." *American Journal of Political Science* 37(1):119–147.

Inglehart, Ronald and Pippa Norris. 2003. *Rising tide: Gender equality and cultural change around the world*. Cambridge Cambridge University Press.

IPU. 2020. "Women in National Parliaments.".
URL: *http://archive.ipu.org/wmn-e/world.htm*

Jamieson, Kathleen. 1988. *Beyond the Double Bind: Women and Leadership*. Oxford: Oxford University Press.

Jones, Jennifer J. 2016. "Talk "Like a Man": The Linguistic Styles of Hillary Clinton, 1992-2013." *Perspectives on Politics* 14(3):625–642.

Kathlene, Lyn. 1994. "Power and influence in state legislative policymaking: The interaction of gender and position in committee hearing debates." *American Political Science Review* 88(3):560–576.

Kincaid, J. Peter, Robert P. Fishburne Jr, Richard L. Rogers and Brad S. Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. Technical report Institute for Simulation and Training.

Koenig, Anne, Alice Eagly, Abigail Mitchell and Tiina Ristikari. 2011. "Are leader stereotypes masculine? A meta-analysis of three research paradigms." *Psychological Bulletin* 137(4):616–642.

Krook, Mona Lena and Fiona Mackay. 2011. *Gender, Politics and Institutions: Towards a Feminist Institutionalism*. Basingstoke, UK: Palgrave Macmillan.

Lamprinakou, Chrysa, Marco Morucci, Rosie Campbell and Jennifer van Heerde-Hudson. 2017. "All change in the house? The profile of candidates and MPs in the 2015 British general election." *Parliamentary Affairs* 70(2):207–232.

Leaper, Campbell and Melanie M Ayres. 2007. "A Meta-Analytic Review of Gender Variations in Adults' Language Use: Talkativeness, Affiliative Speech, and Assertive Speech." *Personality and Social Psychology Review* 11(4):328–363.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *Working Paper* pp. 1–12.

Mo, Cecilia Hyunjung. 2015. "The Consequences of Explicit and Implicit Gender Attitudes and Candidate Quality in the Calculations of Voters." *Political Behavior* 37:357–395.

Okimoto, Tyler G. and Victoria L. Brescoll. 2010. "The Price of Power: Power Seeking and Backlash Against Female Politicians." *Personality and Social Psychology Bulletin* 36(7):923–936.

Pennington, Jeffrey, Richard Socher and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1532–1543.

Proksch, Sven Oliver, Will Lowe, Jens Wäckerle and Stuart Soroka. 2019. "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches." *Legislative Studies Quarterly* 44(1):97–131.

Rice, Douglas R. and Christopher Zorn. 2021. "Corpus-based dictionaries for sentiment analysis of specialized vocabularies." *Political Science Research and Methods* 9(1):20–35.

Saha, Sparsha and Ana Catalano Weeks. 2020. "Ambitious Women: Gender and Voter Perceptions of Candidate Ambition." *Political Behavior* pp. 1–27.

Sayer, Liana C., Suzanne M. Bianchi and John P. Robinson. 2004. "Are parents investing less in children? Trends in mothers' and fathers' time with children." *American Journal of Sociology* 110(1):1–43.

Schneider, Monica C. and Angela L. Bos. 2019. "The Application of Social Role Theory to the Study of Gender in Politics." *Political Psychology* 40(1):173–213.

Seguino, Stephanie. 2007. "PlusÇa Change? Evidence on global trends in gender norms and stereotypes." *Feminist Economics* 13(2):1–28.

Sendén, Marie Gustafsson, Amanda Klysing, Anna Lindqvist and Emma Aurora Renström. 2019. "The (not so) changing man: Dynamic gender stereotypes in Sweden." *Frontiers in Psychology* 10(JAN):1–17.

Shorrocks, Rosalind. 2018. "A Feminist Generation? Cohort Change in Gender-Role Attitudes and the Second-Wave Feminist Movement." *International Journal of Public Opinion Research* 30(1):125–145.

Taylor, Eleanor Attar and Jacqueline Scott. 2018. "Gender: New consensus or continuing battleground?" *British Social Attitudes 35* pp. 56–85.

Twenge, Jean M. 1997. "Attitudes toward women, 1970–1995: A meta-analysis." *Psychology of Women Quarterly* 21(1):35–51.

Twenge, Jean M. 2001. "Changes in women's assertiveness in response to status and roles: A cross-temporal meta-analysis, 1931-1993." *Journal of Personality and Social Psychology* 81(1):133–145.

Yu, Bei. 2013. "Language and gender in congressional speech." *Literary and Linguistic Computing* 29(1):118–132.

Ziv, Jacob and Abraham Lempel. 1977. "A Universal Algorithm for Sequential Data Compression." *IEEE Transactions on Information Theory* 23(3):337–343.

Appendix – No Longer Conforming to Stereotypes? Gender, Political Style, and Parliamentary Debate in the UK *

Contents

***This version**: June 18, 2021.

Word-embedding-based dictionaries

Our word-embedding-based measurement strategy consists of several steps, which we describe in more detail in this section.

First, for each style we define a "seed" dictionary that represents our concept of interest. We use the following sources to construct our seed dictionaries:

1. **Affect** – Linguistic Inquiry and Word Count 2015 (Affect) (Pennebaker et al., 2015)
2. **Fact** – Linguistic Inquiry and Word Count 2015 (Number and Quantitative) (Pennebaker et al., 2015) and all occurrences of any numeric figures
3. **Positive Emotion** – Regressive Imagery Dictionary (Emotions: Positive Affect) (Martindale, 1990)
4. **Negative Emotion** – Regressive Imagery Dictionary (Emotions: Anxiety and Sadness) (Martindale, 1990)
5. **Aggression** – A bespoke dictionary of words (see figure S1 below)
6. **Human Narrative** – A bespoke dictionary of words (see figure S2 below) and the 200 most common names of children born between 1970 and 2019

The final two seed dictionaries – which relate to aggression and human narrative – are our original constructions. These dictionaries were constructed by reading and watching debates from the House of Commons that are known to feature either aggression (for instance, Prime Minister's Questions) or examples of human narrative (for instance, debates on mental health or social policy issues), and selecting words and phrases that we thought were likely to capture these concepts in a broader set of parliamentary debates. We report the full lists of words that feature in these new seed dictionaries in figures S1 and S2

Second, a key component of our approach to measuring style are a set of word-embeddings, which we estimate from the full corpus of parliamentary speeches. Word-

Figure S1: “Aggression” seed dictionary

irritated ; stupid ; stubborn ; accusation ; accuse ; accusations; accusing ; anger ; angered ; annoyance ; annoyed ; attack ; insult ; insulting ; insulted ; betray; betrayed ; blame ; blamed ; blaming ; bitter; bitterly ; bitterness ; complain; complaining; confront ; confrontation; fibber; fabricator ; phoney ; fibber ; sham ; deceived ; deceive ; disgrace; villain; good-for-nothing; hypocrite ; deception; steal ; needlessly; needless; criticise ; criticised ; criticising ; blackened ; fiddled; fiddle; problematic ; lawbreakers ; offenders; offend; unacceptable ; leech; phoney ; appalling ; incapable ; farcical ; absurd ; ludicrous; nonsense ; laughable ; nonsensical ; ridiculous; outraged ; hysterical ; adversarial ; aggressive ; shady ; stereotyping; unhelpful ; unnatural ; assaulted ; assault ; assaulting ; half-truths ; petty; humiliate ; humiliating ; confrontational; hate ; hatred ; furious ; hostile ; hostility ; nasty; obnoxious ; sleaze; sleazy ; inadequacy; faithless; neglectful ; neglect; neglected; wrong ; failure ; failures ; failed ; fail ; scapegoat ; cruel; cruelty ; demonise ; demonised ; tactic ; trick; trickery ; deceit ; dishonest ; deception; devious; deviousness; shenanigans ; fraudulence ; fraudulent ; fraud; swindling; archaic ; sly; slyness; silly; silliness ; scandal; scandalous ; slander ; slanderous ; libellous ; disreputable ; dishonourable ; shameful; atrocious ; gimmick ; immoral; ridicule; antagonistic ; antagonise ; ill-mannered; spiteful ; spite ; vindictive ; prejudice ; prejudices ; disregard ; arrogant ; arrogance ; embarrassment ; embarrass; embarrassing ; distasteful ; provoke; provoked ; petulant ; ignorance ; stupidity ; idiot ; idiotic ; annoying; dodgy ; untrue ; penny-pinching ; attacking ; ironic ; irony ; outrageous; hackery; crass; backchat; rude ; ill-judged ; ragbag; mess; hash ; fiasco; shambles ; shambolic ; farce; botch; botched ; blunder ; mischievous; mischief ; undermine ; straightjacket ; groan; abuse; chaos; chaotic ; dull; predictable ; negligent; grotesque; scapegoats; hypocrisy; bogus; counterproductive; betrayal; patronise ; patronising; reprehensible; fool; foolish; abysmal ; disgraceful; woeful; inferior ; sneaky ; scaremongering; scaremonger; coward; cowardly; ignorant; intolerant; unacceptable ; condemn; short-sighted; ashamed; falsehood; blackmail; clownery; debased; debase; hypocrisy; mislead; misleading; smokescreen; subterfuge; horrendous; despicable; deplorable

Figure S2: “Human narrative” seed dictionary

example; constituent; person; someone; instance; surgery; case; told; illustrate; anecdote; experience; people; individual; cases; man; woman; mother; father; son; daughter; uncle; aunt; cousin; wife; husband; parent; child; say; said; support; discuss; speak; community; local; area; family; issues; remember; recall; married; resolve; authorities ; help; imagine; envisage; lives; sometimes; concerned; heard; circumstance; anyone; nobody; citizens; relationship; girl; boy; believe; listen; problem; inspire; many; comment; authority; conversation; worked; tell; thought; life; home; referred; situation; happened; everyone; concern; recognise; advice; advise; everyday; personal; letter; involve; nephew; niece; learn; local area; my constituents; previous job; tell me; told me; first hand; speaking as; own experience; for example; I recognise; I remember; help people; many years; see me; spoke with; their; them; talk; constituency ; constituents ; mum; dad; rhetoric; mr ; mrs ; know ; wrote ; write; ask ; call; dr; doctor; society; ordinary ; together ; dear; honest; visit; everybody; feel; view; public ; employer ; reflect; born; expect; anybody; responsibility ; youngster; heartbreaking; young; hopeless ; desperate; picture; chat; electorate; provide for; foster; colleague; represent ; neighbourhood; locality ; sympathy ; condolence; grief; bereavement ; trust; serve; communicate; testimony; motherhood; fatherhood; sensitive; remark; couple; brave; lifelong; proud; pride; facilities; quote; real; meet; met; childhood; reminisce ; nostalgia; recollect; hometown; lifetime; email; neighbour; partner; children; teenager; youth; contact; tale; scenario; bred; hard-working; year-old; friend; parent; parents; came; knew; recently; lady; gentleman; families

embedding models, which are of increasing use in political science (Spirling and Rodriguez, 2019), seek to describe any word in a corpus as a dense, real-valued vector of numbers. The construction of the word-embedding vectors, regardless of the specific algorithm used to estimate them, relies centrally on the distributional hypothesis: the idea that words which are used in similar contexts will have similar meanings. Here, a context refers to a window of words around a target word, and the embedding model allows us to *learn* the semantic meaning of each word directly from the use of the word in the corpus.

The main output of embedding models are the word-embeddings themselves. These are vectors that correspond to each unique word in the corpus. The dimensions of the embedding vectors capture different semantic "meanings" that can be used to provide structure to vocabulary. Crucially for our purposes, given this representation, the distances *between* word-vectors have been shown to effectively capture important semantic similarities between different words (Mikolov et al., 2013). We use this property to define the set of words that, *in the context of UK parliamentary debate*, are used in a semantically similar fashion to the seed words.

We follow the estimation procedure outlined in Pennington, Socher and Manning (2014) and estimate a word embedding, W, of length $J = 150$ for each unique word in our corpus. We use a small "context" window size of 3 words either side of the target word to estimate our embeddings. This is consistent with our aim of capturing semantic (rather than topical) relations between words (Spirling and Rodriguez, 2019, 7). We exclude all words that occur very rarely (fewer than 90 times overall), and all words that occur very frequently (in more than 90% of documents). We remove all stop-words, punctuation, and a bespoke list of parliamentary address terms such as "Honourable Friend" or "Home Secretary". We collect the embeddings in a matrix, θ, which we use to calculate the mean word-embedding vector for each of our seed dictionaries. The average word-embedding

of the seed words represents the "location" of the dictionary in the vector-space defined by the embedding model, and allows us to calculate the relative semantic similarity of different words to the dictionary.

Third, we calculate the similarity between *every* word in the corpus and the mean dictionary word-vector using the cosine-similarity metric. Words closely related to the average semantic meaning of the seed words will have a high similarity score, and words that are less closely related will have a low similarity score. We then follow Zamani and Croft (2016) and apply the sigmoid function to the similarity scores, which transforms all similarity scores to the [0,1] interval and shrinks the scores of all but the most similar words to very close to zero. Where x_w^s is the cosine similarity between the word-embedding for word w and the mean word-embedding of the seed dictionary for style s, the sigmoid transformation is given by:

$$Sim_w^s = \frac{1}{1 + e^{-a(x_w^s - c)}} \tag{S1}$$

Here, a and c are free parameters which we set to be equal to 40 and .35, respectively, based on the results in Zamani and Croft (2016, 3). Sim_w^s gives our final score for each word for each style. Words closely related to the average semantic meaning of the seed words for a given dictionary will have a high Sim_w^s, and words that are less closely related will have a low Sim_w^s.

Finally, we use the word-level scores, Sim_w^s, to score each *sentence* in the corpus. As described in the main body of the paper, the score for a given sentence on a given dimension is:

$$Score_i^s = \frac{\sum_w^W Sim_w^s N_{wi}}{\sum_w^W N_{wi}} \tag{S2}$$

where Sim_w^s is the similarity score defined above, and N_{wi} is the (weighted) number of

times that word w appears in sentence i, where the weights are term-frequency inverse-document-frequency weights.[1] $Score_i^s$ represents the fraction of words in sentence i that are relevant to dictionary s. When words with high scores for a given style appear frequently in a given sentence, the sentence will be scored as highly relevant to the style. The score for each *document* is then the weighted average of the relevant sentence level scores, where the weights are equal to the number of words in each sentence.

[1] TF-IDF weighting is used to down-weight very common words, and up-weight relatively rare words.

Validation tests

As with all quantitative text analysis approaches, careful validation of our measures is essential (Grimmer and Stewart, 2013), and we provide two face validation checks in this section, as well results from a human validation task.

Face validity checks

In table S1, we examine the words that are associated with large Sim_w^s values for each of our styles. In particular, the table shows the top 30 words associated with each concept according to our word-embedding measure (*Top*), the words that are high-scoring based on the word-embedding measure, but which do not feature in the seed dictionaries (*Added*), and the words that are low-scoring on the word-embedding measure but which did feature in the seed dictionaries (*Removed*). The *Added* words are particularly important, as they represent words that are used in a similar context to the words in our seed dictionary in the parliamentary setting, but which would be missed by traditional dictionary based approaches.

The tables reveal that high-weight words (*Top*) generally correspond very closely to the style dimensions to which they relate. For instance, the top-loading words in the "Positive Emotion" dimension include "joy", "delight", "eager", and "excitement". Similarly, in the "Aggression" dimension, top words include "disgraceful", "shameful", "outrageous", and "scaremongering". It is also encouraging that the top words in the "Fact" dimension are mostly numeric quantifiers, and the top "Human Narrative" words include "constituent", "told", "wrote", "said", and several words that indicate specific individuals ("son", "father", "wife").

In addition, many words that are not included in the original seed dictionaries are nevertheless given high weights via the word-embedding approach (*Added*). For exam-

ple, the words “shocking”, “incompetence”, “pathetic”, and “deplore” do not appear in the “Aggression” seed dictionary, but nevertheless receive high weights for that style. That these words are consistent with intuitive notions of these broad stylistic categories, although not in the original dictionaries, highlights the fact that the word-embedding approach is successfully finding words that are semantically closely related to our key concepts of interest.

Similarly, the table also shows that some words included in the original seed dictionaries which are not semantically similar to the relevant concepts in the context of parliamentary debate are given low weights by the word-embedding approach (*Removed*). For example, that “terrorism” is removed from the “Negative Emotion” dictionary is encouraging, as within a parliamentary context the use of the word “terrorism” is likely to be from a reference to matters of policy rather than to an expression of emotion.

Overall, the words in table S1 suggest that our word-embedding model is a) accurately associating sensible words with our stylistic concepts; and b) capturing language use that is representative of a given style, even when those words are not included in our seed dictionaries, and so would be missed by traditional dictionary approaches.

Affect			Positive Emotion		
Top	*Added*	*Removed*	*Top*	*Added*	*Removed*
feel	feel	award-winning	joy	eager	gladstone
really	really	admiral	delight	anticipation	reliefs
sometimes	sometimes	securities	eager	pity	satisfied
afraid	undoubtedly	super	enjoyable	liked	relieve
fear	always	approvals	happy	hear	relieving
undoubtedly	frankly	destroyers	excitement	appreciated	satisfactorily
always	think	festival	enjoying	amazed	satisfy
frankly	nevertheless	dwellings	cheer	wonderful	gay
think	often	engagements	celebration	sadness	relief
nevertheless	genuinely	championships	delighted	love	grind
often	believe	championship	celebrate	doubtless	satisfies
genuinely	seem	approving	relieved	birthday	satisfactory
believe	felt	shakespeare	amused	horrified	entertainment
certainly	however	challenger	anticipation	praise	grinding
seem	feeling	treasurer	fun	always	amusement
felt	indeed	pesticides	pity	fascinating	laughed
however	perhaps	risk-based	enjoyed	informative	laughs
feeling	obviously	approved	entertaining	look_forward	satisfaction
indeed	something	harmonise	liked	churlish	gladly
perhaps	find	flexibilities	hear	pleased	cheers
obviously	say	energy-intensive	enjoy	admire	satisfying
something	probably	laughs	appreciated	christmas	laughing
worry	nothing	relaxing	amazed	lovely	entertain
find	deeply	exhaustive	excited	afternoon	rejoice
say	people	approve	wonderful	fascinated	laughable
probably	suspect	festivals	sadness	spirit	enthusiastically
nothing	thing	resignations	celebrating	compliment	cheered
deeply	somehow	approves	love	astonished	enjoyment
people	quite	glamorgan	doubtless	sincerely	celebrates
suspect	much	champagne	birthday	coincidence	jokes

Negative Emotion			Aggression		
Top	*Added*	*Removed*	*Top*	*Added*	*Removed*
upset	upset	painstaking	disgraceful	utterly	inferior
suffering	terrible	painting	shameful	cynical	offenders
terrible	hurt	alarms	outrageous	frankly	assaulted
distressing	deeply	paint	scaremongering	embarrassing	annoyance
hurt	unfortunate	paints	utterly	incompetence	fiddle
distress	angry	terrific	cynical	misguided	fiddled
frightening	feeling	disappointingly	frankly	irresponsible	steal
unhappy	felt	avoidance	scandalous	pathetic	assault
worry	caused	terrorists	dishonest	dreadful	offend
deeply	horrendous	cowardly	embarrassing	bizarre	fail
dreadful	appalling	grievance	absurd	complacency	furious
unfortunate	frustrating	miserably	ridiculous	illogical	deceived
worried	shocked	hopelessly	ludicrous	incompetent	dodgy
suffer	compounded	alarmingly	deplorable	shocking	predictable
anxiety	frustration	lone	incompetence	reckless	fool
frightened	anger	terrorism	misguided	disingenuous	problematic
despair	sometimes	grievances	irresponsible	deliberate	bitterness
fear	experiencing	alarmist	pathetic	complacent	fiasco
angry	horrible	painted	appalling	unfortunate	neglected
suffered	frustrated	timid	dreadful	downright	betray
sad	feel	terrorist	nonsense	wicked	deceive
feeling	appalled	shy	bizarre	deplore	cruelty
felt	shocking	discouraged	complacency	unjust	confrontational
tragic	disturbed	discouraging	ashamed	unacceptable	archaic
caused	understandably	avoids	illogical	horrible	blackmail
horror	unpleasant	lamentable	arrogant	plainly	embarrass
horrendous	terribly	pitiful	incompetent	muddle	mischief
appalling	embarrassing	discourage	shocking	manifestly	smokescreen
tragedy	awful	sufferers	accusation	callous	adversarial
frustrating	imagine	painfully	arrogance	excuse	annoyed

Fact			Human Narrative		
Top	*Added*	*Removed*	*Top*	*Added*	*Removed*
half	nearly	sixthly	constituent	like	poppy
five	year	sevenoaks	told	called	bred
four	whereas	doubly	know	whose	amber
nearly	years	infinitely	wrote	went	florence
three	£	ooost-century	like	think	georgia
ooo	months	double-dip	called	indeed	anecdote
six	just	infinite	said	says	hopeless
year	days	oooth-century	constituents	also	recollect
whereas	weeks	scarce	whose	just	alice
years	moreover	bunch	father	others	aunt
seven	past	seven-day	mr	asked	eve
two	compared	groupings	son	saying	albert
quarter	almost	fifthly	went	week	skye
eight	yet	samples	tell	wanted	chat
£	now	grouped	met	see	spencer
million	next	equalities	remember	perhaps	kate
months	spend	ooo-page	think	former	mohammed
billion	thirds	equalise	indeed	described	rhetoric
just	ago	ooog	says	obviously	ashton
average	figure	group's	dr	one	tale
least	addition	ooond	david	ooo-year-old	roman
days	roughly	sixth-form	wife	mine	inspire
weeks	week	oood	say	knows	nicola
moreover	furthermore	equalisation	also	yesterday	locality
past	number	ooo-to-ooo	just	unfortunately	youngster
compared	within	six-week	others	looked	everyday
third	times	triple	woman	friends	sensitive
almost	equivalent	grouping	family	aware	jamie
yet	april	four-year-old	asked	although	carter
one	probably	oood	man	now	scenario

Table S1: Word-level validation

Tables S2 and S3 assess the face validity of our approach by showing the 10 highest scoring *sentences* for each style, according to the $Score_i^s$ measure described in equation S2. For all styles, the sentences clearly reflect the conceptual definitions we outline in the main paper. For instance, the "fact" category is dominated by statements using numerical language, and the "human narrative" category has many examples of MPs referring to the experiences of specific individuals. This again suggests that our measurement strategy plausibly captures our stylistic dimensions of interest.

Table S2: Top sentences for Affect, Positive Emotion, and Human Narrative

Affect	Positive Emotion	Human Narrative
Others eventually got jobs, although usually far less rewarding, far less secure and far less well paid.	As always, it is an enormous pleasure to follow the hon Member for Bootle , whose speeches are always entertaining and occasionally informative.	Moreover, what happens when an elderly brother and sister live together, or an elderly mother lives with her elderly son?
In others, everyone seems a little depressed - perhaps not greatly upset but a little depressed none the less.	It is always a pleasure to listen to Members' maiden speeches, and I enjoyed his as well.	Last week a friend of mine who works with elderly residents in Ogmore visited four elderly residents in one day.
Some of us believe that the legislation is profoundly unacceptable, profoundly wrong and profoundly damaging to our country.	I am always excited and in a state of eager anticipation to hear what the right hon Gentleman has to say on everything.	Anyone whose wife or partner had a child 20 years ago will remember that the woman spent a week to two weeks in hospital.
We also need to stop trying to blame someone every time something bad happens: sometimes bad things happen and they are no one's fault.	I join hon Members across the House in wishing a happy Pride to all those celebrating London Pride this weekend.	However his father David suffered a stroke 13 years ago since when his mother Sarah has had to care for both son and husband.
Such serious problems have left many facing uncertainty, which can cause severe stress to people who already face incredibly challenging circumstances.	I begin this afternoon by wishing the Secretary of State a very happy birthday - I sincerely hope that it improves from here on.	I speak as someone whose father served in the Metropolitan police for 25 years and whose younger brother is a serving Metropolitan police officer.
Many mentally ill people face sad and painful lives with great courage - more courage than the rest of us may have.	I hope that I have the pleasure of listening to his own speech today, because I enjoy his speeches immensely.	American civilians took leave once every six months; British diplomats took leave every six weeks, for two weeks.
All of us are aware that the Labour party has trouble understanding aspiration and even more trouble in rewarding aspiration.	I also congratulate my hon Friend the Member for Blackpool, North on his most amusing, entertaining and sincere maiden speech.	I have also discovered that a person called Mr Richard Shires subsequently became a paid constable in West Yorkshire police and continues to serve to this day.
Is it any wonder that mentally ill people desperate for help just get lost, sometimes with tragic consequences?	Today's debate has been extremely lively, interesting and, at times, amusing and much good wit and humour have made it a delight.	On 13 March 1942, in New End hospital, the older brother that I never knew, James John Dromey, died at three days old.
But neither can anyone underestimate the anger and sadness among people that things should ever have been allowed to get into this position.	I had a great surprise last Christmas when I received both a birthday card and a Christmas card from John and his family.	On Monday this week, another south Birmingham MP and I met South Birmingham primary care trust to talk about the situation in south Birmingham.
Experience over many years has shown us that that difficulty can too often lead to tragic loss of life.	It was wonderful to hear the shadow Chancellor - it is always wonderful to hear the shadow Chancellor in his marvellous speeches - explaining how cross-party he was.	Yes, another day, another Home Office statement and, sadly, yet another similar response from the shadow Home Secretary.

Table S3: Top sentences for Aggression, Fact, and Negative Emotion

Aggression	Fact	Negative Emotion
I found the attitude of the Conservatives' motion not only hypocritical and incoherent, but profoundly cynical and dishonest.	None the less, social security on average now costs every working person nearly £15 every working day.	People understandably already feel fraught and upset - they are in a situation that they never anticipated, and feel vulnerable and sometimes deeply hurt and angry.
That statement is as barbaric as it is downright stupid; it is nothing more than an ignorant, cruel and deliberate misconception to hide behind.	The growth rate figures are substantially different from the growth rate figures produced in the Budget just four months ago.	However, the indignity, discomfort and inconvenience caused to Brian during this episode understandably left him feeling demoralised and, in his words, depressed.
It is grossly irresponsible and, I am afraid, profoundly and disturbingly misleading, and even ignorant, to go around doing that.	The maximum figure for those costs was $91 billion, although the real extra costs amounted to $26 billion.	This is deeply worrying for families living in those blocks, and is causing huge anxiety, fear and insecurity.
There is something horrible, vindictive and cowardly about the Government's intolerant and ignorant attack on a small minority".	I have primary schools receiving less than £3,500 per pupil and secondary schools receiving less than £4,600 per pupil.	Such serious problems have left many facing uncertainty, which can cause severe stress to people who already face incredibly challenging circumstances.
Of course the situation in Zimbabwe is disgraceful and we condemn utterly the barbaric attacks on farmers, which are totally unacceptable.	Recent figures show the current account deficit running at the much lower level of £0.5 billion per month.	It can cause misery and pain for individuals and their families through serious disease or, worse, death.
They should not be all about blaming people, because blaming individuals for errors and mistakes is unhelpful and counter-productive.	We now spend nearly £11 billion extra each year on pensioners, and almost half that additional spending goes to the poorest third.	In addition to suffering horrendous physical injuries, enormous physical stress and emotional trauma, they had enormous financial stress.
Some of us believe that the legislation is profoundly unacceptable, profoundly wrong and profoundly damaging to our country.	The five Conservative speakers took three hours, five minutes; the six Labour speakers took one hour forty-five minutes.	Children described the extreme distress they experienced: losing weight, having nightmares, suffering from insomnia, crying frequently and becoming deeply unhappy.
Worse even than the failure publicly to criticise and condemn has been the United Kingdom Government's tendency almost to excuse.	They would produce sentences of seven months, six days or nine months, six days and various split months and split days.	If people feel isolated, depressed, lonely, jobless and skill-less, they will feel worse in hospital.
To claim that the financial crisis was somehow caused by the Labour party's mismanagement is complete and utter nonsense.	Approximately 100 people per 1,000 currently receive disability living allowance, compared with 50 people per 1,000 in Britain.	To the families we say: we are deeply sorry for your loss and deeply sorry for the pain you have suffered.
If that happens because of an arrogant and incompetent subordinate should not that arrogant and incompetent subordinate be fired?	Working in early years or later years care in private services means earning minimum wage or minimum wage plus.	Their anger is the anger of pain, the anger of discrimination, and the anger of lack of understanding, as well as the anger of frustration.

Human validation task

In this section, we provide results from a human validation task which assesses whether our text-based measures of style mirror human judgements of the same concepts. We wrote a web app which presented two research assistants with pairs of sentences (sampled from all sentences in our corpus). Coders were asked to complete two tasks. First, a style-*comparison* task required them to select which of the two sentences was more typical of a particular style. Second, a style-*intensity* task required them to rate the degree to which each sentence was representative of the selected style on a 5 point scale.

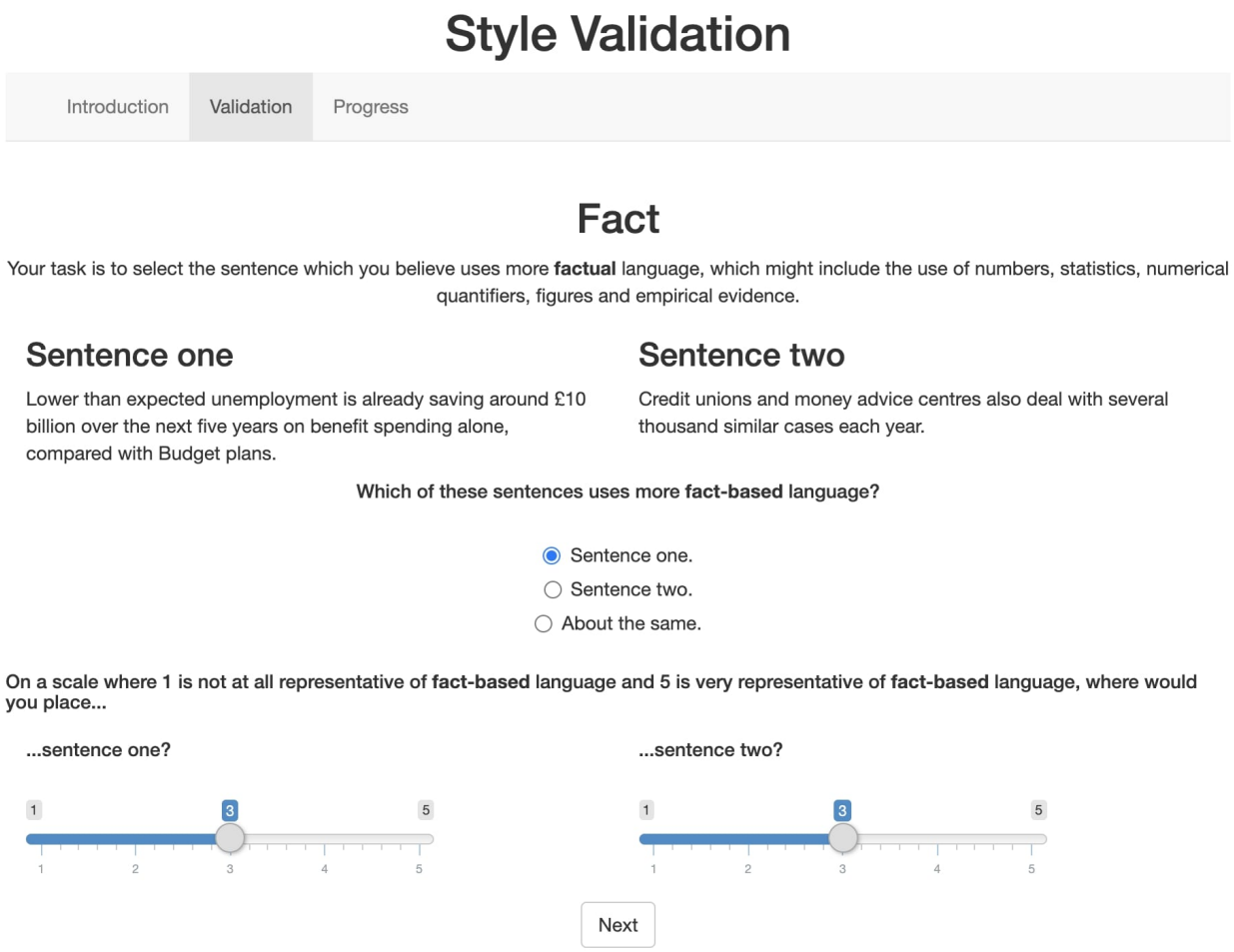

Figure S3: Human validation task prompt

Figure S3 gives an example of the prompt seen by our coders. In addition to the sen-

Table S4: Correlation between text-based measures and human judgments.

Style type	Comparison task	Intensity task
Human Narrative	0.67 (0.5)	0.7 (0.45)
Affect	0.62 (0.46)	0.61 (0.48)
Positive Emotion	0.7 (0.38)	0.71 (0.34)
Negative Emotion	0.75 (0.47)	0.75 (0.45)
Fact	0.77 (0.71)	0.81 (0.74)
Aggression	0.66 (0.32)	0.73 (0.22)
Complexity	0.83	0.85
Repetition	0.8	0.82

tences themselves, we presented coders with minimal definitions of the speech-styles of interest to ensure that the human coding related to the style dimensions identified in the literature review.

Each coder completed 70 comparisons per style, on average, meaning that we have on average 140 individual sentence-ratings per style. We use the distribution of responses to these tasks and compare them to the distribution of text-based style measures described in the main body of the paper for the same sentences as seen by the coders.[2]

We summarise the results in table S4. The "intensity task" column presents the correlation between our sentence-level style measures (equation S2) and our coders' ratings of the same styles. For the "comparison task" column, we calculate the difference in the sentence-level scores for each pair of sentences, and correlate that with the choices made by our coders from the comparison task.

Overall, the results are very encouraging. Across all styles, the correlation between the text-based scores and the human validation is always positive and is never lower than 0.61 for either task. These results suggest that there is a clear correspondence be-

[2]To assess inter-coder reliability, our research assistants both coded an additional common set of 20 comparisons per style. Coders agreed on which of the two sentences was more representative of a given style in 75% of comparisons. The correlation for the "intensity" scores for all sentences across coders was 0.8.

tween the measures of style implied by our text-analysis approach, and human judgements of those concepts in the same set of texts.[3]

Moreover, we can compare our measures with standard dictionary-based measurement approaches. For all styles except for repetition and complexity, we compare our word-embedding approach to an approach that measures style using the proportion of words in each sentence that appears in a pre-defined dictionary. This measurement strategy is more typical of existing applications of dictionaries in political science, and forms the basis of the analysis in several previous studies on gender and political style (e.g., Gleason, 2020; Jones, 2016; Yu, 2013). To maximise comparability, the dictionaries we use for this analysis are the same as the seed dictionaries we use to construct our word-embedding scores:

- *Affect* – Linguistic Inquiry and Word Count 2015 (Affect) (Pennebaker et al., 2015)
- *Fact* – Linguistic Inquiry and Word Count 2015 (Number and Quantitative) (Pennebaker et al., 2015) and all occurrences of any numeric figures
- *Positive Emotion* – Regressive Imagery Dictionary (Emotions: Positive Affect) (Martindale, 1990)
- *Negative Emotion* – Regressive Imagery Dictionary (Emotions: Anxiety and Sadness) (Martindale, 1990)
- *Aggression* – our bespoke dictionary of words shown in figure S1
- *Human Narrative* – our bespoke dictionary of shown in figure S2 and the 200 most common names of children born between 1970 and 2019.

This means that, for each sentence in our corpus, we have a measure of style based

[3]As repetitiveness is a quantity that manifests more clearly *across* rather than *within* sentences, our sentence-based human validation is somewhat less well suited to evaluating this concept. Nevertheless, the sentences that our measure marks as most repetitive do clearly demonstrate high levels of repetitiveness, and, as table S4 indicates, even though detecting repetitiveness at the sentence-level might represent a hard task, we recover a clear correspondence between our measures and human judgements of the same concept.

on our word embedding method (described in equation 1 in the paper), and a measure of style based on counting the fraction of words in the sentence that fall into the relevant style's seed dictionary.

The results are given in table S4. The numbers in parentheses show the correlation between the standard dictionary measure of style described above, and human judgements provided by our coders. Our word-embedding approach clearly outperforms standard dictionary approaches in approximating human judgement. For instance, for positive emotion, standard dictionary measures correlate at 0.38 and 0.34 with human codings for the two tasks, compared to 0.7 and 0.71 for the word-embedding approach. Despite the relatively small sample sizes, the magnitude of the difference in predictive power means that – in all cases except for "fact" – the correlation between our word-embedding measures and human codings is significantly higher than the equivalent correlation for standard dictionary measures.[4] Overall, this exercise provides strong evidence that we can reliably detect our styles of interest in parliamentary speech and outperform the standard measures used in previous studies on gender and political style.

[4]We deterime this difference by using a bootstrap proceedure, in which we sample from our set of sentences 2000 times with replacement and calculate the correlation between our word-embedding measures and human codings, and between the dictionary measures and human codings, on each iteration. We can easily reject the null hypothesis of no difference in these correlations for all styles except for the "fact" dimension.

Controlling for individual-level covariates

In this section we show results of the alternative specification for the dynamic hierarchical model described in the paper in which we expand the model at the second level by including a vector of individual-level covariates, $X^k_{j,t}$:

$$\alpha_{j,t} \sim N(\mu_{0,t} + \mu_{1,t} Female_j + \sum_{k=1}^{k} \lambda_k X^k_{j,t}, \sigma_\alpha) \tag{S3}$$

where $X^k_{j,t}$ includes:

- Party (categorical: Conservative; Labour; Liberal Democrat; Other)
- Government or opposition party status (binary)
- Government or opposition frontbench position (binary)
- Committee chair (binary)
- MP age (in years, continuous)
- Margin of victory in prior election (percentage points, continuous)
- University degree (binary)
- Prior occupation (categorical: manual; professional; political; business; other)

We transform the two continuous predictors such that they have mean zero, and standard deviation one. We present the results for our main quantities of interest ($\mu_{1,t}$) estimated from this model in figure S4.

The figure shows that, in general, we recover very similar patterns of gender differences in style use over time when controlling for individual-level covariates. For human narrative, affect, positive emotion, negative emotion, fact, and aggression the trajectories of the gender differences over time are very similar to those presented in the main body of the paper. The largest differences are for complexity and repetition, where the

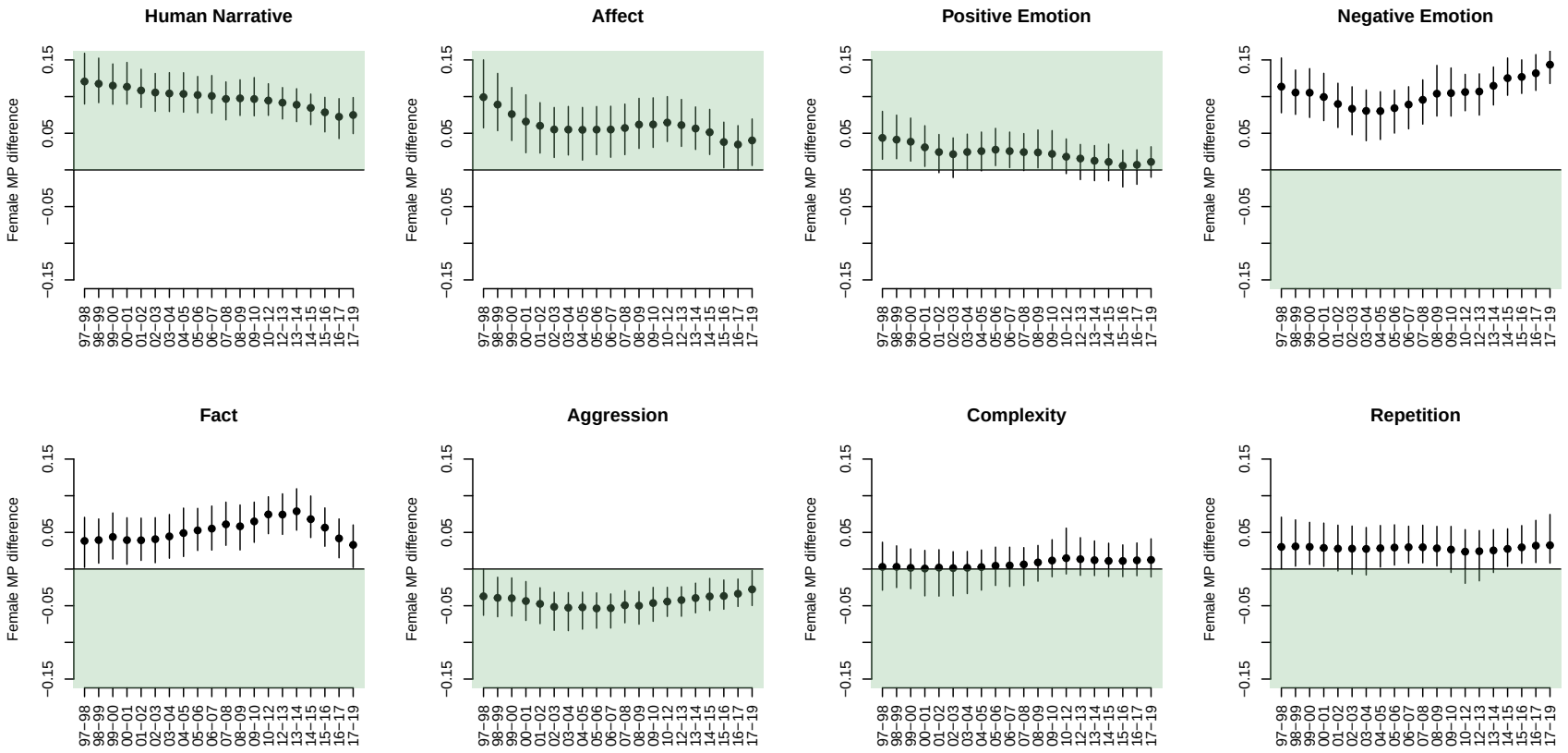

Figure S4: Gender differences in style over time controlling for individual-level confounders

pattern of convergence between men and women is somewhat attenuated in the estimates from the alternative specification. For complexity in particular, the large shift in the gender difference that we observe between 2008 and 2013 is confounded by some of the individual-level covariates, as the gender difference is largely constant (and indistinguishable from zero) for the entire time period once we control for these other factors. Nevertheless, overall, these results suggest that while other MP-level characteristics clearly account for some variation in style use, our central finding – that the debating styles of male and female MPs have diverged from gender-based stereotypes over time – is not affected by these estimates.

Figure S5 presents the estimates for each of the individual-level covariates for each style. Although these are not our primary quantities of interest, there are several patterns that are of substantive interest. First, we find, consistent with other work (Proksch et al., 2019), that MPs from government parties use significantly less negative and more positive language than MPs from opposition parties. Government MPs are also less aggressive and tend to rely more on human narrative and less on fact-based arguments than their opposition counterparts. Second, compared with backbench MPs, politicians in leadership positions are less likely to use human narrative, more likely to make fact-based arguments, use substantially less emotive language, and are more repetitious in their speeches. We also see some evidence of partisan differences. Compared to Conservative Party MPs, Labour MPs use more human narrative, more factual language, and are somewhat less complex in their speeches. Liberal Democrat MPs, by contrast, make less use of human narrative, more use of fact, and are substantially less aggressive than Conservative MPs. There are also interesting patterns in speech styles according to the education and occupation variables. For instance, university-educated MPs tend to make less use of human narrative, and less use of negative emotional language, but deliver speeches that are more complex and more repetitious than their non-university edu-

cated counterparts. With regard to prior employment, MPs from manual occupations do appear to have distinct speechmaking styles, as they employ more human narrative, and less aggressive and repetitive language than MPs from other employment backgrounds. Overall, it is clear that there are many factors that influence the political styles that MPs adopt and, while these are not directly relevant to the substantive questions in our study, we think that these findings may be profitably investigated in future work.

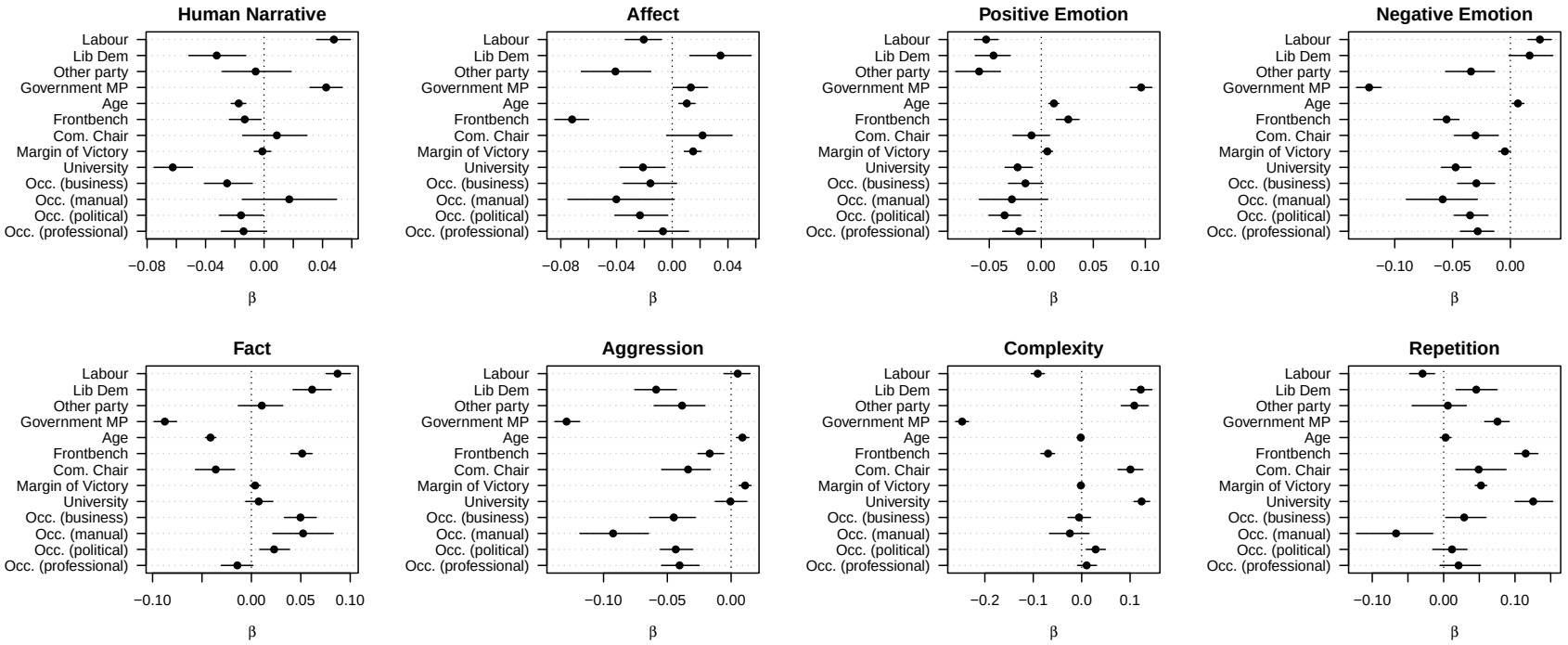

Figure S5: Individual-level covariate effects

Style use and debate-type

Our model accounts for aggregate differences in style use across debates via the δ_d random-effects described in equation 2 in the main body. The inclusion of these parameters means that gender differences in style use cannot be attributed to men and women participating in systematically different types of debates, as the gender effects we estimate are based on within-debate variation in the style outcomes. However, it is possible that the magnitude of gender differences nevertheless varies across debates of different types. We investigate this possibility here. Specifically, we separate the debates in our data into common types that occur regularly in the UK House of Commons (for more detail, see Blumenau and Damiani, 2020):

1. **All**: all debates in our dataset.
2. **Ministerial Question Time**: the routine questioning of Ministers, occurs four times a week.
3. **Prime Minister's Question Time**: the Prime Minister answers questions from the Leader of the Opposition, opposition members and government backbenchers, occurs once a week.
4. **Procedural debates**: a compound category that includes debates that are not substantive in nature, but deal with matters of parliamentary procedure or scheduling. For example, Business of the House or Points of Order.
5. **Legislation**: debates on legislation, includes all stages of the process that occur in the Commons' chamber, such as second and third reading.
6. **Opposition Days and Backbench Business**: this includes business for debate that is placed on the parliamentary agenda by opposition members or backbenchers.
7. **Other**: all other forms of debate that are not captured by the above categories.

This categorisation captures important substantive differences between different types

of debates in the House of Commons, some of which have been shown to be predictive of MPs' style in previous work (Osnabrügge, Hobolt and Rodon, 2021).

We run a series of OLS models for each of our outcomes, where our main explanatory variable of interest is the gender of the MP, and where we also control for party, age, years in parliament, margin of victory in the previous election, degree education, previous occupation, and whether the MP was a) a member of the cabinet, b) a member of the shadow cabinet membership, c) a government minister, d) a shadow minister, or e) a committee chair. For each outcome, we subset the data to only debates of a certain type, estimate the model, and record the coefficient on the gender variable at each iteration. Figure S6 shows, for each style, the gender differences in the seven different debate types.

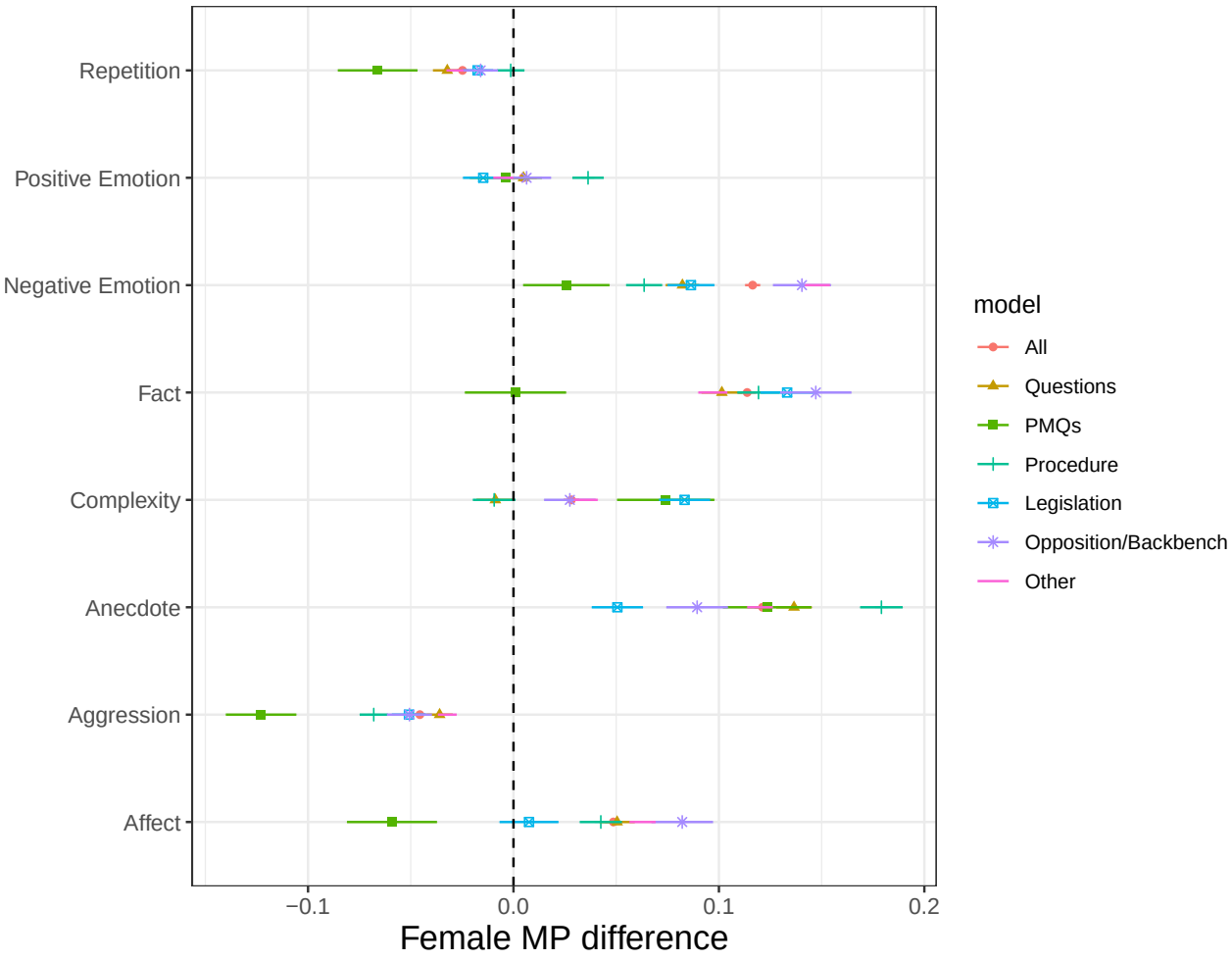

Figure S6: Debate type models

The analysis reveals that the magnitude of average gender differences are relatively constant across the debate types. In the debate types we identify, Prime Minister's Questions seems to be the only type of debate that significantly effects the gender coefficients. We see that, relative to the model which pools across all debates, the magnitude of gender differences is increased for repetition, aggression, and affect; decreased for negative emotion; and reduces gender differences in fact to statistically indistinguishable from zero. Overall, however, while there is some variation in the magnitude of gender differences across debate types, these differences are for the most part very small.

In figure S7 we show additional descriptive information on the average level of each style in speeches used across the different debate types. The patterns in style use across

debates generally conform with standard intuitions. For instance, the figure shows that both Question Time and Prime Ministers Questions (PMQ) debates are substantially less positive than debates on legislation, which is consistent with the idea that these settings are used by the opposition parties to interrogate – and often castigate – the government on issues of the day. Similarly, both PMQ debates and debates initiated by the Opposition parties in parliament are more aggressive than other debates, which again follows the intuition that these debates are mainly used as a vehicle for criticising government policy. In general, these descriptive figures bolster the results from our validation exercises above, as they imply that our measures accurately capture expected differences in speech style across different types of parliamentary debate.

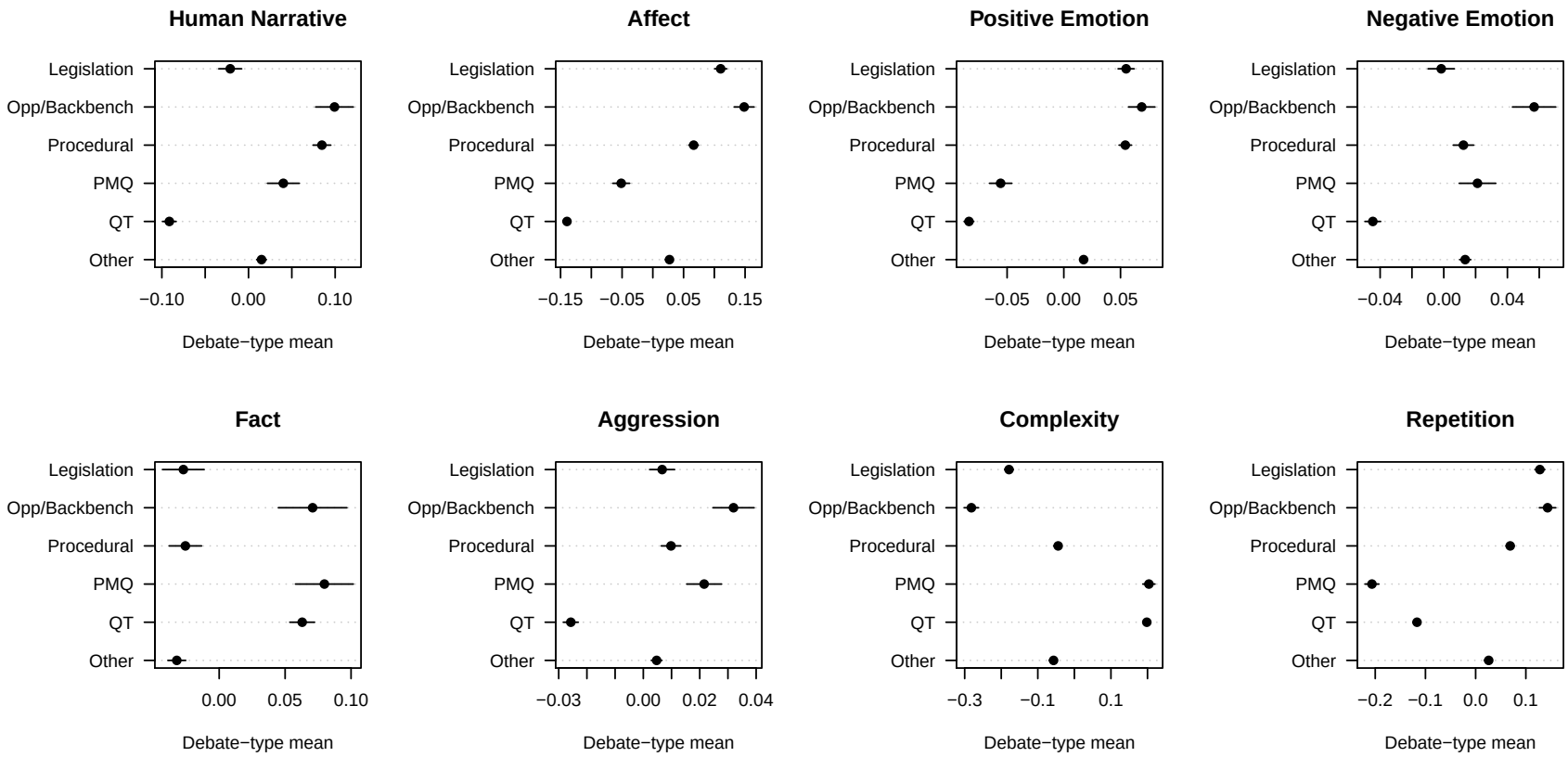

Figure S7: Style type average by debate type

Within-MP and replacement effects

Does gender explain less variation in aggregate style use over time because of a gradual convergence in styles of female and male MPs throughout their careers in parliament? Or do gender gaps decrease because the men and women entering parliament over time are systematically different from those leaving parliament? Which of these two explanations – which we refer to as "within-MP" and "replacement" effects – is responsible for the aggregate patterns we document in the main body of the paper? Our modelling approach allows us to decompose the evolving gender differences that we report in the section above into these two mechanisms of change.

Given the model described by equations 2 and 3 in the main body of the paper, we can decompose the shifting patterns of gendered style use into those changes that stem from within-MP change over time, and those that come from replacement. Our goal is to specify a decomposition of $\mu_{0,t} - \mu_{0,t-1}$, which is the change in average style use for men between parliamentary session t and session $t-1$ (we can then provide an equivalent approach for female MPs). We begin by distinguishing between three types of MP, which we label as "remainers", "joiners", and "leavers":

- J_m^R is the set of male MPs who appear in both session t and $t-1$ (Remainers)
- J_m^J is the set of men who appear in t but not in $t-1$ (Joiners)
- J_m^L is the set who appear in $t-1$ and not in t (Leavers)

We also will require the fraction of men who are "remainers" in t and $t-1$:

- π_t^R is the fraction of male MPs in t who also served in $t-1$
- π_{t-1}^R is the fraction of male MPs in $t-1$ who also served in t

Note that the proportion of male MPs who are "remainers" in t may be different from the proportion in $t-1$, because some male MPs who leave parliament in $t-1$ will be replaced by women in t (and vice versa).

Given these definitions, we can write the mean style use for men in each period as a function of the MP-period effects ($\alpha_{j,t}$):

$$\mu^m_{0,t-1} = \underbrace{\pi^R_{t-1}\frac{1}{|J^R_m|}\sum_{j\in J^R_m}\alpha_{j,t-1}}_{\text{Remaining MPs}} + \underbrace{(1-\pi^R_{t-1})\frac{1}{|J^L_m|}\sum_{j\in J^L_m}\alpha_{j,t-1}}_{\text{Leaving MPs}} \tag{S4}$$

$$\mu^m_{0,t} = \underbrace{\pi^R_{t}\frac{1}{|J^R_m|}\sum_{j\in J^R_m}\alpha_{j,t}}_{\text{Remaining MPs}} + \underbrace{(1-\pi^R_{t})\frac{1}{|J^J_m|}\sum_{j\in J^J_m}\alpha_{j,t}}_{\text{Joining MPs}} \tag{S5}$$

Here, $\mu_{0,t-1}$ is a weighted average of the finite-sample average of the "remainers" and "leavers" in $t-1$, where the weights are given by the relative proportion of those groups in that parliamentary session. $\mu_{0,t}$ is constituted from the equivalent averages for "remainers" and "joiners" in time period t, again weighted by the size of those two groups in t.

Taking the difference between S4 and S5 and rearranging reveals an additive decomposition which separates the two effects of interest:

$$\begin{aligned}\mu^m_{0,t} - \mu^m_{0,t-1} = & \underbrace{\pi^R_{t}\frac{1}{|J^R_m|}\sum_{j\in J^R_m}\alpha_{j,t} - \pi^R_{t-1}\frac{1}{|J^R_m|}\sum_{j\in J^R_m}\alpha_{j,t-1} +}_{\text{"Within-MP" effect } (S_m)} \\ & \underbrace{(1-\pi^R_{t})\frac{1}{|J^J_m|}\sum_{j\in J^J_m}\alpha_{j,t} - (1-\pi^R_{t-1})\frac{1}{|J^L_m|}\sum_{j\in J^L_m}\alpha_{j,t-1}}_{\text{"Replacement" effect } (R_m)}\end{aligned} \tag{S6}$$

We denote the within-MP effect for men as W_m and the replacement effect as R_m. We

can also, of course, define the same quantities for female MPs, and therefore can describe the changing gender difference in terms of replacement and socialisation effects:

$$(\mu^{w}_{0,t} - \mu^{m}_{0,t}) - (\mu^{w}_{0,t-1} - \mu^{m}_{0,t-1}) = \underbrace{(W_w - W_m)}_{\text{“Within-MP” difference}} - \underbrace{(R_w - R_m)}_{\text{“Replacement” difference}} \tag{S7}$$

Turning to our results, we plot these quantities in the left (for male MPs) and centre (for female MPs) panels of figure S8. The x-axis describes the average direction and magnitude of changes between parliamentary sessions for each style for men and women, respectively. The right-hand panel reports *the difference* in the effects for women and men. In each panel, hollow points show changes that occur because of replacement, and solid points show changes that occur due to within-MP shifts.

We use these plots to understand whether replacement or within-MP change is a stronger determinant of the aggregate shifts we observe. Overall, neither differential replacement nor within-MP change alone explain the convergence that we document across multiple different styles in the main body of the paper, though there is some evidence that replacement is more important as a mechanism for explaining the changing gender dynamics we observe for the “agentic” styles while within-MP change is somewhat more important for explaining change for more “communal” styles.

For example, figure 2 in the main body of the paper shows that women are much more likely than men to use negative emotion in their speeches in later years, but only somewhat more likely in the earlier years. The middle panel of figure S8 shows that the replacement effect for women for negative emotion is positive (the hollow point for negative emotion is greater than zero), which implies that newly elected women are more negative than the women leaving parliament, on average. However, the left panel of figure S8 suggests that this is *not* true for male MPs: male MPs joining parliament use negative language at the same rate on average as male MPs leaving parliament (the hol-

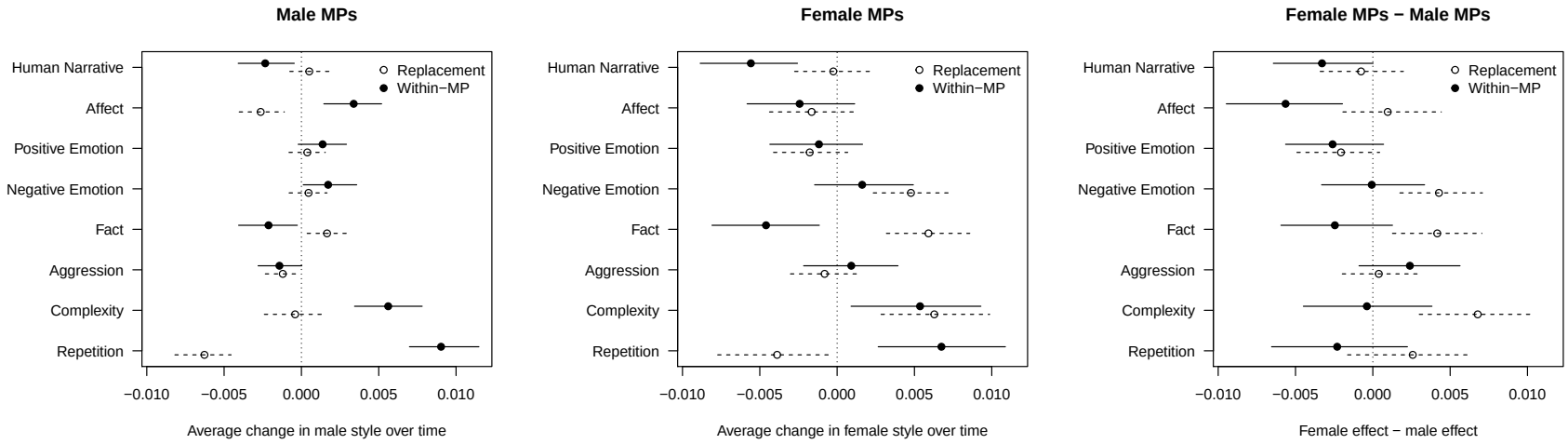

Figure S8: Within-MP and replacement over time change by gender

low point for negative emotion is close to zero). Consequently, the right panel suggests a (positive) differential replacement effect for negative emotion. Note that the difference between male and female *within-MP* effects is close to zero for negative emotion. This suggests that the divergence between men and women that we note at the aggregate level is almost entirely driven by differential replacement between male and female MPs, rather than existing MPs becoming more alike in their behaviour over time.

The right-hand panel of S8 indicates that, beyond negative emotion, replacement effects also account for a greater share of the aggregate change in gender differences for factual language and complexity. For both, while the women entering parliament are significantly more likely to use these styles than the women leaving, newly elected male MPs employ these styles at broadly similar rates as the men that they replace. By contrast, both male *and* female MPs are less likely to use factual language as their careers in parliament progress, and the speeches of both men *and* women become more complex the longer they spend in parliament. Consequently, the large aggregate shifts that we observe for these styles are largely driven by the fact that the women newly elected to parliament adopted a legislative debating style that was more factual, complex, and negative than the women they replaced.

For other styles, we see that within-MP change accounts for a greater share of the variation in gender differences. For instance, the gradually decreasing gap in the use of human narrative in figure 2 in the main body of the paper is mostly attributable to women MPs using this style less the longer that they stay in parliament, but the decreasing use of human narrative for male MPs is much smaller. Similarly, on average women employ less positive emotion over time, whereas the positive language use of male MPs remains relatively constant. Conversely, within-MP change in affect for men is positive, implying that men become more emotional overall in their speeches over time, but there is very little average within-MP change in affect for women. These results imply that, for these

style types, the convergence that we see in the main analysis is driven by the different stylistic trajectories than male and female MPs appear to follow throughout their tenure in parliament.

Topic-based confounding

We present evidence of convergence between men and women with respect to several debating styles over time. One potential concern for the interpretation of our results is that the parliamentary agenda is not fixed, and changes to the set of issues under discussion may result in convergence between men and women even in the absence of behaviour change.

Consider, for instance, a style like human narrative, where we observe a large convergence between men and women over time. Women are significantly more likely to use human narrative in their parliamentary speeches at the beginning of the time period than they are at the end. If, however, women are more likely to use human narrative than men in certain *topics*, and those topics become less prevalent over time, then the convergence we document might in fact be attributable to changes to the parliamentary agenda. For changes in topic prevalence to be responsible for convergence, it would have to be the case that the topics on which we observe women using *more* human narrative than men are becoming *less* prevalent, or that the topics on which we see women using *less* human narrative than men are become *more* prevalent over time. For example, perhaps women use more human narrative than men when discussing education policy, and education policy is more frequently discussed in the early period in our data than the later period in our data. If this were true, then our results might be subject to topical confounding, as changes in topical prevalence over time would account for the aggregate changes we observe in the main analysis.

To address this concern, in this section we use statistical topic models to evaluate whether topics on which we observe notable stylistic differences between men and women become more or less prevalent over time. We begin by estimating a correlated topic model (Blei and Lafferty, 2006) (CTM) for all speeches in our data. The CTM is an

unsupervised learning approach which assumes that the frequency with which words co-occur within different speeches provides information about the topics that feature in those speeches. As with other topic models, the CTM requires the analyst to choose the number of topics, K. Given that our results might be sensitive to this choice, we choose to present results from a series of models, where we vary the number of topics: $K \in 10, 20, ..., 80$. We implement the CTM as the null form of the Structural Topic Model, which we implement in R (Roberts et al., 2014).

The key output of the topic model is θ, a $N * D$ matrix of topic proportions that measures the degree to which each speech (i) in the data features each of the estimated topics (d). $\theta_{i,k}$ therefore gives the proportion of speech i devoted to topic d. With these topics in hand, we then evaluate – for each of our 8 styles – the size of the stylistic gender gap between men and women on each topic. To do so, we estimate models where we interact the gender of the MP delivering a speech with the topic proportions that pertain to that speech:

$$y^s_{i(j)} = \alpha + \beta^1 Female_j + \sum_{k=2}^{K} \beta^2_k \theta_{i,k} + \sum_{k=2}^{K} \beta^3_k (Gender_i \cdot \theta_{i,k}) + \epsilon_{i(j)} \tag{S8}$$

We use the coefficients of this model to calculate estimated average differences between men and women on speeches devoted to each topic, which we denote as:

$$\delta^s_k = \begin{cases} \beta^1 & \text{if } k = 1 \\ \beta^1 + \beta^3_k & \text{if } k \neq 1 \end{cases} \tag{S9}$$

The average difference in style s between men and women on speeches that are entirely devoted to topic 1 is given by β^1 (i.e. the baseline), and $\beta^1 + \beta^3_k$ captures the average gender difference in style on speeches entirely devoted to topic k. We denote the gender difference on each topic and style as δ^s_k. This specification allows us to capture the

aggregate differences between male and female use of a style on each topic. Positive values for δ_k^s indicate that women use the style more than men in a given topic, and negative values suggest that women use the style less than men in a given topic.

We then estimate a second set of regression models to capture, for each topic, the relationship between time and topic prevalence. To do so, we first multiply the number of words in each speech by the vector of topic proportions for that speech, giving us the weighted number of words dedicated to a given topic for each speech in the data. We then sum these topic-weighted word counts across all speeches within a given calendar month, and use the summed word counts as the dependent variable for regressions of the form:

$$y_t^k = \alpha + \gamma_k YearMon_t + \epsilon_t \tag{S10}$$

Here, y_t^k is the number of words on topic k in time period t, and γ_k captures the linear relationship between time and topic prevalence for topic k. Positive values of γ_k imply that topic k becomes more prevalent in parliamentary debate throughout the study period, and negative values suggest that the topic becomes less prevalent over time.

If the topical confounding argument is correct, then for a style like human narrative – where we observe average convergence between men and women over time – it must be the case that there is a negative relationship between the gender gap on that topic and the relationship between topic and time. That is, topics where women use human narrative more than men (positive coefficient from equation S8) should be becoming less prevalent over time (negative coefficient from equation S10).

The topical-confounding hypothesis implies different relationships between topical gender-gaps and changes in topic prevalence over time for different styles. For instance, for human narrative, our main analysis shows that women are more likely to use this

style in the early period of our data and less in the later period. For this style, topical confounding would occur if topics where women use narrative *more* on average than men (positive δ_k^s from equation S9) became *less* prevalent over time (negative γ_k from equation S10), or the topics where women use narrative *less* than men (negative δ_k^s from equation S9) became *more* prevalent over time (positive γ_k from equation S10). For human narrative, then, the topical-confounding hypothesis implies a negative relationship between the two sets of coefficients.

On the other hand, our aggregate results suggest that women are less aggressive than men in the early period of the data but are equally as aggressive later in the period. Accordingly, if this convergence can be explained by changes to the topics under discussion, it must be the case that the topics on which women tend to be less aggressive than men (negative δ_k^s from equation S9) become less prevalent over time (negative γ_k from equation S10), or that the topics on which women tend to be more aggressive than men (positive δ_k^s from equation S9) become more prevalent over time (positive γ_k from equation S10). Therefore, for aggression, the topical confounding hypothesis implies a positive relationship between the two sets of coefficients.

Following this logic through all eight style types, the topical-confounding explanation suggests that we should observe a positive relationship between γ_k and δ_k^s for aggression, complexity, fact and negative emotion, and a negative relationship between γ_k and δ_k^s for human narrative, affect, positive emotion, and repetition.

In figure S9 we evaluate these expectations by plotting the estimated values of γ_k and δ_k^s against each other for each style. In this plot, each point represents a single topic from our $K = 40$ topic model: the x-axis measures the gender gap in the use of a given style (δ_k^s), and the y-axis measures the changing prevalence of the topic over time (γ_k). We also fit a regression line between the sets of coefficients, which is coloured in red if the slope of the line is associated with a p-value of less than 0.05, and otherwise

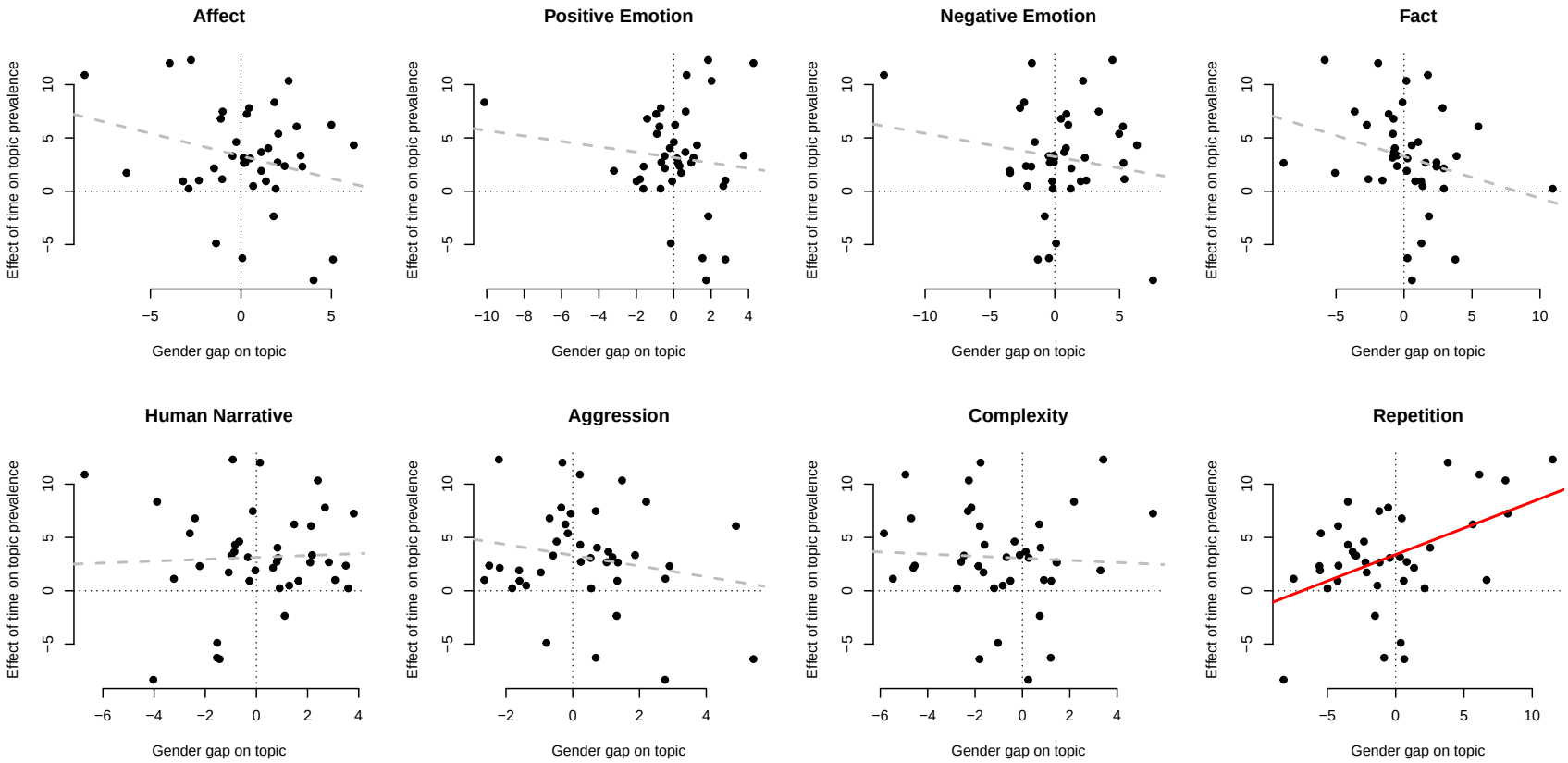

Figure S9: **Topical-confounding**: The figure shows the relationship between the gender gap in the use of a given style on a given topic (x-axis), and the change in the prevalence of a given topic over time (y-axis).

is coloured in grey.

The main implication of this analysis is straightforward: we find very little evidence to support the topical-confounding hypothesis. The size of the gender gap measured for a given style on a given topic largely does not predict the degree to which that topic becomes more or less prevalent over time. For three of the styles – aggression, negative emotion, and fact – the relationships in figure S9 are negative, where they would need to be positive for topical changes to explain the stylistic convergence we document in the main body of the paper. We also find a relationship that is in the "wrong" direction for repetition (that is, although statistically significant, the relationship would need to be negative to cause concern), and there is also essentially no relationship between the gender gap in human narrative on different topics and the changing prevalence of those topics over time. For the remaining styles – affect, positive emotion, and complexity – we do find some evidence that topics on which women display more of these styles become more prevalent over time, but the relationships are very noisy and in none of those cases are we able to reject the null hypothesis of a relationship of zero.

As there is no *a priori* reason to base our inferences on the $K = 40$ topic model, in figure S10 we summarise the relevant results from all 8 topic model specifications. In this plot, the x-axis measures the value for K, and the y-axis measures the slope of the regression line for the changing prevalence of a topic over time (γ_k) as a function of the gender gap in the use of a given style in that topic (δ_k^s). The results clearly demonstrate that our findings are not sensitive to the number of topics used in the analysis. For all models, we find patterns that are very similar to those depicted in figure S9. The only exception is that we find a significant coefficient for the "fact" style in the $K = 80$ topic model. However, again, this relationship is in the "wrong" direction as it suggests a negative relationship between the topic-specific gender-gap in factual language and over-time topic prevalence, where the topical confounding story implies a positive rela-

tionship between these quantities for the factual language style.

Taken together, these analyses imply that the aggregate patterns we observe in the main body of the paper cannot be convincingly explained by changes to the parliamentary agenda over time.

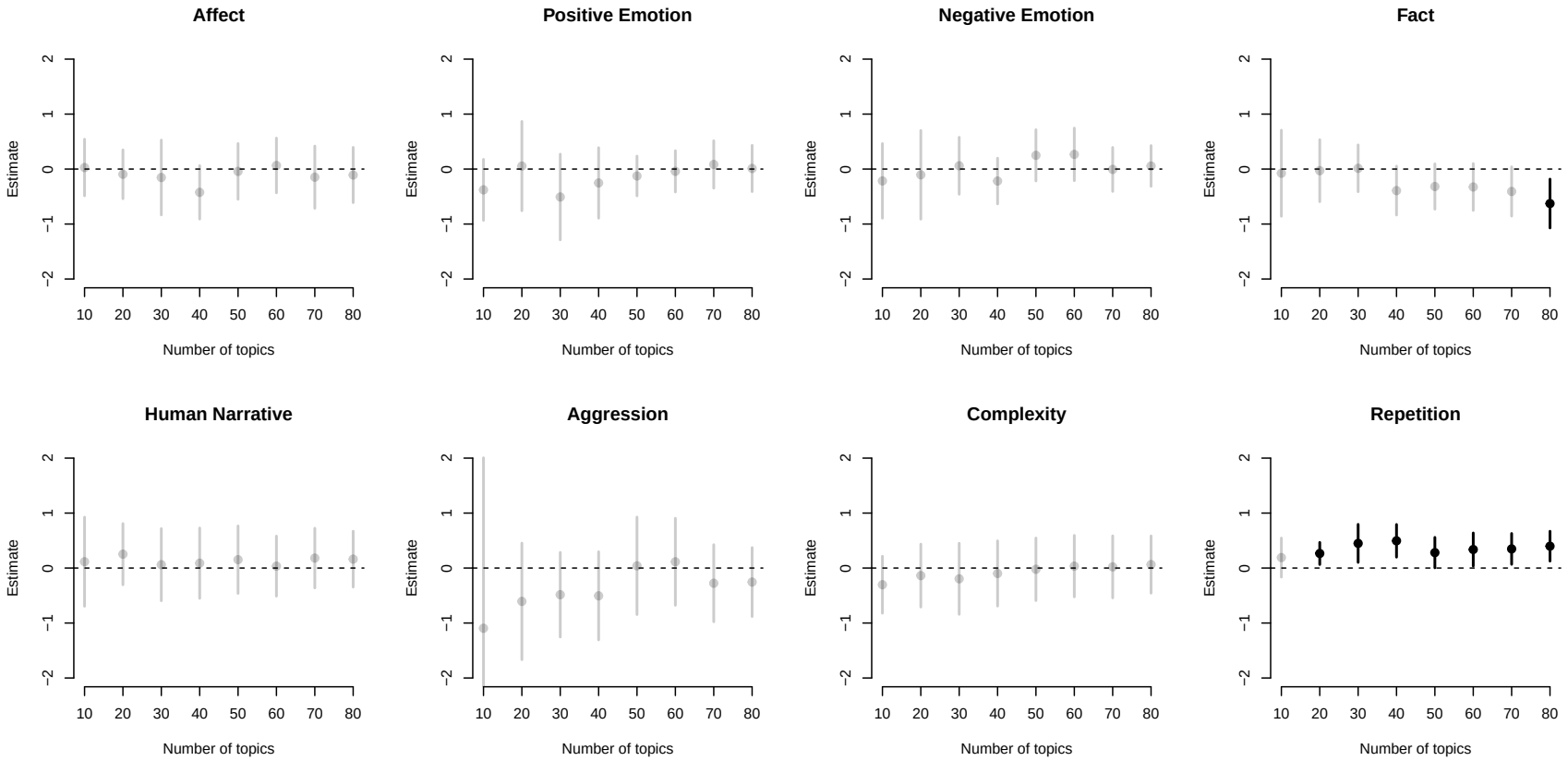

Figure S10: **Topical-confounding, varying** K: On the y-axis, the figure summarises the linear relationship between the gender gap in the use of a given style on a given topic (δ_k^s), and the change in the prevalence of a given topic over time (γ_k). The x-axis measures the number of CTM topics, K, used to estimate these relationships.

Style use and debate participation

Our results show that, on average, female MPs deliver speeches that are less likely to be marked by "communal" styles and more by "agentic" style over time. One potential alternative explanation for our results is that male and female MPs who employ different speaking styles might have become differentially likely to *participate* in parliamentary debate over time. We might imagine, for instance, that female MPs who tend to deliver highly agentic speeches gave more speeches in parliament over the course of the study period, and that women who tend to deliver highly communal speeches participated less in debate over time. If that were the case, differential participation might drive the changing gender speechmaking dynamics that we document in the paper, rather than within-MP changes.

To investigate this alternative explanation, we assess whether the average style of an MP across all speeches in a given parliamentary term predicts the number of speeches that the MP delivers. We begin by measuring the number of speeches delivered by each MP in each parliamentary term (# Speeches$_{i(t)}$), which we then model as a function of the gender of the MP, the average style of speeches given by the MP in that term ($Style^{s}_{i(t)}$), and the interaction between these two variables. Specifically, for each parliamentary term, t, and each style, s, we estimate a model of the following form:

$$\text{\# Speeches}_{i(t)} = \alpha + \beta_1 Female_i + \beta_2 Style^{s}_{i(t)} + \beta_3 (Female_i \cdot Style^{s}_{i(t)}) + \epsilon_{i(t)} \qquad \text{(S11)}$$

Our key quantities of interest here are β_2, which measures the effect of a standard deviation increase in the use of a given style on the number of speeches delivered by men, and $\beta_2 + \beta_3$, which gives the same quantity for female MPs. If our results are driven by a selection-based story about the types of MPs who choose to participate in debate, then we should find that these two quantities broadly mirror the aggregate patterns

we document in figure 2 of the paper. For example, if differential participation is the explanation for the decreasing average use of "human narrative" by female MPs, then we should observe a weaker relationship between the degree to which a female MP's speeches tend to feature human narrative and the number of speeches delivered by that MP over time. Similarly, for "negative emotion", if selection into debate drives the increasing use of that style by women, we would expect to see the relationship between the use of negative emotion and the number of speeches delivered by female MPs to have strengthened over time. We present our quantities of interest for each style in each parliamentary term in figure S11.

In general, we find very little evidence that the average style of an MP predicts participation in debate at any point during the study period. Across almost all styles, the effects are indistinguishable from zero, implying that it is very unlikely that our results are driven by which MPs choose to speak in debate. Moreover, there are no clear over-time trends in these coefficients, which undermines the idea that, for example, women with more agentic speaking styles participate more over time. In other words, this analysis suggests that the sample of speeches that we observe do *not* appear to be disproportionately delivered by the more "communal" female MPs in the early period, and by more "agentic" female MPs in the later period. Rather, this analysis suggests that the changes over time that we document in the paper are largely driven by within-MP changes in speaking style, and the replacement of MPs with different style-types over time (see figure S8 above).

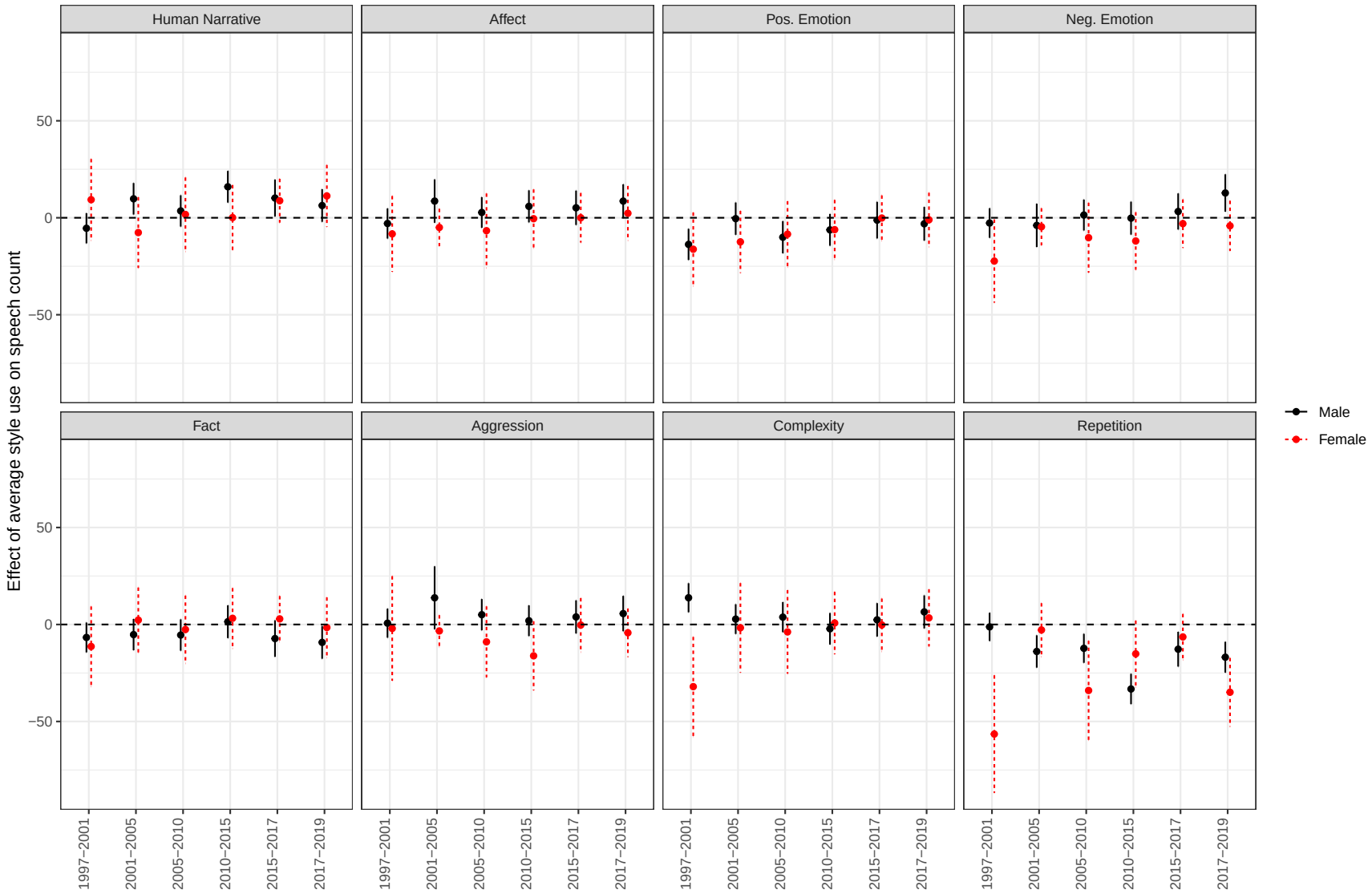

Figure S11: **Participation as a function of average style use, by parliamentary term**: The figure illustrates the average marginal effect of a one standard deviation increase in the average style use on the number of times an MP speaks in a given parliamentary term.

References

Blei, David and John Lafferty. 2006. "Correlated topic models." *Advances in neural information processing systems* 18:147.

Blumenau, Jack and Roberta Damiani. 2020. Parliamentary Debate in the UK House of Commons. In *The Politics of Legislative Debate*, ed. Hanna Bäck, Marc Debus and Jorge M. Fernandes. Oxford: Oxford University Press p. (Forthcoming).

Gleason, Shane A. 2020. "Beyond Mere Presence: Gender Norms in Oral Arguments at the U.S. Supreme Court." *Political Research Quarterly* 73(3):596–608.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21:267–297.

Jones, Jennifer J. 2016. "Talk "Like a Man": The Linguistic Styles of Hillary Clinton, 1992-2013." *Perspectives on Politics* 14(3):625–642.

Martindale, Colin. 1990. *The clockwork muse: The predictability of artistic change*. New York: Basic Books.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *Working Paper* pp. 1–12.

Osnabrügge, Moritz, Sara B. Hobolt and Toni Rodon. 2021. "Playing to the Gallery: How Politicians Use Emotive Rhetoric in Parliaments." *American Political Science Review* pp. 1–15.

Pennebaker, James W., Ryan L. Boyd, Kayla Jordan and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Austin, US: University of Texas at Austin.

Pennington, Jeffrey, Richard Socher and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1532–1543.

Proksch, Sven Oliver, Will Lowe, Jens Wäckerle and Stuart Soroka. 2019. "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches." *Legislative Studies Quarterly* 44(1):97–131.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley and Others. 2014. "stm: R package for structural topic models." *Journal of Statistical Software* 10(2):1–40.

Spirling, Arthur and Pedro L. Rodriguez. 2019. "Word Embeddings What works, what doesn't, and how to tell the difference for applied research." pp. 1–51.

Yu, Bei. 2013. "Language and gender in congressional speech." *Literary and Linguistic Computing* 29(1):118–132.

Zamani, Hamed and W. Bruce Croft. 2016. "Embedding-based Query Language Models." *Proceedings of the 2016 ACM international conference on the theory of informational retrieval* pp. 147–156.