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

Contents

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 ; acuse ; accusations; accusing ; anger ; angered ; annoyance ; annoyed ; attack ; insult ; insulting ; insulted ; betray; betrayed ; blame ; blamed ; blaming ; bitter; bitterly ; bitterness ; complain; complaining; confront ; confrontation; fibber; fabricator ; phony ; fibber ; sham ; deceived ; deceive ; disgrace; villain; good-for-nothing; hypocrite ; deception; steal ; needlessly; needless; criticise ; criticised ; criticising ; blackened ; fiddled; fiddle; problematic ; lawbreakers ; offenders; offend; unacceptbale ; leech; phoney ; appalling ; incapable ; farcical ; absurd ; ludicrous; nonsense ; laughable ; nonsensical ; ridiculous; outraged ; hysterial ; adversarial ; aggressive ; shady ; stereotyping; unhelpful ; unnatural ; assaulted ; assault ; assaulting ; half-truths ; petty; humiliate ; humiliating ; confrontational; hate ; hatred ; furious ; hostile ; hostility ; nasty; obnoxious ; sleeze; sleezy ; inadequacy; faithless; neglectful ; neglect; neglected; wrong ; failure ; failures ; failed ; fail ; scapegoat ; cruel; cruelty ; demonise ; demonised ; tactic ; trick; trickery ; deceit ; dishonest ; deception; devious; deviouness; 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 ; embarrasment ; embarrass; embarrasing ; 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; unacceptbale ; condemn; short-sighted; ashamed; falsehood; blackmail; clownery; debased; debase; hypocracy; 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	frankly	destroyers	happy	hear	relieving
undoubtedly	always	approvals	excitement	appreciated	satisfactorily
frankly	think	festival	enjoying	amazed	satisfy
always	nevertheless	dwellings	cheer	wonderful	relief
think	often	engagements	celebration	sadness	gay
nevertheless	genuinely	championships	delighted	love	grind
often	believe	championship	relieved	doubtless	satisfies
genuinely	seem	approving	celebrate	birthday	satisfactory
believe	felt	shakespeare	amused	horrified	entertainment
certainly	however	challenger	anticipation	always	grinding
seem	indeed	treasurer	fun	praise	amusement
felt	feeling	pesticides	entertaining	informative	laughed
however	perhaps	approved	pity	fascinating	laughs
indeed	obviously	risk-based	enjoyed	churlish	satisfaction
feeling	something	harmonise	liked	pleased	gladly
perhaps	say	flexibilities	hear	admire	cheers
obviously	probably	energy-intensive	appreciated	christmas	satisfying
worry	find	relaxing	enjoy	look_forward	laughing
something	deeply	laughs	excited	afternoon	entertain
say	nothing	approve	amazed	lovely	laughable
probably	people	festivals	wonderful	fascinated	rejoice
find	thing	exhaustive	sadness	spirit	cheered
deeply	suspect	glamorgan	love	compliment	enthusiastically
nothing	somehow	approves	celebrating	astonished	enjoyment
people	quite	resignations	doubtless	coincidence	celebrates
thing	much	praises	glad	sincerely	joke

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	felt	disappointingly	frankly	irresponsible	steal
unhappy	feeling	terrorists	scandalous	pathetic	assault
worry	caused	avoidance	dishonest	dreadful	offend
deeply	horrendous	cowardly	embarrassing	bizarre	furious
dreadful	appalling	grievance	absurd	complacency	fail
unfortunate	shocked	hopelessly	ridiculous	illogical	deceived
worried	frustrating	lone	ludicrous	incompetent	predictable
suffer	compounded	miserably	deplorable	shocking	dodgy
anxiety	anger	terrorism	incompetence	reckless	fool
despair	frustration	alarmingly	misguided	disingenuous	problematic
fear	sometimes	grievances	irresponsible	complacent	bitterness
frightened	horrible	alarmist	pathetic	unfortunate	fiasco
angry	experiencing	painted	appalling	downright	neglected
suffered	feel	timid	dreadful	deliberate	betray
sad	frustrated	terrorist	nonsense	wicked	cruelty
felt	appalled	discouraged	bizarre	unjust	confrontational
feeling	shocking	shy	complacency	deplore	deceive
tragic	understandably	discouraging	ashamed	unacceptable	archaic
caused	disturbed	avoids	illogical	plainly	blackmail
horror	unpleasant	lamentable	arrogant	horrible	embarrass
horrendous	embarrassing	pitiful	incompetent	manifestly	mischief
appalling	terribly	discourage	arrogance	callous	smokescreen
shocked	frankly	sufferers	accusation	somehow	needlessly
frustrating	imagine	painfully	shocking	muddle	adversarial

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	whose	also	recollect
whereas	weeks	scarce	constituents	just	alice
years	moreover	bunch	father	others	aunt
seven	past	seven-day	mr	asked	eve
quarter	compared	groupings	son	saying	skye
two	almost	fifthly	tell	week	albert
eight	yet	samples	went	see	chat
£	now	grouped	met	wanted	spencer
million	next	ooo-page	remember	perhaps	kate
months	spend	equalities	think	former	mohammed
billion	ago	equalise	indeed	described	rhetoric
just	thirds	ooog	says	obviously	ashton
average	figure	group's	dr	one	tale
least	addition	ooond	wife	ooo-year-old	roman
days	roughly	sixth-form	david	mine	inspire
weeks	week	ooob	say	knows	youngster
moreover	furthermore	equalisation	also	yesterday	nicola
past	number	ooo-to-ooo	just	unfortunately	locality
compared	within	six-week	family	friends	everyday
third	times	triple	others	looked	sensitive
almost	april	four-year-old	woman	aware	jamie
one	probably	grouping	asked	although	carter
yet	equivalent	ooord	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 begin this afternoon by wishing the Secretary of State a very happy birthday - I sincerely hope that it improves from here on.	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 join hon Members across the House in wishing a happy Pride to all those celebrating London Pride this weekend.	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.
Is it any wonder that mentally ill people desperate for help just get lost, sometimes with tragic consequences?	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.
All of us are aware that the Labour party has trouble understanding aspiration and even more trouble in rewarding aspiration.	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.
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.	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.
But neither can anyone underestimate the anger and sadness among people that things should ever have been allowed to get into this position.	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.	I have primary schools receiving less than £3,500 per pupil and secondary schools receiving less than £4,600 per pupil.	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".	The maximum figure for those costs was $91 billion, although the real extra costs amounted to $26 billion.	Such serious problems have left many facing uncertainty, which can cause severe stress to people who already face incredibly challenging circumstances.
They should not be all about blaming people, because blaming individuals for errors and mistakes is unhelpful and counter-productive.	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.
Of course the situation in Zimbabwe is disgraceful and we condemn utterly the barbaric attacks on farmers, which are totally unacceptable.	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.	Working in early years or later years care in private services means earning minimum wage or minimum wage plus.	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?	Approximately 100 people per 1,000 currently receive disability living allowance, compared with 50 people per 1,000 in Britain.	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.

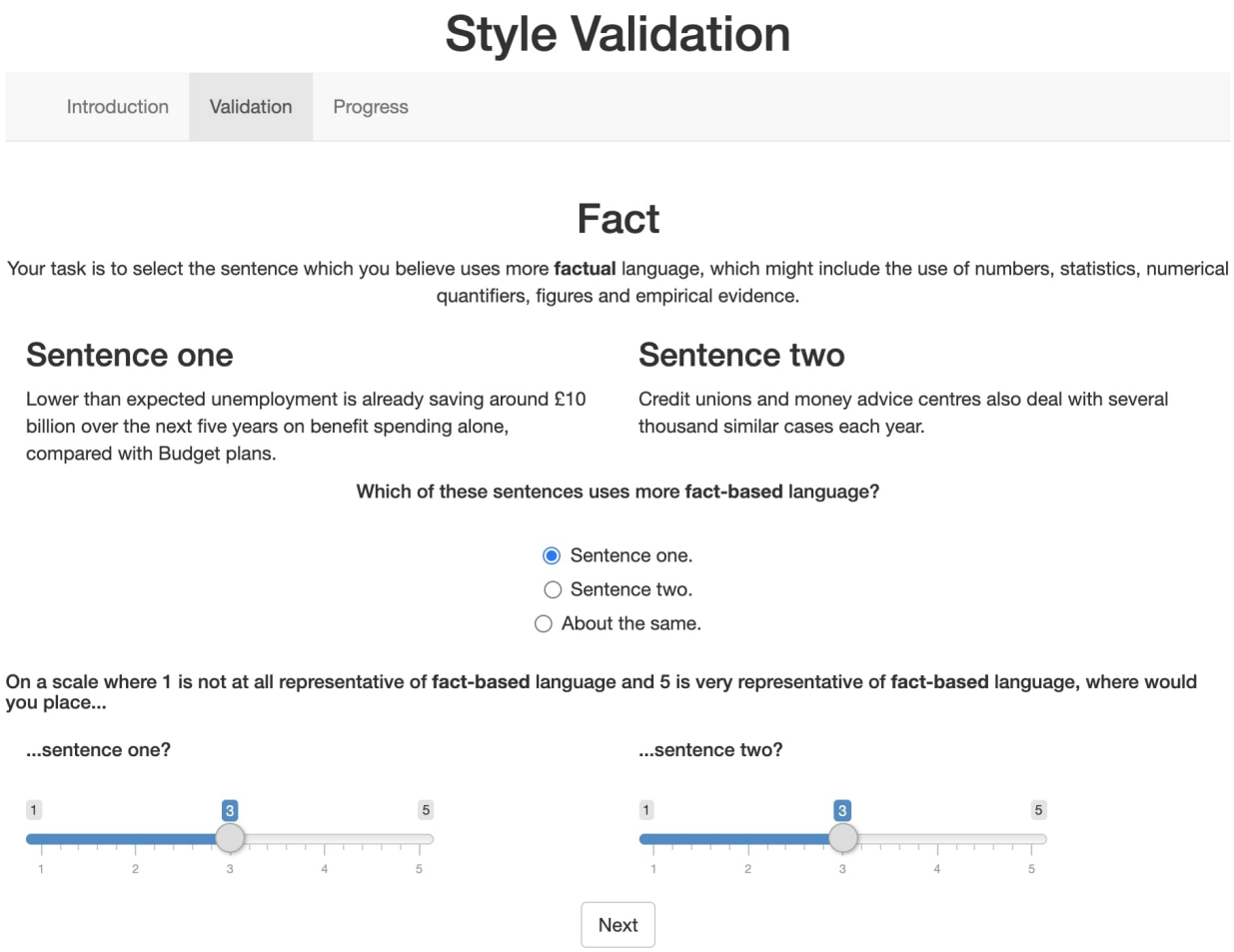

Figure S3: Human validation task prompt

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)	0.72 (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 determine this difference by using a bootstrap procedure, 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

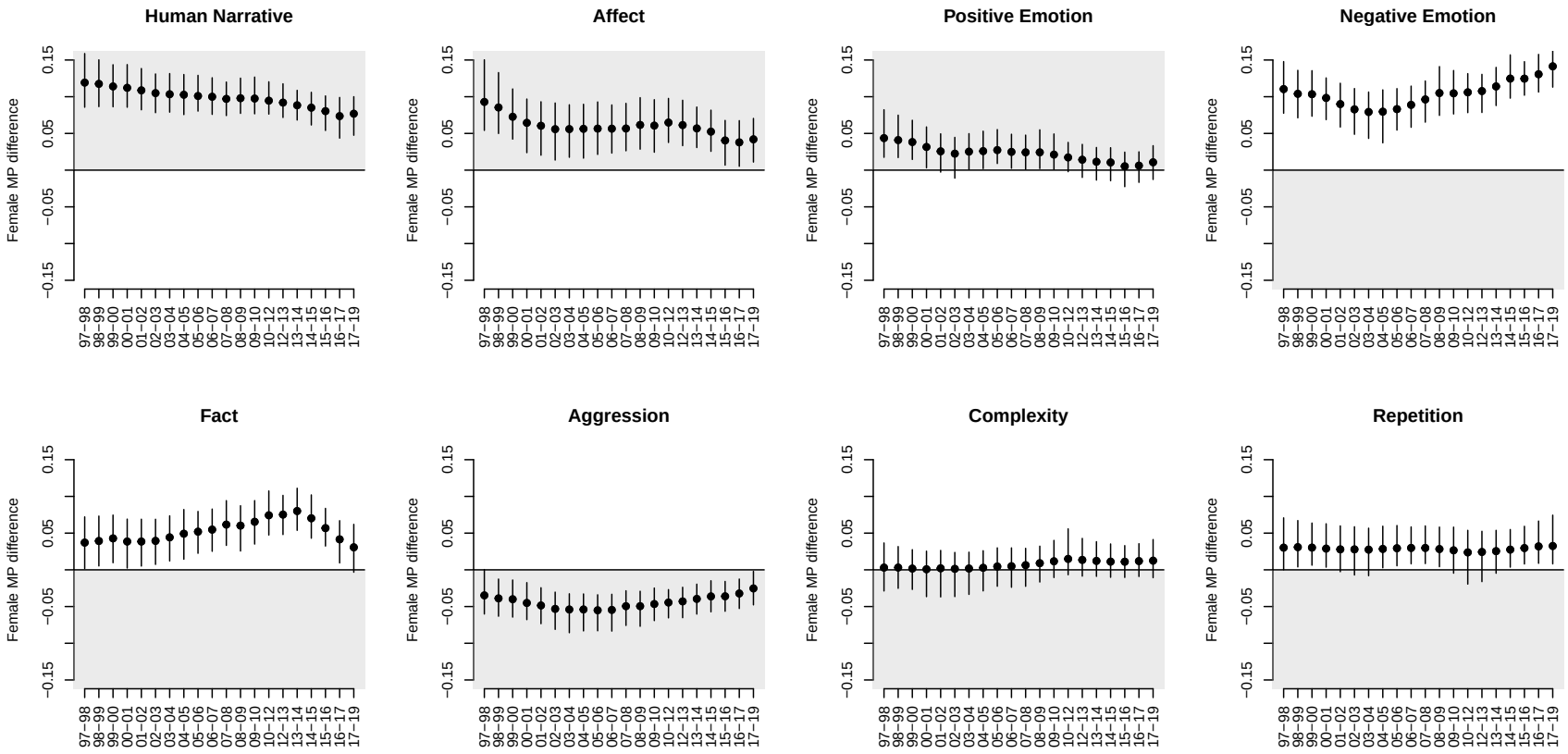

Figure S4: Gender differences in style over time controlling for individual-level confounders

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.

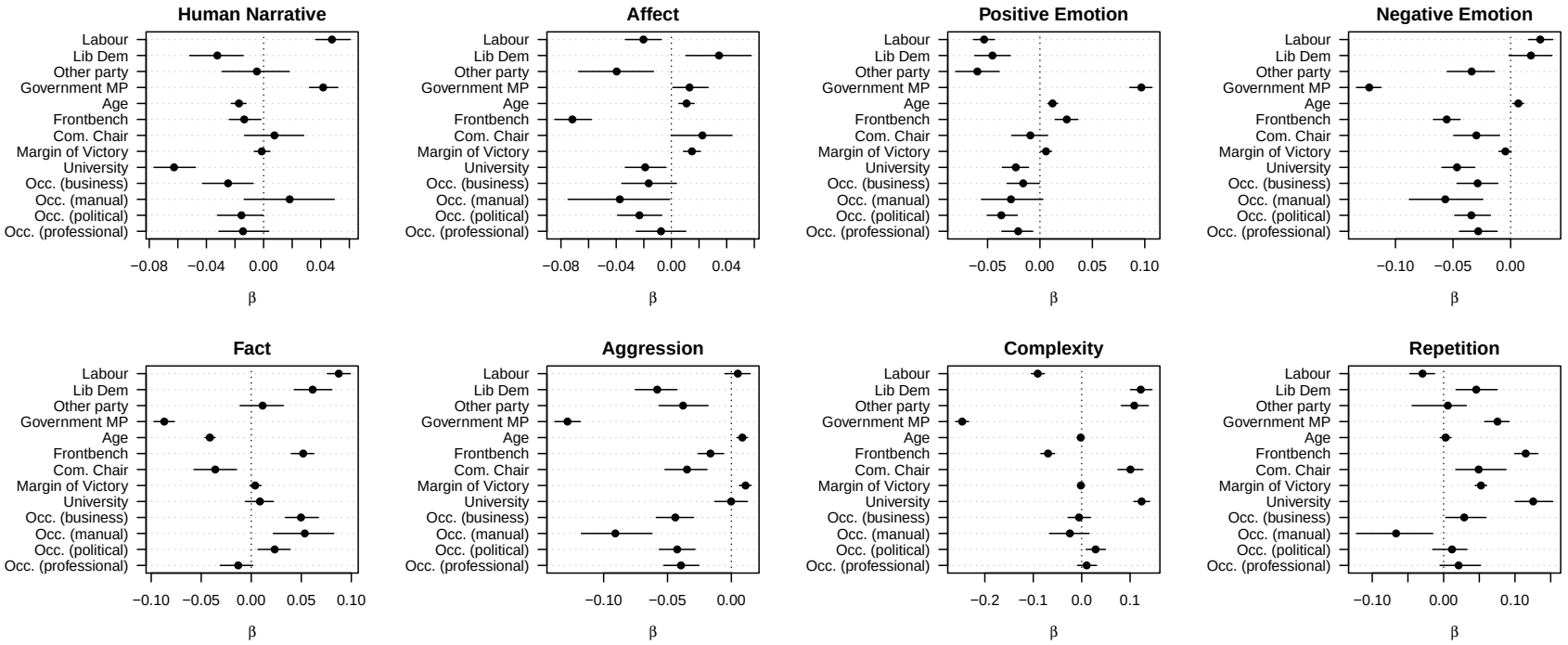

Figure S5: Individual-level covariate effects

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, 2021):

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.

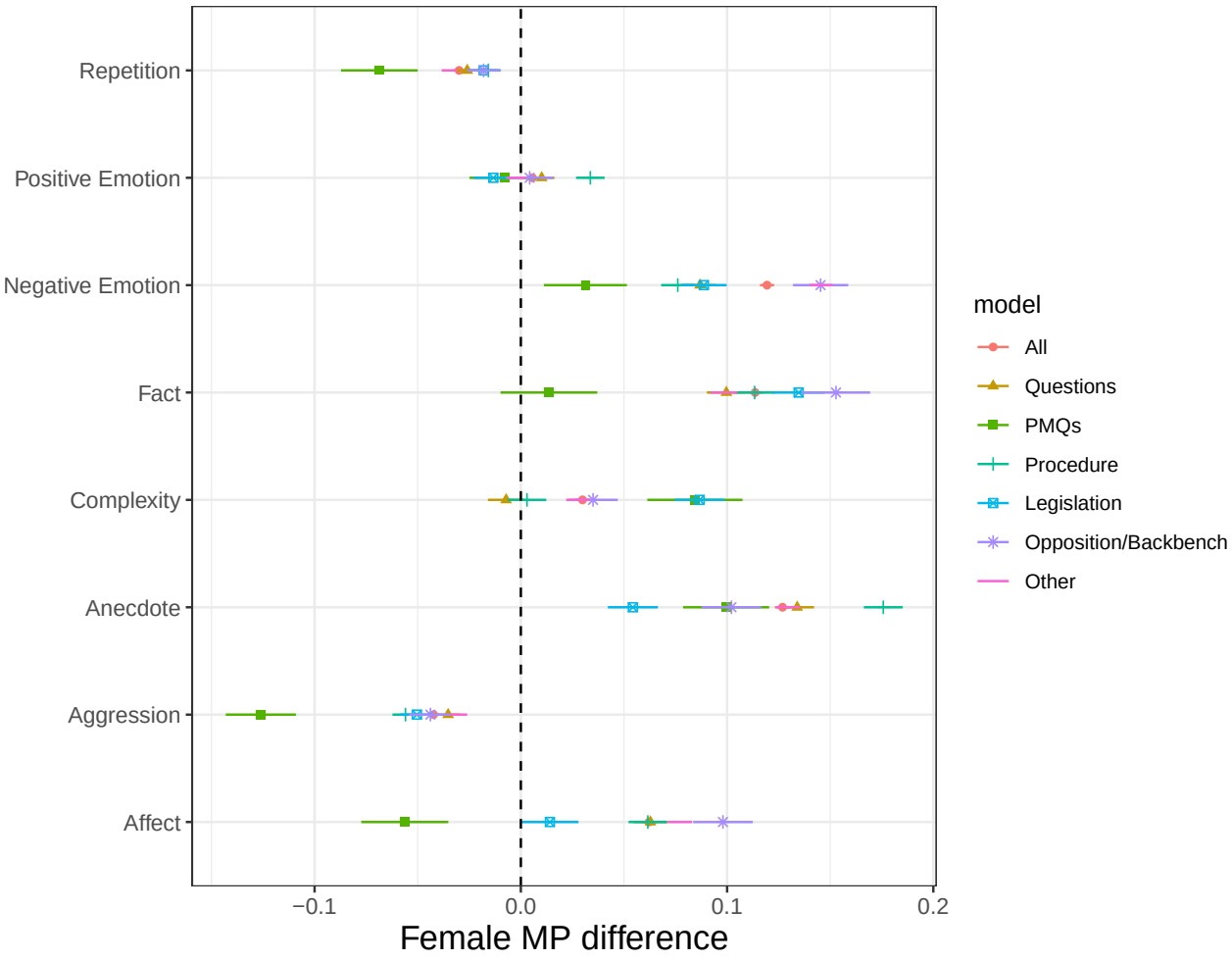

Figure S6: Debate type models

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.

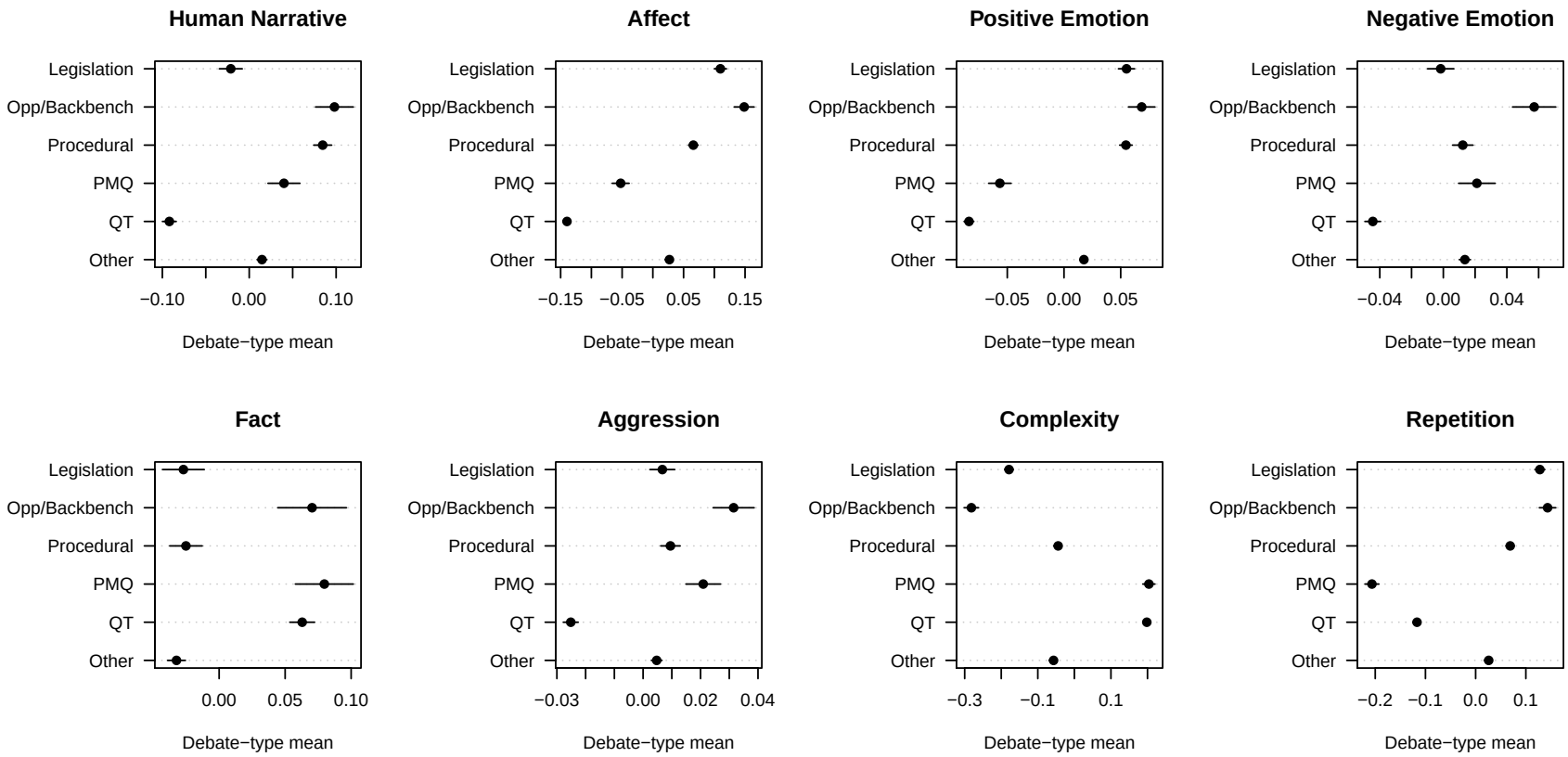

Figure S7: Style type average by debate type

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-

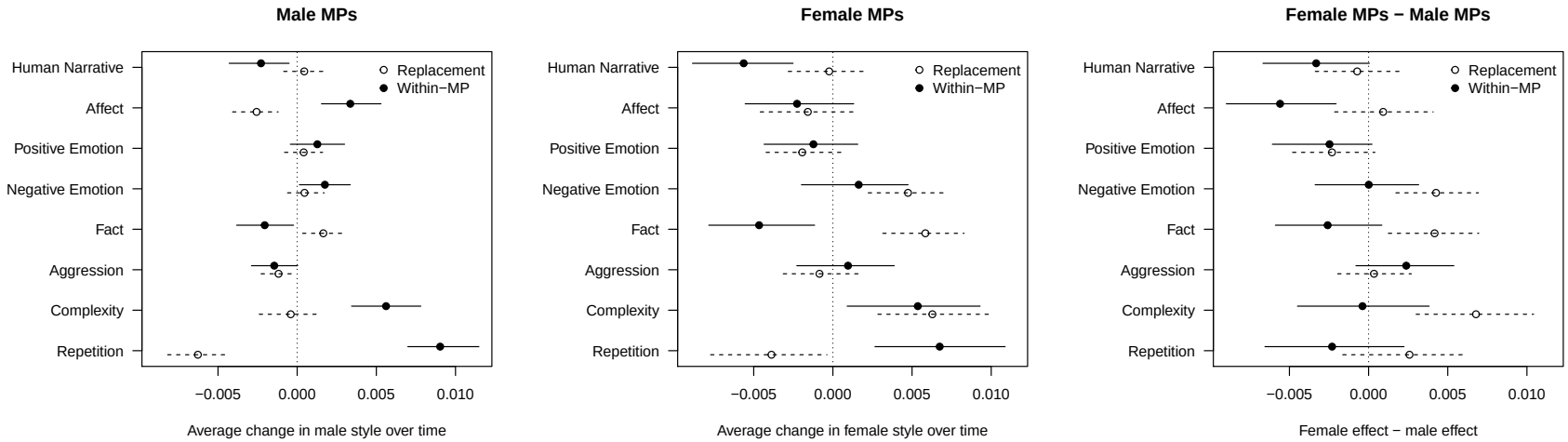

Figure S8: Within-MP and replacement over time change by gender

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

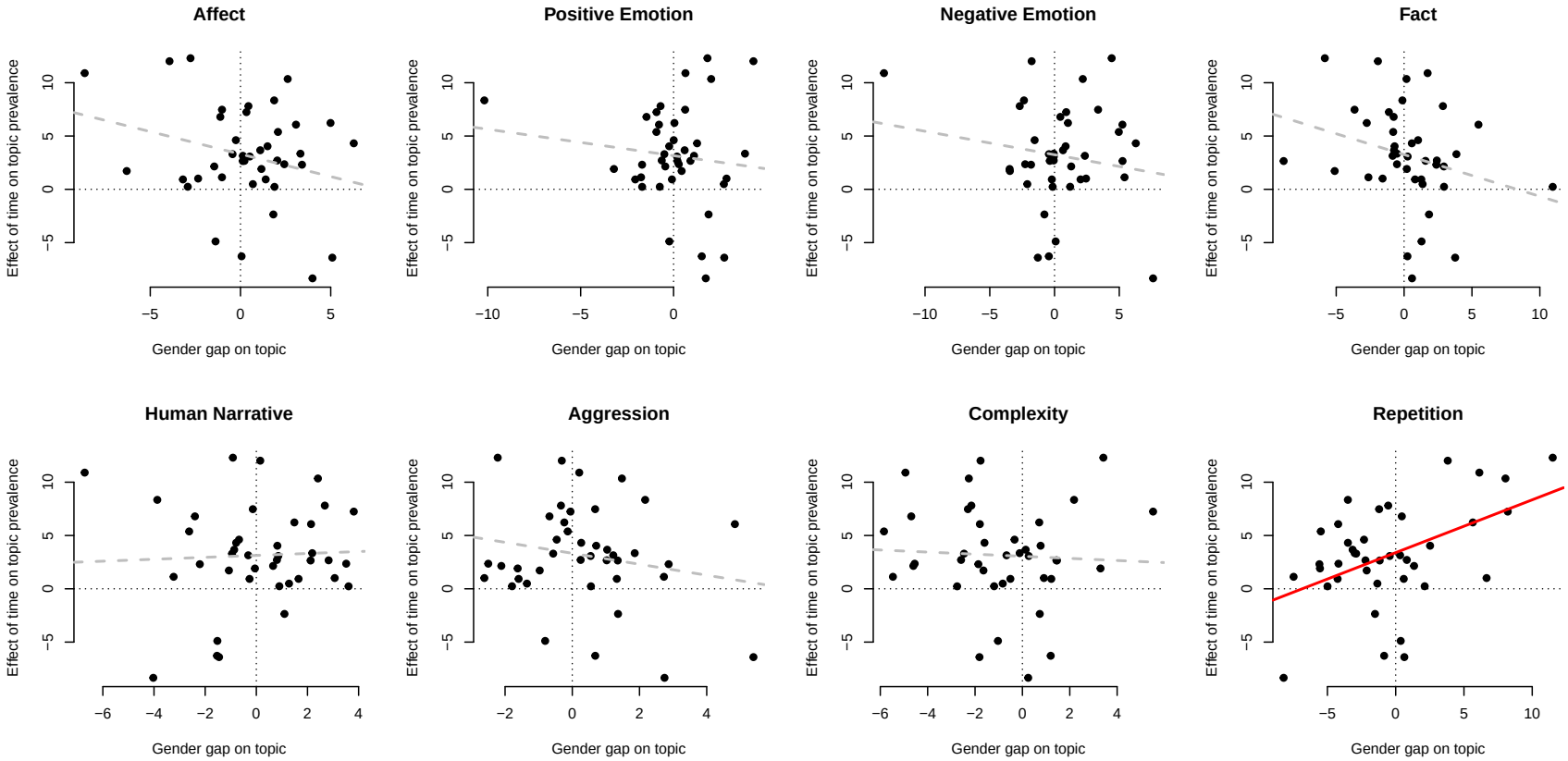

Figure S9: **Topical-confounding**: The figure shows the relationship between the gender gap in the use of a given style on a given topic (x-axis), and the change in the prevalence of a given topic over time (y-axis).

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.

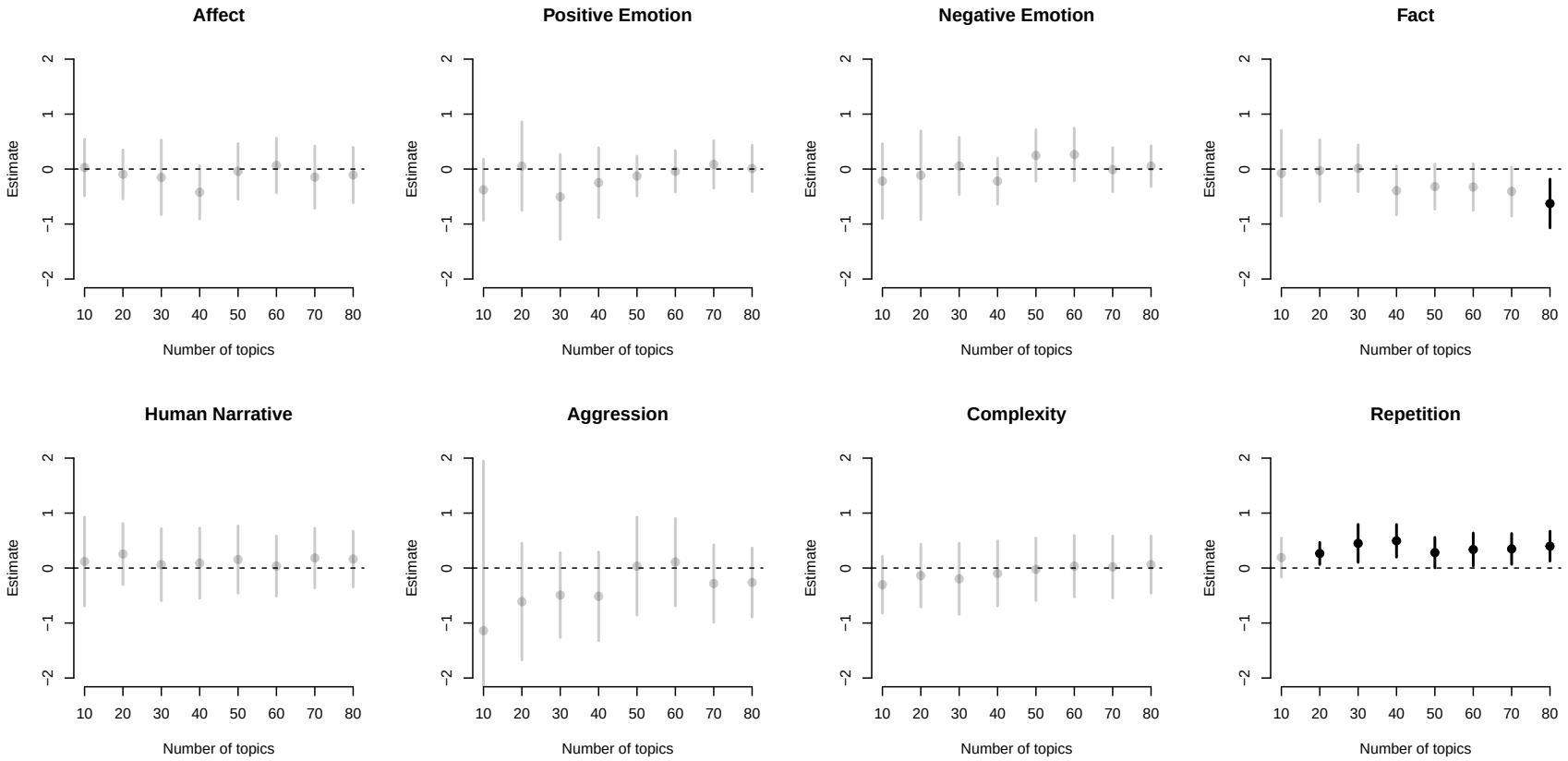

Figure S10: **Topical-confounding, varying** K: On the y-axis, the figure summarises the linear relationship between the gender gap in the use of a given style on a given topic (δ_k^s), and the change in the prevalence of a given topic over time (γ_k). The x-axis measures the number of CTM topics, K,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

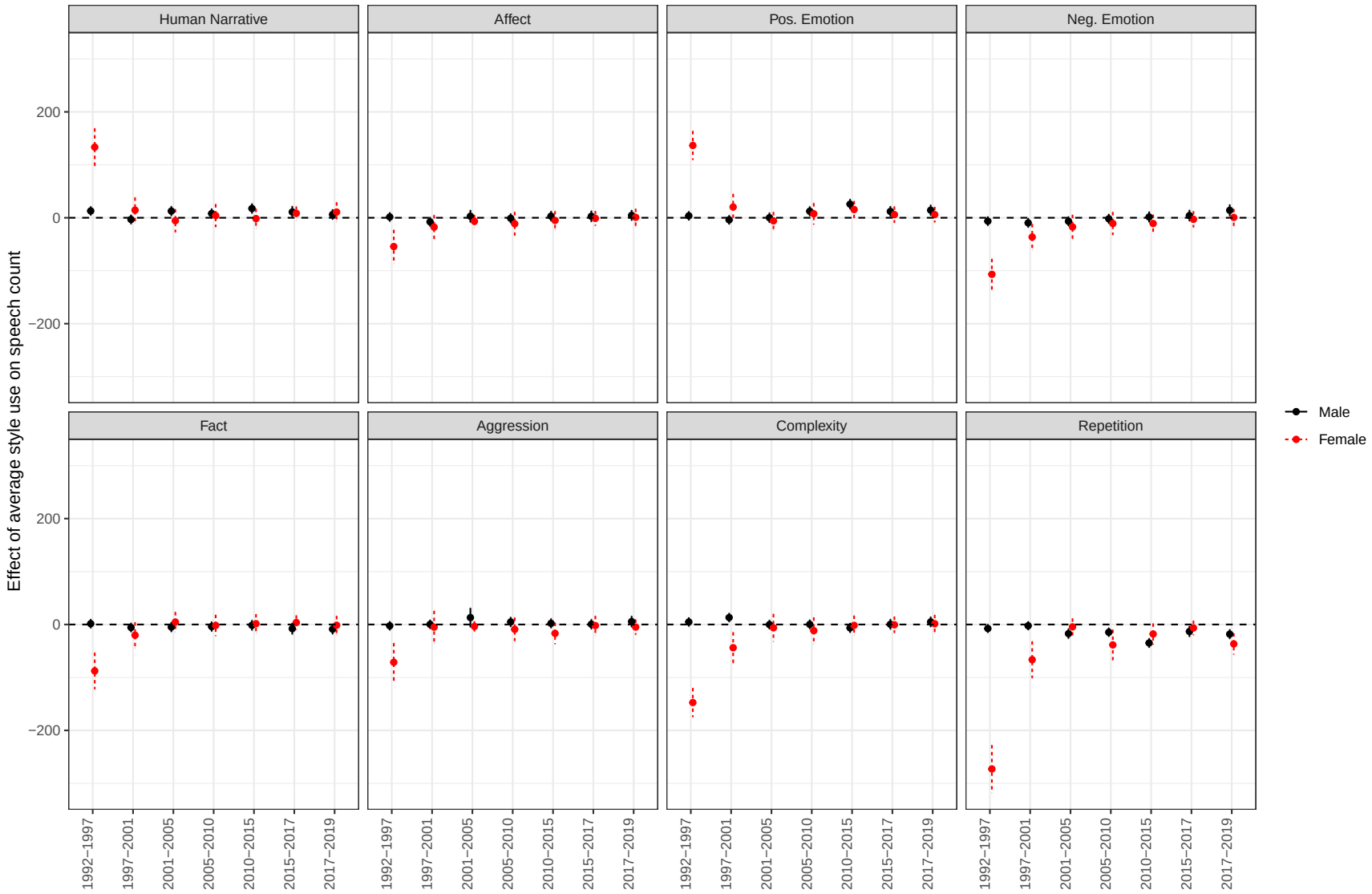

Figure S11: **Participation as a function of average style use, by parliamentary term**: The figure illustrates the average marginal effect of a one standard deviation increase in the average style use on the number of times an MP speaks in a given parliamentary term.

References

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

Blumenau, Jack and Roberta Damiani. 2021. Legislative Debates in the British House of Commons. In *The Politics of Legislative Debates*, ed. Hanna Bäck, Marc Debus and Jorge M. Fernandes. Oxford: Oxford University Press.

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.