Gender, Candidate Emotional Expression, and Voter Reactions During Televised Debates
Online Appendix

Constantine Boussalis Travis G. Coan Mirya R. Holman Stefan Müller

American Political Science Review

Research documentation and data that support the findings of this study are openly available at the American Political Science Review Dataverse:
https://doi.org/10.7910/DVN/NVVVUV

Author Information

Constantine Boussalis (https://orcid.org/0000-0002-0609-6272), Assistant Professor, Department of Political Science, Trinity College Dublin, boussalc@tcd.ie.
Travis G. Coan (https://orcid.org/0000-0002-4587-3396), Senior Lecturer, Department of Politics and the Exeter Q-Step Centre, University of Exeter, t.coan@exeter.ac.uk.
Mirya R. Holman (https://orcid.org/0000-0001-6648-4122), Associate Professor, Department of Political Science, Tulane University, mholman@tulane.edu.
Stefan Müller (https://orcid.org/0000-0002-6315-4125), Assistant Professor and Ad Astra Fellow, School of Politics and International Relations, University College Dublin, stefan.mueller@ucd.ie.

Contents

Appendix A Televised Debates Around the World

Figure A1 shows the 130 countries that normally conduct televised debates candidates or party representatives during legislative or presidential campaigns.

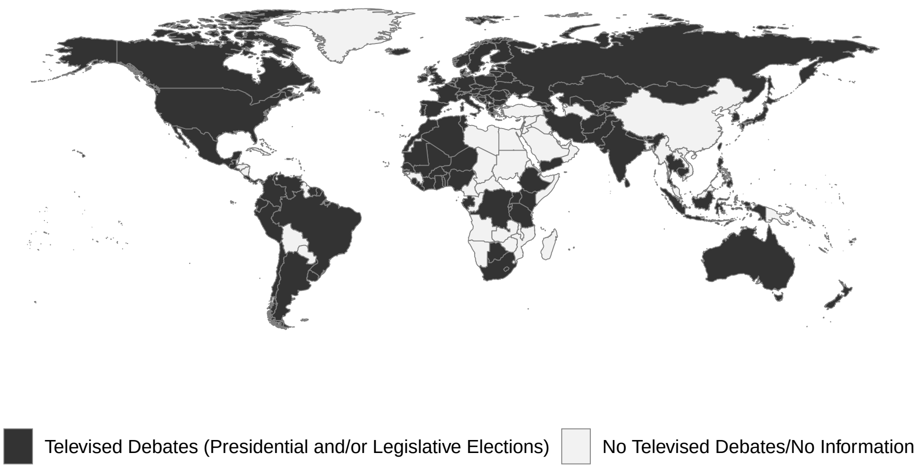

Televised Debates (Presidential and/or Legislative Elections)
No Televised Debates/No Information

Figure A1: The prevalence of televised debates across the world. Visualization based on data provided by ACE Electoral Knowledge Network (2021) and own extensions.

Appendix B The German Debates

B.1 Schröder v Merkel (2005)

Starting in 1998, a coalition between the Social Democratcs (SPD) and the Green Party under Chancellor Gerhard Schröder governed Germany. The "red-green" coalition ended the 18-year tenure of CDU Chancellor Helmut Kohl. In 2002, Schröder was re-elected by a very narrow margin (only 6,000 votes separated the SPD and the CDU/CSU) as the strongest party (Roberts 2006). Under Schröder's tenure, Germany suffered a severe economic crisis.

In May 2005, Chancellor Schröder announced an early election to strengthen his position of further reforms of the economy and labour-market. Schröder lost the artificially engineered vote of no confidence (Roberts 2006 669) resulting in early elections in September 2005. He competed against CDU candidate Angela Merkel, who was not only the first candidate for chancellor from the former GDR, but also the first ever female chancellor candidate. While the media and pundits expected a landslide victory for Merkel, Schröder made a very strong comeback in the weeks before the election and almost levelled with the CDU/CSU as the strongest party on election day.

The TV debate between Schröder and Merkel was the most watched TV debate up to that point. As Roberts (2006 637) summarizes "[c]ommentators in the press and on television thought it was more of an equal outcome, though since expectations of Merkel's rhetorical abilities before the debate had been rather low, the fact that she made no obvious mistakes and managed to score some points against Schröder may have induced an over-estimation of her performance."

The CDU/CSU ended up with 35.1% of list votes (second votes) followed by the SPD with 34.3%. As a coalition including the SDP and the Left Party was ruled out categorically (Proksch and Slapin 2006), the only viable coalition option was a government between the CDU/CSU and SPD—the second ever "grand coalition" in Germany since 1945. Merkel became Germany's first woman chancellor.

B.2 Merkel v Steinmeier (2009)

The "grand coalition" under the leadership of Angela Merkel coalition worked pragmatically and smoothly,[1] but was overshadowed in the last year of the alliance by the global financial crisis. Angela Merkel and Peer Steinbrück (Minister of Finance) received a lot of praise for how the parties handled the challenging economic circumstances. Yet, most voters attributed credit for these developments to Merkel and the CDU/CSU, while the SPD struggled to profit electorally from their crisis management.

In 2009, the SPD selected Frank-Walter Steinmeier, Secretary of State and vice-chancellor. Steinmeier's closeness to Merkel's administration meant he struggled to criticize Merkel and her policies. The televised leaders' debate mirrored this dilemma. As Faas (2010 897) notes: "advertised as a 'duel' by the organising media, with Merkel and Steinmeier as the main contenders, it turned out instead to be quite a harmonious 'duet'."

The 2009 election resulted in the SPD's worst election result of all time. The party obtained only 23 percent of the list votes (–11.2 percentage points), losing 76 seats. The CDU/CSU lost only 1.4 percentage points of list votes. The Liberals (FDP) emerged as the winner of the election, reaching their historically best result with 14.6% of list votes, and subsequently joining a coalition with the CDU/CSU.

B.3 Merkel v Steinbrück (2013)

After the 2009 election, both the FDP and CDU dropped in the public opinion polls. The FDP failed to keep a central electoral promise of tax reductions, while the CDU also made a poor impression with a number of ministers having to resign throughout the term. The SPD presented Peer Streinbrück[2] as their contender at a hastily called press conference in the autumn of 2012, but the party lacked a clear strategy and campaign. Moreover, Steinbrück faced public pressure after journalists revealed that he delivered many private and semi-public talks between 2009 and 2013 with honoraria that summed up to over EUR 1 million. (Faas 2015).

The SPD ran an extensive door-to-door campaign and, for the first time, deployed a comprehensive social media strategy. Yet, these measures did not translate into an increase in public support. The televised debate, however, was regarded as a success for Steinbrück. It increased his popularity and support for the SPD (Faas 2015 242). Despite the promising performance during the TV debate, the election result for the SPD was disappointing. The party gained 2.7 percentage points, but the 25.7% of list votes were nothing close to the 41.5% of the CDU/CSU (+7.8 percentage points).

While the election result was a success for the CDU/CSU, the coalition partner FDP did not pass the five percent threshold. Even though the "left block" of SPD, Greens, and the Left Party

[1] Unemployment fell below 3 million, Germany was moving towards a balanced budget, and social security contributions were lowered.

[2] Steinbrück left politics after the 2009 election, but was endorsed by several former SPD politicians and enjoyed high popularity because he worked very convincingly as the Minister of Finance during the 2005–2009 period which coincided with the height of the global financial crisis.

would have had a majority of seats, the SPD ruled out a coalition with the Left Party. As a result, the only feasible remaining option with a majority of seats was another grand coalition between the CDU/CSU and the SPD with Merkel as chancellor.

B.4 Merkel v Schulz (2017)

In January 2017, SPD party leader Sigmar Gabriel announced that he did not intend to run as the main candidate for the party. The party nominated Martin Schulz, the former President of the European Parliament. Shortly after this announcement, the support for the SPD increased drastically. Many SPD supporters and experts believed Schulz had a realistic chance of becoming chancellor (Faas and Klingelhöfer 2019) resulting in a sheer "Schulz hype." However, after his nomination, the party lost several important subnational State elections and the honeymoon period ended abruptly.

Merkel and the CDU tried deliberately to reduce political conflict before the election. The SPD claimed that the party delivered on a lot of their central promises during the "grand coalitions". Yet, it was mainly Merkel and the CDU who received credit for these policy changes. The TV debate between Merkel and Schulz mirrored this confrontational style. "During the TV debate with Merkel, Schulz vigorously attacked and tried to undermine Merkel's credibility" (Faas and Klingelhöfer 2019 918).

Both the CDU/CSU and the SPD suffered from massive electoral losses in 2017, with the lowest combined vote share in the history of the state. Moreover, the right-wing populist party AfD gained representation in the Bundestag for the first time. Having failed to pass the 5% threshold of list votes by a small margin in 2013, the AfD obtained 9.6% of list votes.

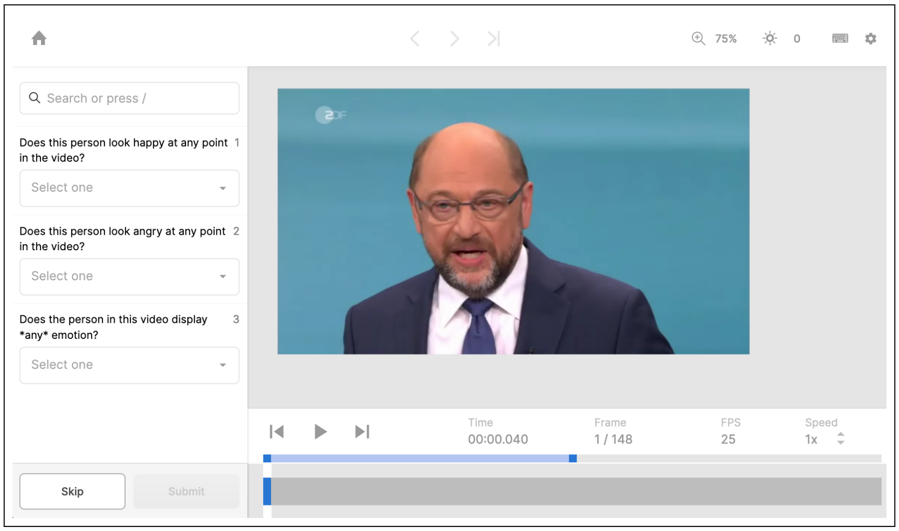
Appendix C Validating Displays of Emotion and Sentiment

C.1 Comparing the Automated Detection of Emotions with Human Coding of Emotional Displays

While both vocal pitch and sentiment have been extensively validated elsewhere ((Dietrich, Hayes and O'Brien 2019, Rauh 2018, Proksch et al. 2019), respectively), few studies validate automatic detection of facial displays and no studies to our knowledge do so in the context of German televised leaders' debates. We begin by examining the validity of the Face API predictions by comparing them to human annotations. We draw on two different samples of coders to assess the API predictions, each with their own strengths and weaknesses: 1) a large sample of roughly 5-second video clips of the debate annotated by two trained coders and 2) a small sample ($N = 50$) of clips each annotated by nearly 500 crowd-workers.

C.1.1 Validation Using Trained Annotators

We started by collecting a large sample of human coded clips across the four debates ($N = 1,341$). To generate the validation set, we recruited two research assistants (both women) to code a random sample of roughly 5-second clips for whether the candidate in the clip "displays any emotion", looks "angry at any point", or looks "happy at any point" (see Figure A2 for an illustration of the annotation tool). Following Boussalis and Coan (2021), the coders were asked to rate the level of



Does this person look happy at any point in the video? 1

Does this person look angry at any point in the video? 2

Does the person in this video display *any* emotion? 3

Time 00:00.040 Frame 1 / 148 FPS 25 Speed 1x

Figure A2: Example of the annotation tool. The coding of debate clips was carried out using software from Labelbox (see `https://labelbox.com`).

emotion expressed on a five point scale ranging from "not at all" to "extremely". After completing a training session, both annotators coded a sample of 75 clips that had been evaluated by the PIs as representing the full set of possible emotions. The inter-coder reliability between the annotators (Krippendorff's $\alpha = 0.81$) indicated reliability across our coders. The coders then annotated an additional $1,134$ clips which are used to establish the correspondence between the model and human judgements.

Given that we are comparing a continuous model prediction (z-score of the average confidence score for the relevant emotion) to a 5-point numerical scale of emotional expression, we first examine the association between the predictions and human annotations by assessing the root mean squared error (RMSE), where zero error is represented by an RMSE value of zero. To assess out-of-sample performance, we employ five-fold, repeated cross-validation. Turing first to the estimates for "all" candidates (i.e., Merkel and her male opponents), we find an RMSE of 0.83 for the expression of "any emotion," suggesting that predictions based on the model are within less than a point on the scale of 1-5. Consistent with Boussalis and Coan (2021), we find that the model does a better job at predicting happiness ($RMSE = 0.82$) than it does for anger ($RMSE = 1.03$).

Next, we examine whether there are systematic differences—or biases—in model performance when comparing Merkel versus her opponents, finding very similar levels of performance for each emotion. When considering the expression of "any emotion" and "happiness", we find only slight differences in the estimated RMSE for Merkel (roughly 0.85 in both cases) and her male opponents (roughly 0.80 in both cases). The model continues to perform better for the male candidates than Merkel when considering anger, but the differences are more pronounced: we find an estimated RMSE of 1.08 for Merkel and 0.97 for her opponents.

In addition to examining performance via the RMSE, we examine *classification* performance by transforming the Likert scale measure of emotions into a binary measure (Boussalis and Coan 2021). We recode each emotion measure (anger, happiness, and any emotion) to equal 1 for clips coded as "very much" or "extremely" and 0 otherwise. We fit a logistic regression classifier and examine held-out model performance via 5-fold repeated cross-validation. The results are generally consistent with the RMSE: the F1 score for happiness 0.96 (precision = 0.94, recall = 0.97) exceeds the score for anger 0.77 (precision = 0.64, recall = 0.96) and any emotion 0.58 (precision = 0.57, recall = 0.58). When stratifying the sample across Merkel and her opponents, the classification

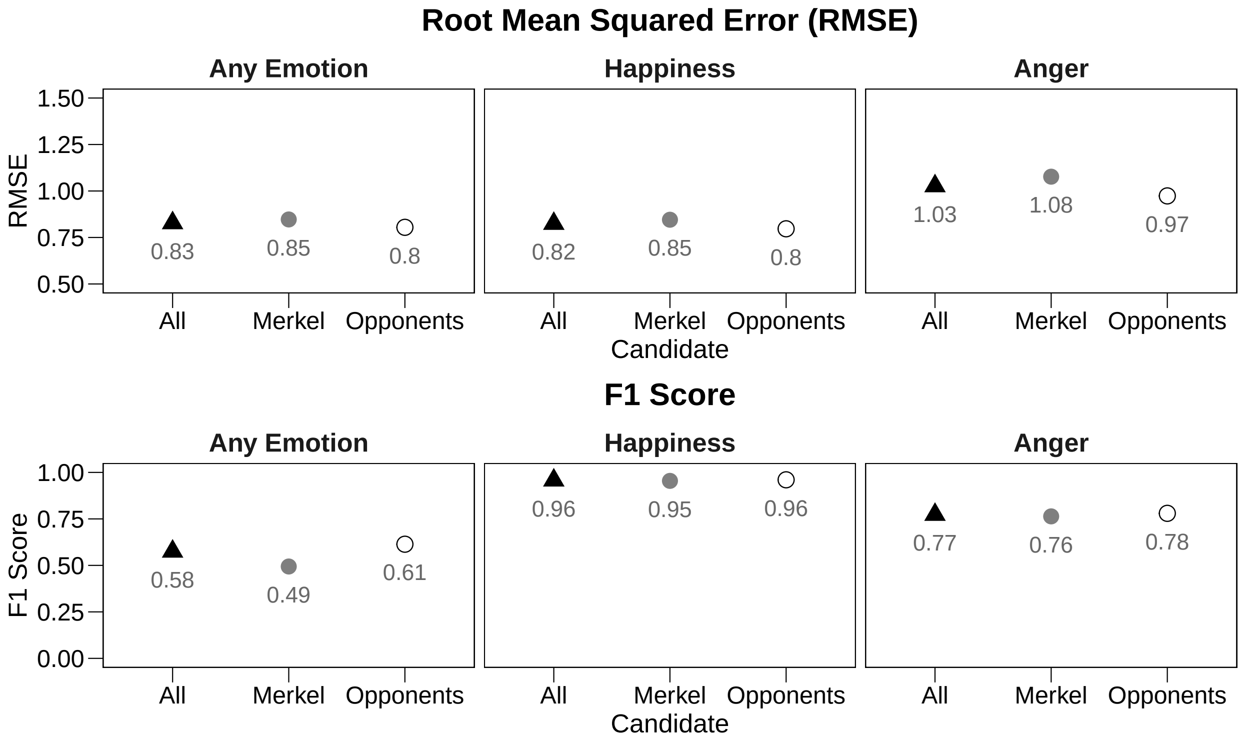


Figure A3: Root Mean Square Error (RMSE) and F1 scores for codings by trained annotators. A lower RMSE (range from 0 to 4) and a higher F1 scores (ranging from 0–1) imply more congruence between human coding of emotions and Face API predictions.

results confirm the main findings from the RMSE analysis: the classification performance is slightly better for the male candidates, and these differences are more pronounced for anger.

C.1.2 Validation using Crowd-sourced Annotations

While our trained annotators sample provides a useful benchmark to assess the performance of the Face API in the context of German debates, the sample is because two coders were both women and we cannot examine whether the *annotator's gender* is important when assessing model performance. To examine the role of annotator gender in model performance and further explore the potential for gender biases, we draw on a sample of 467 respondents using the the crowd-sourcing platform Lucid (https://luc.id). The sample has slightly more women (%54) than men (%46). After an attention check, each respondent annotated 50 clips for whether the candidate "displays any emotion", looks "angry at any point", or looks "happy at any point" using the same answer format as the expert annotators. This exercise resulted in over 70,000 codings (50 clips × 3 emotions questions × 467 respondents).

Figure A4 examines the average correspondence between the crowd annotations and the model-based estimates across candidate gender ("all candidates", "Merkel", and "opponents") and the gender of the coders ("all coders", "Female" coders, and "Male" coders) for each emotion ("any emotion", "happiness", and "anger"). To ensure comparable out-of-sample performance estimates across the various respondents, we followed the following four-step cross-validation procedure. (1): we carry out 5-fold cross-validation at the *clip level*, resulting in 5 training and testing sets with 40 and 10 clips, respectively. (2): we fit a linear model using the training clips for each fold and then estimate the prediction errors for each respondent using their annotations for the corresponding test set in each fold. Note that we pool observations from respondents with the same gender when fitting the model and thus estimate two separate models, one for males and another for female annotators. (3): we calculate the root mean squared error (RMSE) for each annotator and average over the 5 folds to estimate performance per coder. (4): we average the per coder performance estimates within relevant groups (i.e., candidate gender and respondent gender) for each emotion of interest (any emotion, anger, or happiness).

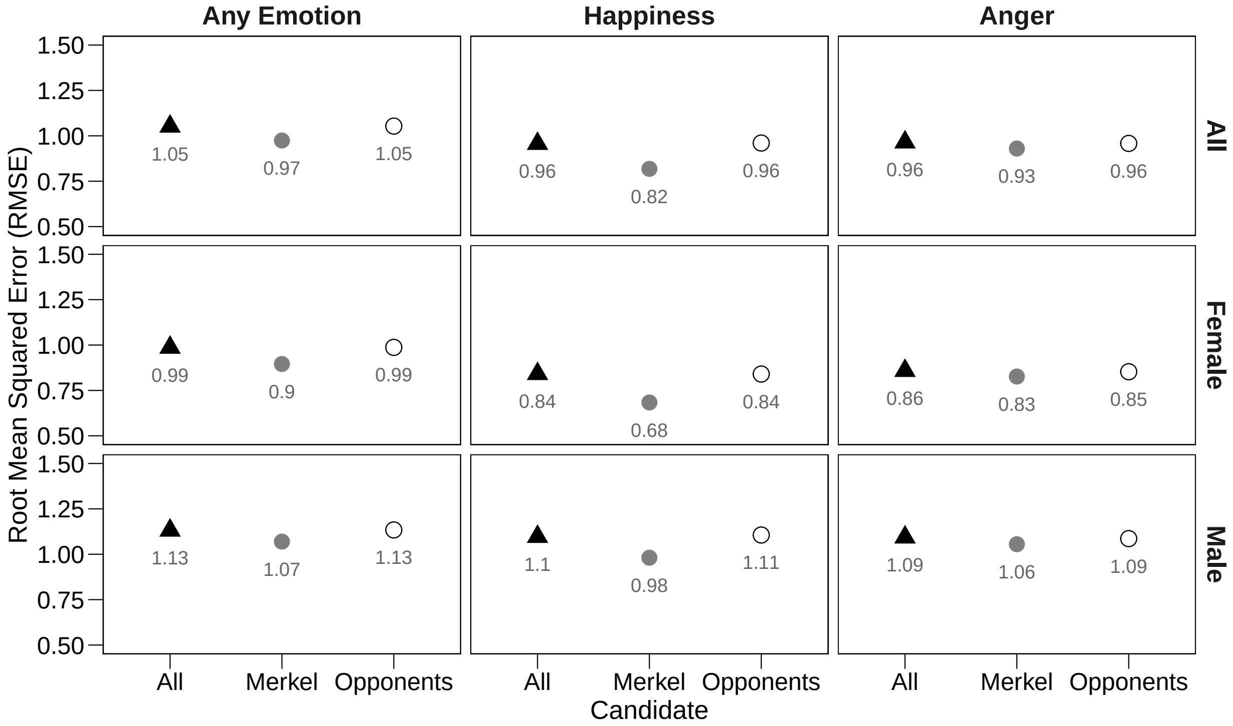


Figure A4: Root Mean Square Error (RMSE) for crowd codings of emotional displays. Lower values imply more congruence between human coding of emotions and Face API predictions.

Overall, the RSME scores indicate relatively good performance of the computer vision model, as compared to the crowd-sourced coders. We continue to find a consistent pattern regarding performance across emotion: the model provides the most accurate predictions for happiness, while providing comparatively poorer predictions for anger and any other emotion. When considering the *candidate* gender, we continue to find slight differences between Merkel and her male opponents. However, in the crowd-sourced sample, we generally find a closer correspondence with the model predictions for Merkel, rather than her opponents, and these differences are most pronounced for female respondents and happiness. Finally, across all emotions and candidates, there is a closer correspondence between the annotations of female coders and the model predictions.

While the descriptive measures provided in Figure A4 are suggestive, the annotator-level observations of performance allow us to specifically test for gender biases in model performance. Figure A5 plots the results of a linear regression in which the gender of the coder (female = 1, male = 0), the gender of the candidate (Merkel = 1, male opponents = 0), and an interaction between these two measures are regressed on annotator-level RMSE estimates. The results confirm the descriptive analysis. First, there is strong evidence for a closer correspondence between the model estimates and female annotations, and these results hold across all relevant emotions categories. Second, when considering "any emotion" or "anger," we find that observed differences in performance across Merkel and her male opponents are insignificant at traditional levels, while also find little support for the interactive influence of coder gender (female) and Merkel clips. However, the results differ when considering happiness, where the model performance for happiness is better for both male and female coders.

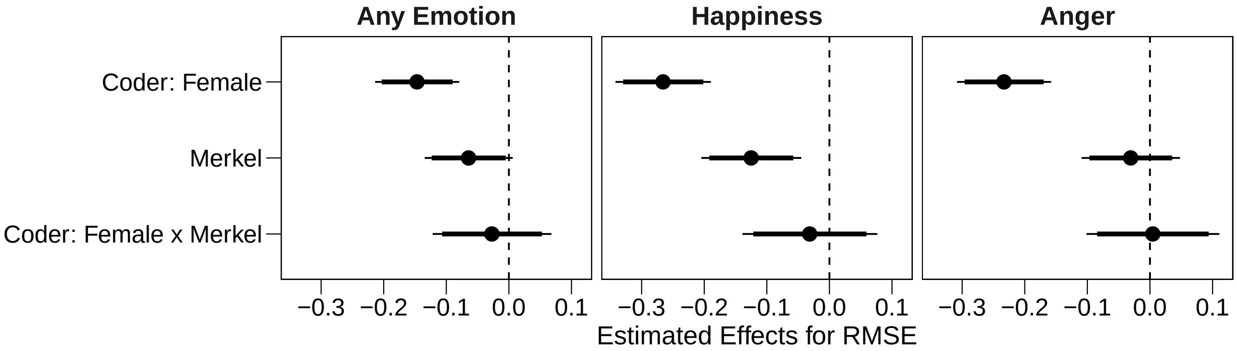


Figure A5: Regression estimates comparing model performance across groups and sub-groups.

C.1.3 Validating Happiness Using Smiles

Happiness is often expressed with a smile (Hess et al. 2009). We use this to validate our happiness measure. Coding every second of the debate, coders trained by Nagel, Maurer and Reinemann (2012) assessed the presence of smiles, distinguishing between no smile, light smile, and strong smile. We align the automated detection of happiness with the human-coded measure. We would expect that the automated measure has the highest values in seconds that the human coders labeled as containing a strong smile, followed by light smiles. Figure A6 corresponds to our expectation. The boxplots show the distribution of the standardized happiness values for each second for Schröder and Merkel for each of the 'smile' categories. The average happiness values for seconds labelled as 'strong smiles' amounts to 2.8 for Merkel and 3.18 for Schröder. In seconds coded as 'light smile' we still observe positive values of happiness, but the mean is considerable lower. For seconds coded as 'no smile' the averages are lowest.

C.2 Validating Speech Sentiment

C.2.1 Alternative Sentiment Dictionaries

We use the sentiment dictionary developed by Rauh (2018) and the German version of the Lexicoder Sentiment dictionary (Proksch et al. 2019) to measure speaker sentiment. Both dictionaries have been validated extensively for the analysis of German political text and political speech (Rauh 2018, Proksch et al. 2019).

Figure A7 plots the correlation between the sentiment scores based on Rauh's sentiment dictionary and the German version of the Lexicoder Sentiment Dictionary (Rauh 2018, Proksch et al. 2019). The correlation does not differ based on the gender of the speakers. For male and female politicians, the correlation amounts to 0.5. The correspondence between both dictionaries does not depend on the gender of a speaker. The similarity between the two measures is promising because the number of words in each dictionary differs substantively. The translated Lexicoder Sentiment Dictionary contains 3,998 terms labeled as 'positive' and 5,849 terms labeled as 'negative', whereas Rauh's Sentiment dictionary contains 17,330 'positive' and 19,750 'negative' words. In addition,

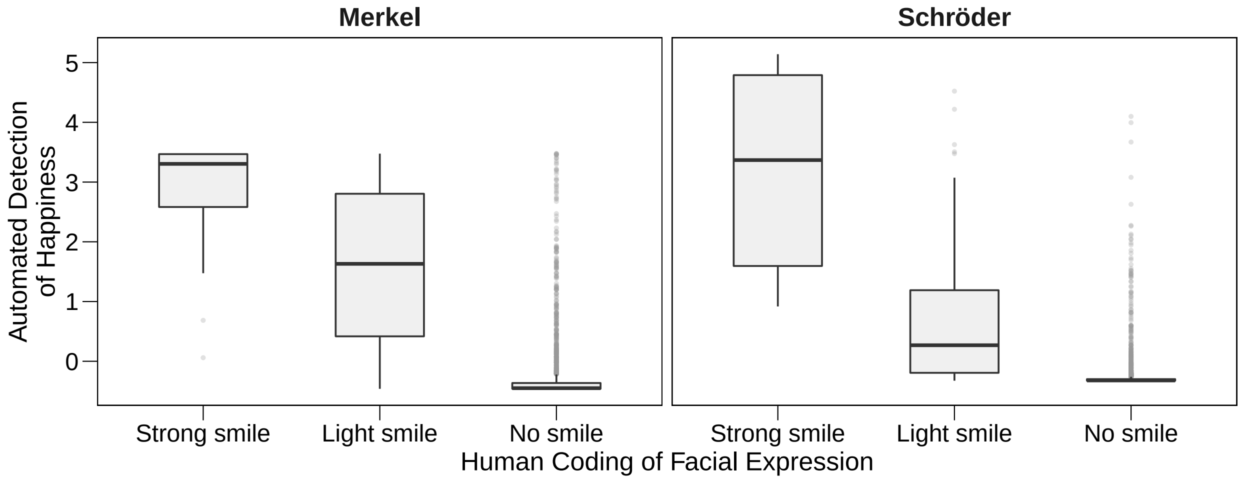


Figure A6: Comparing the coding of smiling in the 2005 debate, with the automated detection of emotional displays (y-axis), standardized by speaker. Smiles from manual coding by Nagel, Maurer and Reinemann (2012).

most statements are rather short, making it a difficult exercise for classifying sentiment using a simple bag-of-words approach.

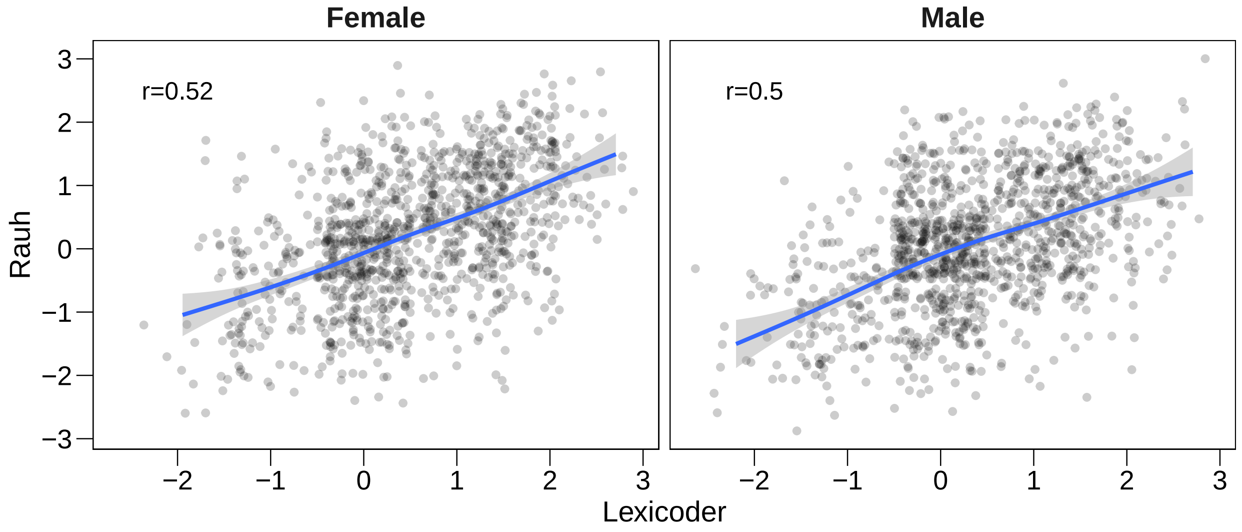


Figure A7: Comparing the correlation of statement-level sentiment using the dictionary by Rauh (2018) and Proksch et al. (2019).

We also assess the face validity of the sentiment scores. For the debates in 2009, 2013, and 2017, human coders assessed whether a statement described the societal situation in a positive or negative way, or if a statement was neutral or did not describe the social situation. We would expect that politicians' statements that describe the current situation positively tend to have higher sentiment scores than statements describing the social situation in a neutral or negative way. Figure A8 shows the averages and confidence intervals of the sentiment scores for each of the three classes, separately for both dictionaries. We observe that sentiment is indeed substantively more positive in utterances that have been coded as a positive description of the social situation. The 2017 debate between minor parties is an exception, which further strengthens the validity of our measure. The correspondence is lower given that only candidates from opposition parties participated in this debate. The politicians usually do not describe the social situation in positive terms and tend to express more negative sentiment.

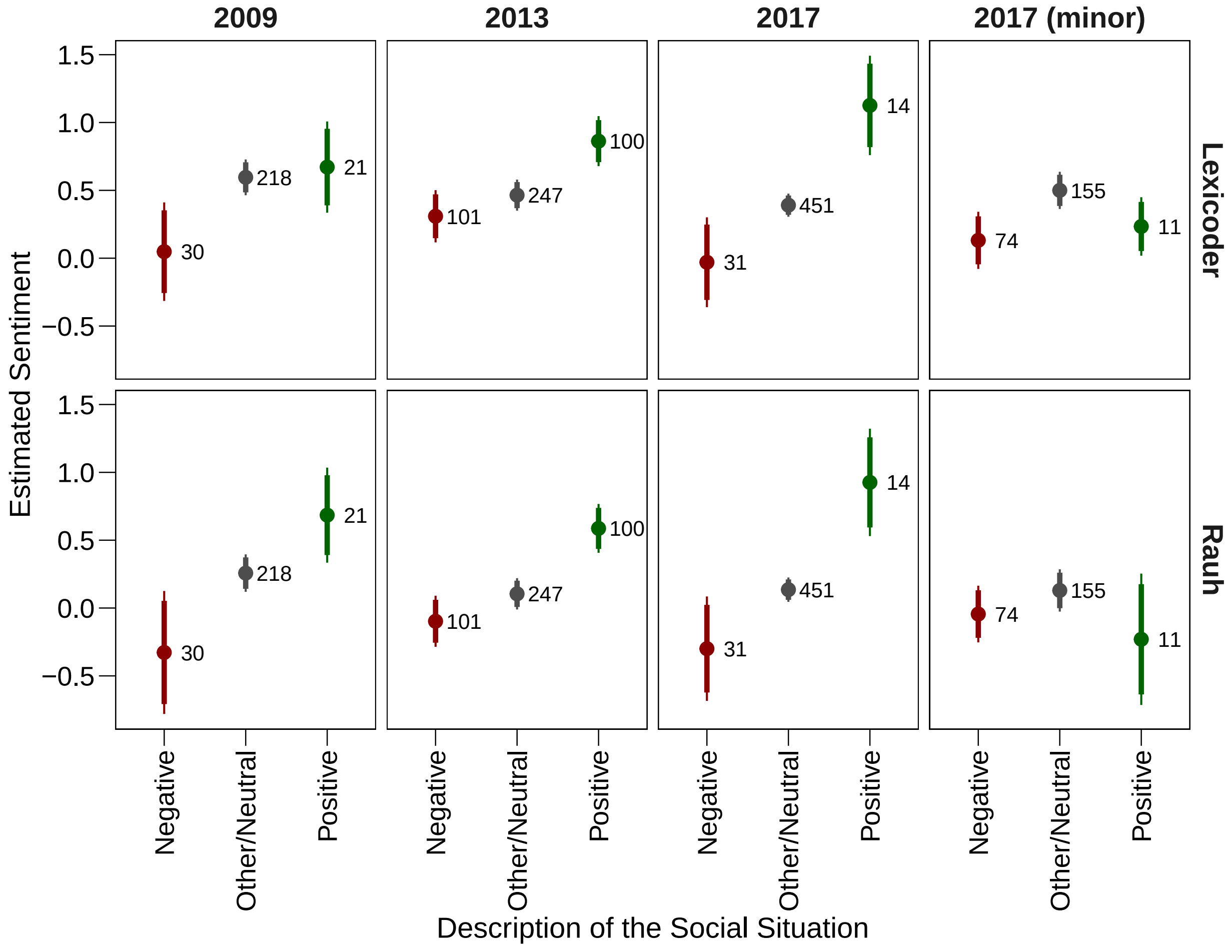


Figure A8: Speaker sentiment for statements coded as a negative, positive, or neutral/other description of the social situation. Vertical bars show 90% and 95% confidence intervals. The numbers beside each point show the number of statements classified into each of the three categories.

C.2.2 Masculine vs Feminine Sentiment

We also test explicitly whether gendered language in the debates is distributed unevenly across male and female candidates. We machine-translate the statement-level debates from 2009, 2013, and 2017 to English and follow the approach by Roberts and Utych (2020). Having created a document-term frequency matrix of the text translated to English, we select terms from a list of gendered terms. Each of these terms was scored by the coders that Roberts and Utych (2020) recruited through MTurk. Higher scores indicate a more 'masculine' language. The word scores range from 1.46 to 6.4. 198 of the 701 coded words appear in our text corpus. We apply the word scores to terms that appeared in the list of gendered language and calculate the sums for all scores

	(1) All coders	(2) Female	(3) Male	(4) All coders	(5) Female	(6) Male
Intercept	1.39***	1.35***	1.42***	1.39***	1.35***	1.42***
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
Speaker: Merkel (ref.: Male)	0.26**	0.24**	0.27**			
	(0.09)	(0.09)	(0.10)			
Speaker: Female (ref.: Male)				0.25**	0.24**	0.27**
				(0.09)	(0.09)	(0.09)
R^2	0.04	0.04	0.04	0.04	0.04	0.04
Adj. R^2	0.04	0.04	0.04	0.04	0.04	0.04
Num. obs.	2022	2022	2022	2262	2262	2262

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A1: Predicting the masculinity of language using method introduced by Roberts and Utych (2020). Models 1–3 limit the analysis to the four debates involving Angela Merkel. Models 4–6 also consider the minor party debate. All models include debate fixed effects.

per statement.[3] Higher values imply that a statement included more ‘masculine’ language.

We then ran a linear regression with the statement-level scores of masculine language as the dependent variable (Table A1). We run the analysis for the scores based on the assessment from all crowd coders (M1 and M4), female coders (M2 and M5) and male coders (M3 and M6). Models 1–3 limit the analysis to debates that include Angela Merkel. The regression models suggest that the Angela Merkel tends to use slightly *more* masculine terms than the male candidates. Yet, this difference of 0.24–0.26 is substantively small and the model fit is poor (R^2 of 0.04). Models 4–6 also include the minor party debate. The substantive conclusions do not change: male speakers do not consistently employ more masculine rhetoric.

[3] To account for differences in the length of statements, we calculate relative frequencies before applying the word scores. Results do not change when using an unweighted document-term frequency matrix instead.

Appendix D Descriptive Statistics about the Debates, Topics, and Audience Characteristics

Here we present descriptive data on the correlation between our candidate emotions, how we code the issue areas for the topics of debate discussions, and how emotions relate to topics. We then provide descriptives on the audience members and respondents in representative German elections studies and information on the RTR data.

D.1 Emotions and Topics Descriptives

Figure A9 provides the correlations across our key measures of emotion, moving from facial displays (non-neutral, anger, and happiness), sentiment, and pitch. Four key patterns emerge: first, we have face validity in our measures that happiness is a more frequent emotion and is positively correlated with non-neutral facial displays and negatively correlated with anger. Second, we see that pitch is positively correlated with non-neutral displays and happiness (consistent with research from Giannakopoulos and Pikrakis (2014)). Third, two surprising correlations: pitch is *negatively* correlated with anger and sentiment negatively associated with happiness; this suggests to us that candidates may use facial, pitch, and sentiment to offset each other at times. And finally, the general *low level* of correlation across most of the measures reaffirms the importance of engaging in multimodal investigations.

Table A2: Summary of coded policy areas

Policy area	Category
Crime	Neutral
Economy	Neutral
Economy	Masculine
Education	Feminine
Environment	Neutral
Foreign Policy	Masculine
Government Affairs (General)	Neutral
Health	Feminine
Immigration	Neutral
Infrastructure	Masculine
Labor	Neutral
Misc	Neutral
Taxes	Neutral
Welfare	Feminine
Women	Feminine

Table A2 provides the coding scheme for how we assigned gendered topics to more general topics within the debates. These masculine, feminine, neutral, and none categories are based on work on gendered stereotypes of issues (Cassese and Holman 2018, Schneider and Bos 2019).

Figure A10 presents the frequency with which candidates discuss specific topics while displaying emotions. We also provide the gendered classification of topics. As the figure displays, anger is most likely to be observed when the speakers are discussing crime, foreign policy, and taxes, all topics where anger would be situationally appropriate. Similarly, we see happiness and non-neutral facial displays during the ”none” topics—which reflects smiles at the beginning and ends of each debate—and for government affairs and a miscellaneous category. Sentiment and vocal pitch change less depending on topic.

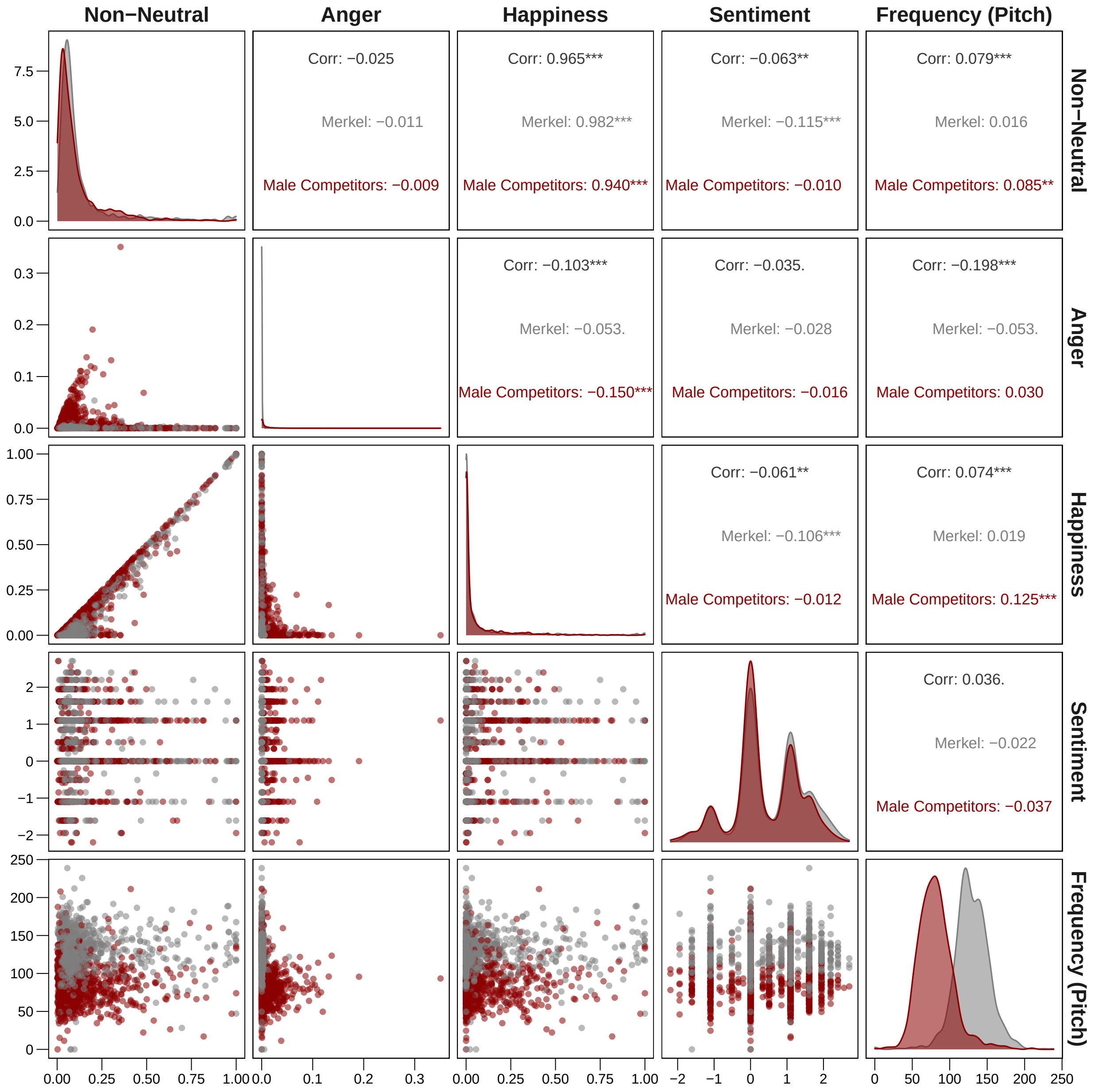


Figure A9: Correlation between unstandardized segment-level emotional expressions for Merkel and her male competitors. Each dot reports the average emotional expression across a statement in a debate.

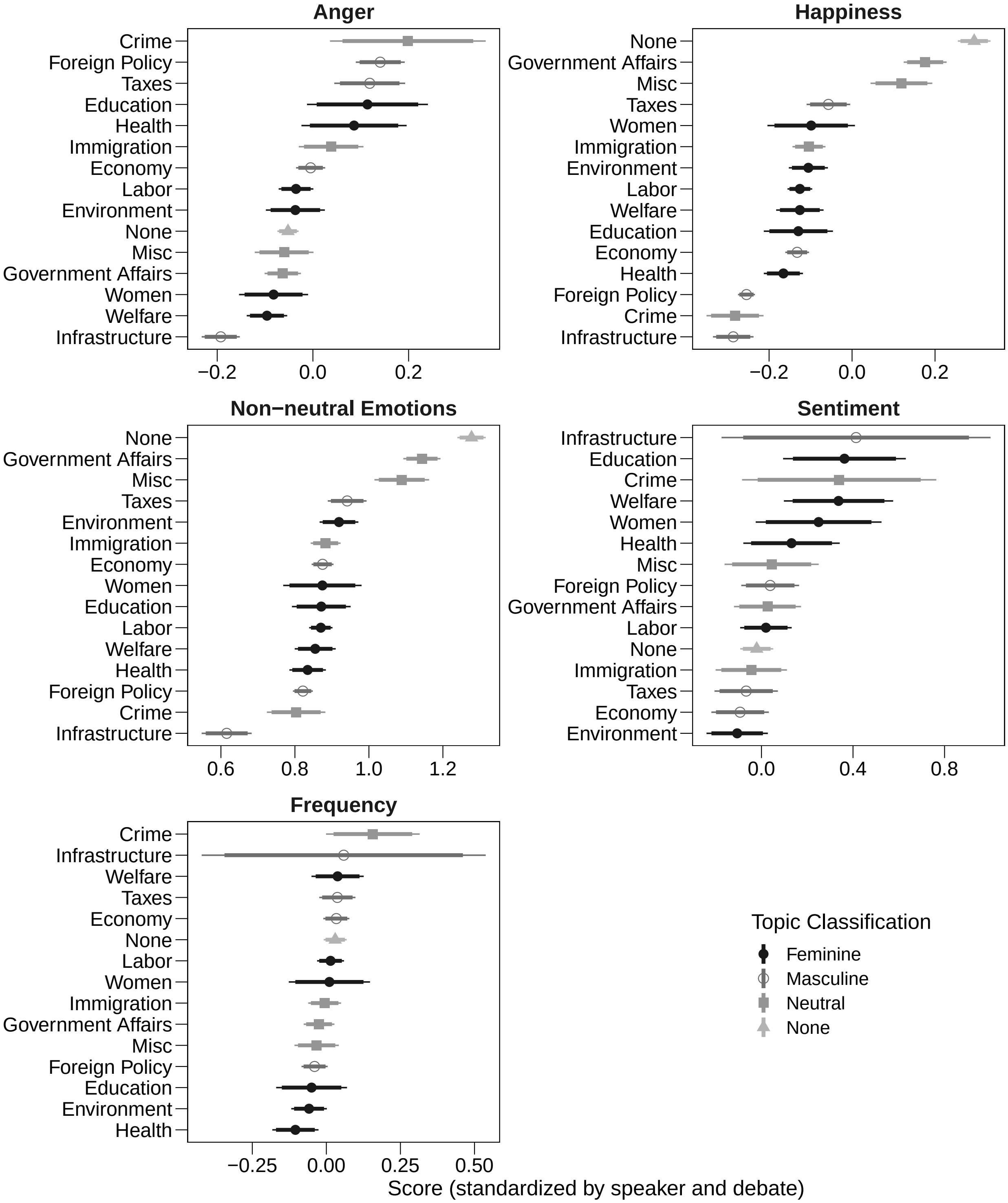


Figure A10: Topic-specific averages of standardized emotions (anger, happiness, non-neutral emotions, frequency, and sentiment) or Merkel and her opponents and proportions of sentences with a fundamental frequency at least one standard deviation above a speaker's average frequency. The Horizontal bars show 90% and 95% confidence intervals.

D.2 Audience Descriptives

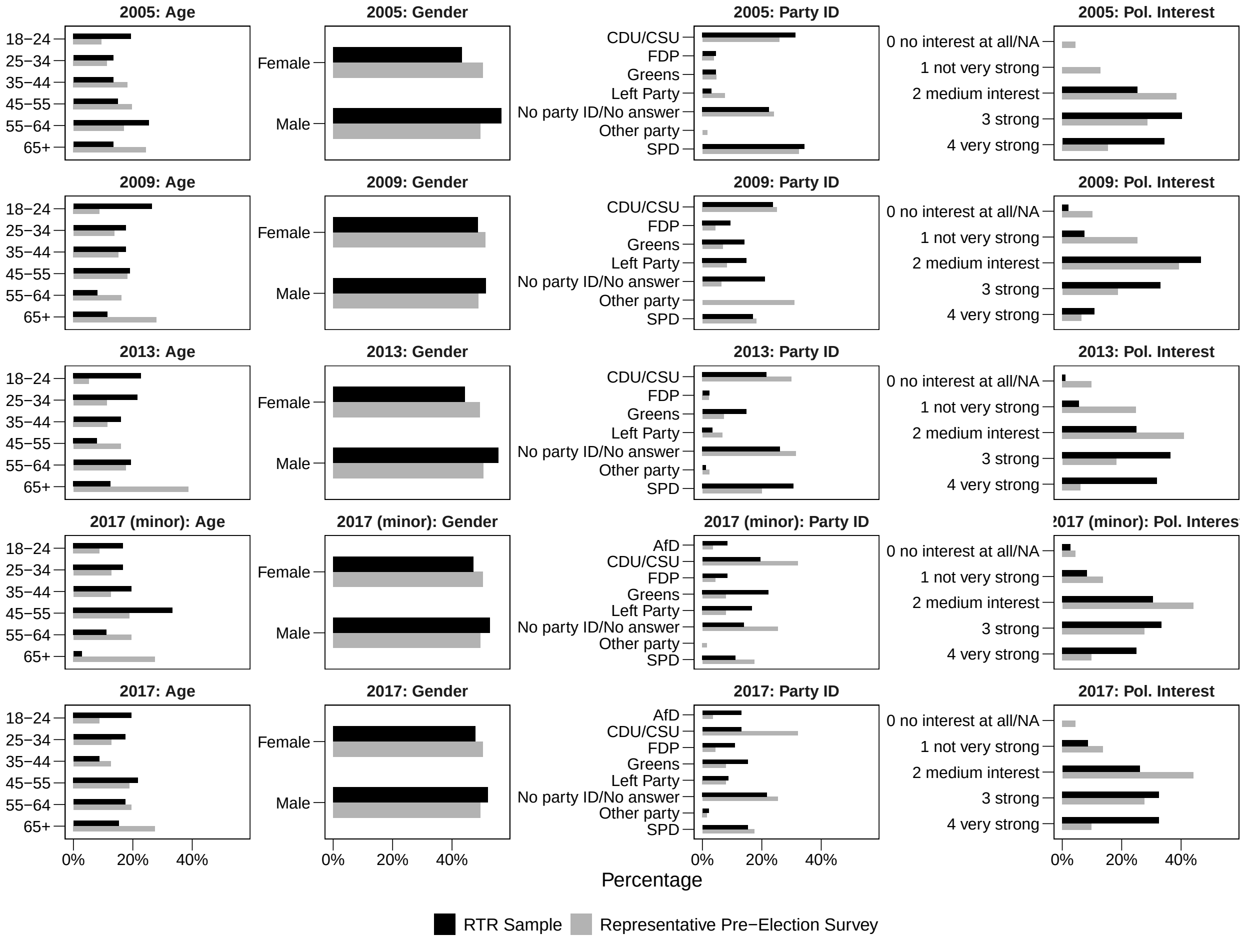


Figure A11: Comparing audience members of with respondents of representative German pre-election studies in the same year, conducted prior to the respective general election. The bars shows the percentages of audience members (black) and participants in election studies (gray) falling into each category.

Figure A11 assesses whether the samples of RTR respondents across the five debates are representative. We retrieve the raw data of the German pre-election studies in 2005, 2009, 2013, and 2017 (Kühnel, Niedermayer and Westle 2012, GLES 2019*a*;*b*;*c*). For each election study, we aggregate the characteristics of respondents in terms gender, party identification, age, and political interest. We repeat the same analysis for the five RTR samples and calculate the same proportions across the same survey items. The RTR respondents show high similarities to respondents in the election studies. RTR respondents tend to be younger and slightly more interested in politics than respondents in the election study. Importantly, we do not observe any systematic biases in terms of party identification across the debates involving Angela Merkel. However, as described in the paper, the number of RTR respondents is substantively lower for the debate between the minor parties, and respondents who identify with one of the five smaller parties are over-represented.

Figure A12 summarizes how often and to what extent debate audience members change the position of their response dial. Each dot in the plot shows an audience member. Recall that the dial ranges from 1–7 in all debates. As the left-hand panel indicates, the standard deviation of the dial on the level of respondents usually ranges between 0.7 and 1.2. The right-hand panel shows how often respondents, on average, change the dial position per minute. The boxplots underscore that most respondents move the dial only around 2–4 times per minute. To sum up, while the scale ranges from 1–7, respondents do not change the dial position very often, and large, abrupt changes on the dial are rare (see also Maier and Faas 2019 79).

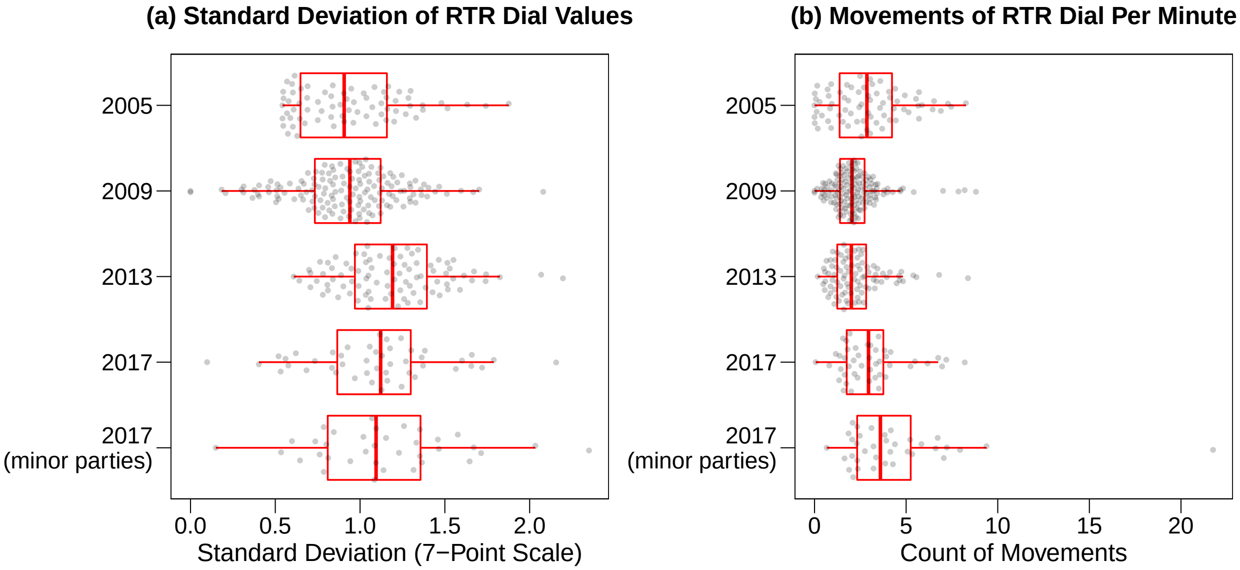


Figure A12: Comparing the standard deviations of the RTR dial values (panel a) and the movements of the RTR dial per minute. Each dot marks the changes in standard deviation by one respondent.

Appendix E Supplementary Tables

Tables A3, A4, A5, A6, and A11 show the regression tables corresponding to the coefficients in Figures 3 and 4. All models include utterance fixed effects and control for the gendered topic indicator.

Tables A7–A12 show the cumulative effects for the coefficients of interest for the voter-level models. We report the cumulative effects across four lags, along with the standard errors, p-value, and 95% confidence intervals. The estimates from these tables are displayed in Figures 5–7.

E.1 Tables for Candidate Results

Table A3: Candidate level regression results for 2005 debate

	(1)	(2)	(3)	(4)	(5)	(6)
	Happiness	Anger	Non-neutral Emotions	Sentiment	Pitch (+1 SD)	Pitch (+1.5 SD)
Merkel	0.0370*	-0.00314***	0.0502***	0.0806	0.352	-0.382
	(0.0153)	(0.000649)	(0.0146)	(0.0573)	(0.590)	(0.623)
Observations	5172	5172	5172	929	3957	3603
R^2	0.093	0.041	0.099	0.009		
Pseudo R^2					0.055	0.071

Standard errors in parentheses
All models include utterance fixed effects and gendered topic indicator variables.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4: Candidate level regression results for 2009 debate

	(1)	(2)	(3)	(4)	(5)	(6)
	Happiness	Anger	Non-neutral Emotions	Sentiment	Pitch (+1 SD)	Pitch (+1.5 SD)
Merkel	-0.0160	-0.0152***	-0.0247	0.332**	0.710*	0.230
	(0.0172)	(0.00175)	(0.0161)	(0.103)	(0.351)	(0.419)
Observations	4583	4583	4583	340	3576	3254
R^2	0.186	0.071	0.187	0.041		
Pseudo R^2					0.046	0.057

Standard errors in parentheses
All models include utterance fixed effects and gendered topic indicator variables.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5: Candidate level regression results for 2013 debate

	(1)	(2)	(3)	(4)	(5)	(6)
	Happiness	Anger	Non-neutral Emotions	Sentiment	Pitch (+1 SD)	Pitch (+1.5 SD)
Merkel	0.0256	-0.0290***	-0.0608***	0.159	0.0445	-0.0685
	(0.0141)	(0.00230)	(0.0143)	(0.0861)	(0.257)	(0.315)
Observations	5117	5117	5117	491	3471	2966
R^2	0.143	0.260	0.188	0.026		
Pseudo R^2					0.068	0.079

Standard errors in parentheses
All models include utterance fixed effects and gendered topic indicator variables.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A6: Candidate level regression results for 2017 debate

	(1) Happiness	(2) Anger	(3) Non-neutral Emotions	(4) Sentiment	(5) Pitch (+1 SD)	(6) Pitch (+1.5 SD)
Merkel	0.0165	-0.00675***	0.0738***	0.0916	0.185	0.167
	(0.0122)	(0.00139)	(0.0115)	(0.0781)	(0.578)	(0.362)
Observations	5598	5598	5598	555	3436	2410
R^2	0.270	0.174	0.273	0.018		
Pseudo R^2					0.059	0.071

Standard errors in parentheses
All models include utterance fixed effects and gendered topic indicator variables.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

E.2 Tables for Voter Results

Debate	Emotion	Coefficient	Std error	p-value	Lower 95% CI	Upper 95% CI
2005	Anger	0.0136	0.0166	0.4130	-0.0194	0.0467
2005	Happiness	0.0340	0.0248	0.1740	-0.0154	0.0835
2005	Fear	0.0610	0.0122	0.0000	0.0366	0.0853
2005	Disgust	-0.0183	0.0156	0.2450	-0.0494	0.0129
2005	Contempt	-0.0987	0.0234	0.0000	-0.1454	-0.0520
2005	Sadness	-0.0855	0.0325	0.0110	-0.1504	-0.0206
2005	Surprise	0.0866	0.0149	0.0000	0.0569	0.1162
2005	Frequency	0.0598	0.0159	0.0000	0.0280	0.0915
2005	Sentiment	-0.0162	0.0093	0.0850	-0.0347	0.0023
2009	Anger	-0.1027	0.0215	0.0000	-0.1451	-0.0602
2009	Happiness	0.0748	0.0119	0.0000	0.0514	0.0983
2009	Fear	0.1036	0.0262	0.0000	0.0519	0.1554
2009	Disgust	0.0638	0.0117	0.0000	0.0407	0.0868
2009	Contempt	-0.0494	0.0138	0.0000	-0.0766	-0.0222
2009	Sadness	-0.3621	0.0806	0.0000	-0.5214	-0.2027
2009	Surprise	0.0450	0.0192	0.0210	0.0070	0.0829
2009	Frequency	0.0253	0.0162	0.1200	-0.0066	0.0572
2009	Sentiment	0.0420	0.0083	0.0000	0.0256	0.0583
2013	Anger	-0.3218	0.0287	0.0000	-0.3789	-0.2648
2013	Happiness	0.0435	0.0165	0.0100	0.0106	0.0763
2013	Fear	-0.2631	0.0448	0.0000	-0.3522	-0.1740
2013	Disgust	0.1227	0.0240	0.0000	0.0750	0.1705
2013	Contempt	-0.0330	0.0321	0.3070	-0.0968	0.0308
2013	Sadness	0.3763	0.0564	0.0000	0.2641	0.4884
2013	Surprise	-0.0057	0.0195	0.7680	-0.0444	0.0329
2013	Frequency	0.0075	0.0238	0.7520	-0.0397	0.0547
2013	Sentiment	0.0081	0.0114	0.4800	-0.0146	0.0309
2017	Anger	0.0270	0.0161	0.1010	-0.0055	0.0595
2017	Happiness	0.1288	0.0321	0.0000	0.0641	0.1935
2017	Fear	0.0236	0.0207	0.2600	-0.0180	0.0653
2017	Disgust	-0.0791	0.0239	0.0020	-0.1271	-0.0310
2017	Contempt	-0.3161	0.0944	0.0020	-0.5063	-0.1259
2017	Sadness	0.0628	0.0199	0.0030	0.0227	0.1028
2017	Surprise	0.0264	0.0234	0.2650	-0.0207	0.0735
2017	Frequency	0.0765	0.0261	0.0050	0.0240	0.1291
2017	Sentiment	0.0336	0.0122	0.0080	0.0090	0.0582

Table A7: Cumulative effects across 4 lags for anger, happiness, sentiment, and avg. fundamental frequency.

Debate	Emotion	Coefficient	Std error	p-value	Lower 95% CI	Upper 95% CI
2005	Noneutral	0.0021	0.0190	0.9110	-0.0358	0.0400
2005	Frequency	0.0759	0.0170	0.0000	0.0419	0.1099
2005	Sentiment	-0.0167	0.0098	0.0930	-0.0363	0.0029
2009	Noneutral	0.0711	0.0130	0.0000	0.0453	0.0968
2009	Frequency	0.0383	0.0157	0.0160	0.0073	0.0693
2009	Sentiment	0.0410	0.0083	0.0000	0.0246	0.0574
2013	Noneutral	0.0589	0.0172	0.0010	0.0248	0.0930
2013	Frequency	0.0244	0.0242	0.3150	-0.0236	0.0725
2013	Sentiment	-0.0024	0.0117	0.8370	-0.0257	0.0208
2017	Noneutral	0.1225	0.0322	0.0000	0.0577	0.1874
2017	Frequency	0.0693	0.0263	0.0120	0.0163	0.1224
2017	Sentiment	0.0392	0.0126	0.0030	0.0137	0.0646

Table A8: Cumulative effects across 4 lags for non-neutral emotions, sentiment, and avg. fundamental frequency.

Candidate	Year	Emotion	Coefficient	Std error	p-value	Lower 95% CI	Upper 95% CI
Merkel	2005	Anger	-0.0286	0.0145	0.0520	-0.0576	0.0003
Merkel	2005	Happiness	0.0274	0.0134	0.0440	0.0007	0.0542
Merkel	2005	Frequency	0.0993	0.0119	0.0000	0.0756	0.1230
Merkel	2005	Sentiment	0.0105	0.0074	0.1630	-0.0043	0.0253
Merkel	2009	Anger	-0.0707	0.0140	0.0000	-0.0985	-0.0430
Merkel	2009	Happiness	0.0251	0.0094	0.0080	0.0066	0.0436
Merkel	2009	Frequency	0.0018	0.0117	0.8770	-0.0213	0.0249
Merkel	2009	Sentiment	0.0137	0.0063	0.0320	0.0012	0.0262
Merkel	2013	Anger	-0.1266	0.0200	0.0000	-0.1664	-0.0867
Merkel	2013	Happiness	0.0295	0.0147	0.0480	0.0002	0.0587
Merkel	2013	Frequency	0.0198	0.0220	0.3710	-0.0240	0.0636
Merkel	2013	Sentiment	-0.0031	0.0110	0.7820	-0.0250	0.0188
Merkel	2017	Anger	0.0447	0.0135	0.0020	0.0176	0.0718
Merkel	2017	Happiness	0.0469	0.0196	0.0210	0.0074	0.0864
Merkel	2017	Frequency	0.1078	0.0212	0.0000	0.0651	0.1506
Merkel	2017	Sentiment	0.0166	0.0097	0.0960	-0.0031	0.0362
Male Opponent	2005	Anger	-0.0324	0.0087	0.0000	-0.0498	-0.0150
Male Opponent	2005	Happiness	-0.0268	0.0212	0.2110	-0.0691	0.0155
Male Opponent	2005	Frequency	0.0492	0.0135	0.0000	0.0223	0.0760
Male Opponent	2005	Sentiment	0.0066	0.0070	0.3480	-0.0073	0.0205
Male Opponent	2009	Anger	0.0397	0.0184	0.0330	0.0033	0.0761
Male Opponent	2009	Happiness	-0.0920	0.0091	0.0000	-0.1099	-0.0740
Male Opponent	2009	Frequency	-0.0271	0.0138	0.0510	-0.0543	0.0001
Male Opponent	2009	Sentiment	-0.0232	0.0062	0.0000	-0.0354	-0.0109
Male Opponent	2013	Anger	0.2486	0.0298	0.0000	0.1893	0.3079
Male Opponent	2013	Happiness	-0.0220	0.0143	0.1290	-0.0504	0.0065
Male Opponent	2013	Frequency	0.0049	0.0140	0.7250	-0.0229	0.0328
Male Opponent	2013	Sentiment	-0.0085	0.0104	0.4190	-0.0292	0.0123
Male Opponent	2017	Anger	0.0262	0.0143	0.0740	-0.0026	0.0550
Male Opponent	2017	Happiness	-0.0939	0.0328	0.0060	-0.1600	-0.0278
Male Opponent	2017	Frequency	0.0323	0.0185	0.0870	-0.0049	0.0696
Male Opponent	2017	Sentiment	-0.0238	0.0111	0.0380	-0.0461	-0.0014

Table A9: Cumulative effects across 4 lags for anger, happiness, sentiment, and avg. fundamental frequency. Separate models for each candidate in the four debates.

Candidate	Year	Emotion	Coefficient	Std error	p-value	Lower 95% CI	Upper 95% CI
Merkel	2005	Non-Neutral Emotion	0.0175	0.0119	0.1480	-0.0063	0.0412
Merkel	2005	Frequency	0.1134	0.0132	0.0000	0.0870	0.1399
Merkel	2005	Sentiment	0.0094	0.0076	0.2200	-0.0057	0.0244
Merkel	2009	Non-Neutral Emotion	0.0315	0.0113	0.0060	0.0090	0.0539
Merkel	2009	Frequency	0.0196	0.0122	0.1100	-0.0045	0.0436
Merkel	2009	Sentiment	0.0140	0.0063	0.0270	0.0016	0.0264
Merkel	2013	Non-Neutral Emotion	0.0230	0.0131	0.0820	-0.0030	0.0491
Merkel	2013	Frequency	0.0368	0.0227	0.1090	-0.0084	0.0821
Merkel	2013	Sentiment	-0.0076	0.0110	0.4900	-0.0294	0.0142
Merkel	2017	Non-Neutral Emotion	0.0415	0.0207	0.0500	-0.0001	0.0832
Merkel	2017	Frequency	0.1192	0.0213	0.0000	0.0763	0.1622
Merkel	2017	Sentiment	0.0179	0.0100	0.0800	-0.0022	0.0379
Male Opponent	2005	Non-Neutral Emotion	0.0026	0.0162	0.8750	-0.0297	0.0349
Male Opponent	2005	Frequency	0.0444	0.0135	0.0020	0.0174	0.0714
Male Opponent	2005	Sentiment	0.0036	0.0069	0.6020	-0.0102	0.0174
Male Opponent	2009	Non-Neutral Emotion	-0.0765	0.0091	0.0000	-0.0945	-0.0585
Male Opponent	2009	Frequency	-0.0220	0.0134	0.1030	-0.0485	0.0045
Male Opponent	2009	Sentiment	-0.0227	0.0066	0.0010	-0.0358	-0.0097
Male Opponent	2013	Non-Neutral Emotion	-0.0398	0.0176	0.0260	-0.0747	-0.0049
Male Opponent	2013	Frequency	0.0052	0.0141	0.7130	-0.0228	0.0332
Male Opponent	2013	Sentiment	0.0022	0.0103	0.8350	-0.0183	0.0226
Male Opponent	2017	Non-Neutral Emotion	-0.0953	0.0312	0.0040	-0.1582	-0.0325
Male Opponent	2017	Frequency	0.0575	0.0192	0.0040	0.0188	0.0962
Male Opponent	2017	Sentiment	-0.0231	0.0115	0.0500	-0.0462	0.0001

Table A10: Cumulative effects across 4 lags lags for non-neutral facial emotions, sentiment, and avg. fundamental frequency. Separate models for each candidate in the four debates.

Table A11: Candidate level regression results for 2017 minor party debate

	(1)	(2)	(3)	(4)	(5)	(6)
	Happiness	Anger	Non-neutral Emotions	Sentiment	Pitch (+1 SD)	Pitch (+1.5 SD)
Female	0.0183	-0.00571***	0.000933	-0.0840	0.853*	-0.310
	(0.0196)	(0.00128)	(0.0181)	(0.107)	(0.404)	(0.502)
Observations	5023	5023	5023	278	1922	1237
R^2	0.107	0.038	0.112	0.048		
Pseudo R^2					0.144	0.146

Standard errors in parentheses

All models include utterance fixed effects and gendered topic indicator variables.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Debate	Emotion	Coefficient	Std error	p-value	Lower 95% CI	Upper 95% CI
2017 (minor parties)	Anger	-0.2853	0.0606	0.0000	-0.4084	-0.1623
2017 (minor parties)	Happiness	0.0235	0.0493	0.6370	-0.0766	0.1236
2017 (minor parties)	Frequency	0.0486	0.0360	0.1860	-0.0244	0.1216
2017 (minor parties)	Sentiment	0.0814	0.0304	0.0110	0.0196	0.1432

Table A12: Cumulative effects across 4 lags for anger, happiness, sentiment, and avg. fundamental frequency

Appendix F Robustness Checks

Figure A13 compares the averages for our six emotional expressions of interest for the two candidates in each debate. The y-axes differ for each facet in order to allow for better comparability of Merkel and her male opponents. Merkel expressed considerably more happiness in her first debate, compared to debates she contested as the chancellor. The panel on anger confirms the results from the main text. Merkel displays almost no anger at all. The levels of expressed anger by her male competitors are also on a low level, but still higher. Peer Steinbrück (2013) expressed by far the highest levels of anger. Merkel's levels of non-neutral emotional displays are comparable over time and on similar levels to her male opponents. When considering a frequency exceeding one standard deviation of the candidate's mean as a benchmark for high voice pitch, we observe that Merkel was slightly more emotional than her opponents. Merkel's pitch was one standard deviations above her mean in around 15% of all seconds. The values for the opponents range between 13% and 14%. When using a more restrictive measure of 1.5 standard deviations above the mean, the difference between Merkel and her opponents decreases, with the debate in 2017 being an exception. Merkel's aggregated sentiment was slightly more positive than the sentiment of her opponents. These descriptive results corroborate with the findings from the candidate-level regression models. In addition, this figure underscores that Merkel's emotional displays have been remarkably stable between 2005 and 2017, with the exception of happiness displays.

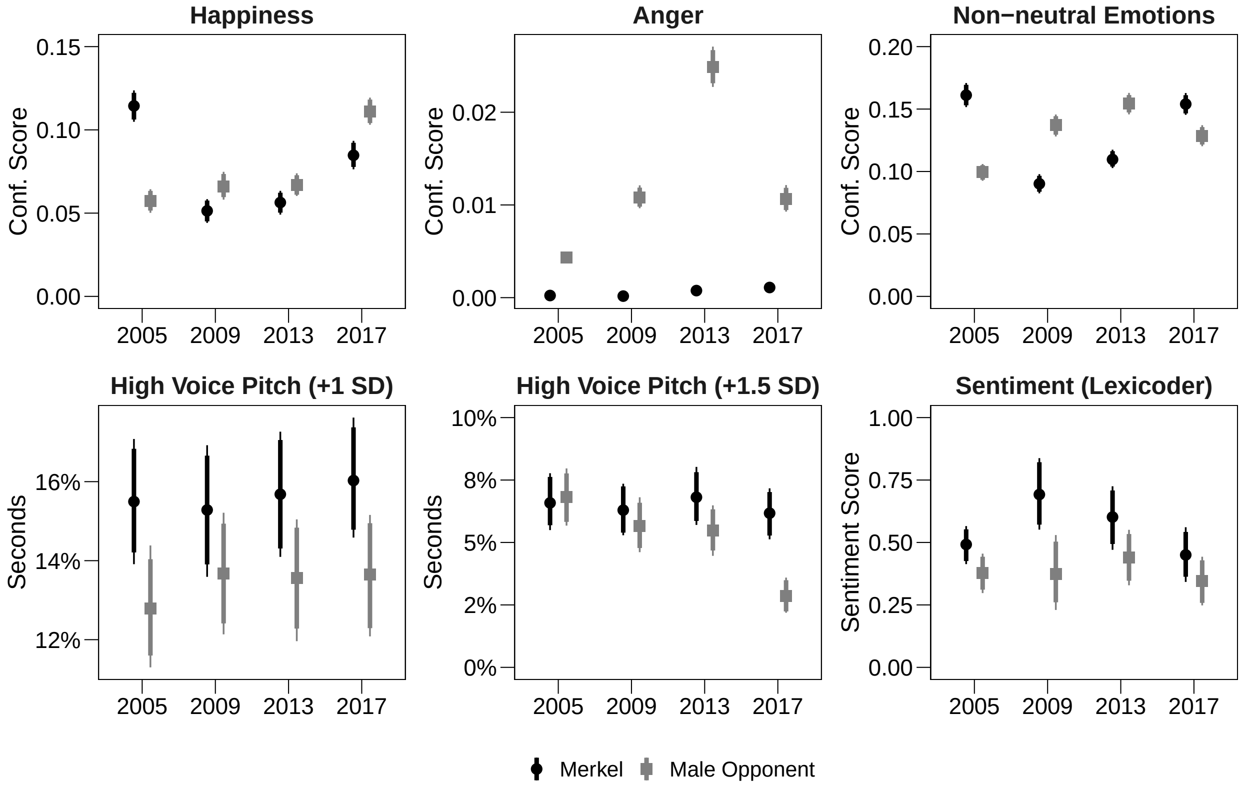


Figure A13: Merkel vs her opponents: Comparing the average values of anger, happiness, non-neutral facial emotions, sentiment, and voice pitch. Vertical bars show 90% and 95% bootstrapped confidence intervals. The y-axes are rescaled for each facet to ease comparability between candidates and Merkel's emotions over time.

Figure A14 plots men's and women's reactions to the relevant facial, verbal, and textual emotions. The figure reports two separate models: a model that only includes female audience members (left-hand panel) and a model restricted to male audience members (right-hand panel). Overall, we observe high correspondence in the cumulative effects across the variables, indicating that our findings are not driven by the gender of audience members.

Figure A15 reproduces the main analysis using an an alternative sentiment dictionary (Rauh 2018). The coefficients for sentiment larger than the baseline dictionary (Lexicoder), suggesting that the results reported in the main paper are conservative estimates of the relationship between textual sentiment and voter reactions. The cumulative effects of the remaining variables do not depend on the sentiment dictionary.

Figure A16 shows the cumulative effects for the minor debate based on our measure of non-neutral emotional facial expressions instead of specific facial emotions (as reported in Figure 7).

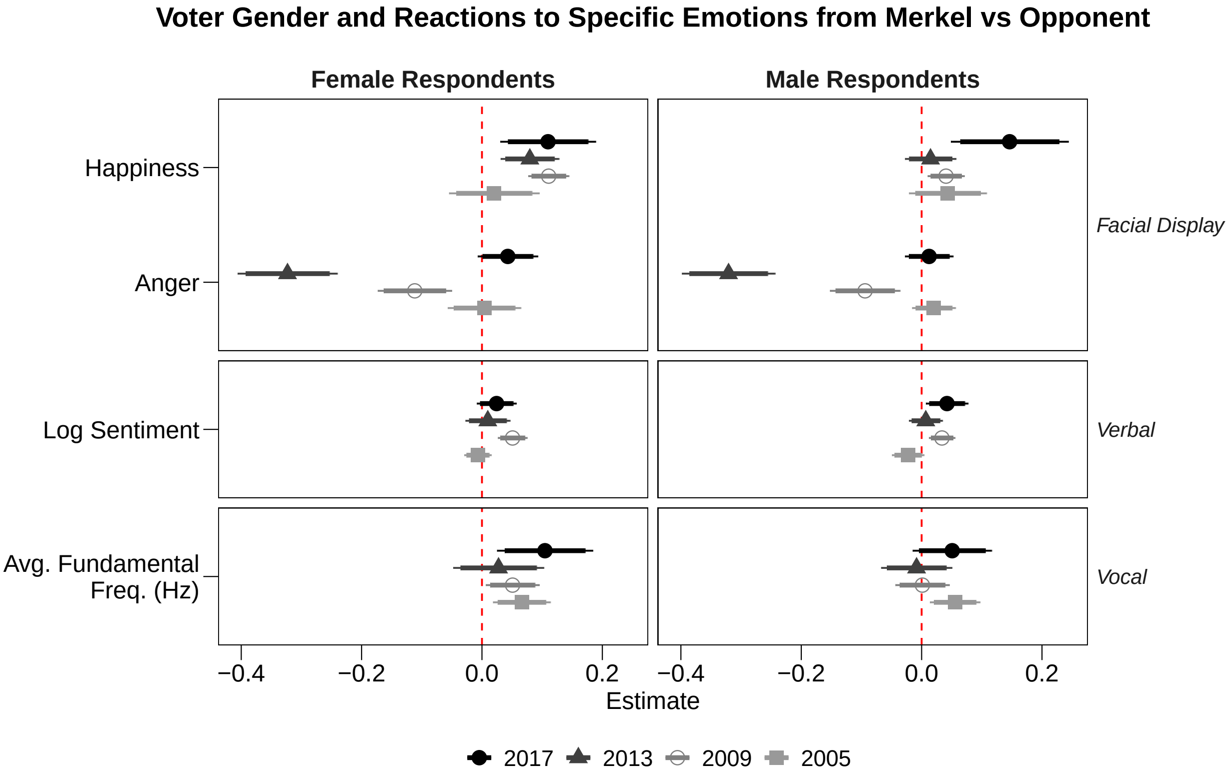


Figure A14: Men's and women's reactions to the cumulative effect (across four lags) of the key textual, vocal, and facial variables of interest. The models include control variables for the party identification, political knowledge, and political interest of respondents. The left-hand panel reports results of four models that only female respondents. The right-hand panel reports models that include only the male respondents of each debate. Horizontal bars show 90% and 95% confidence intervals.

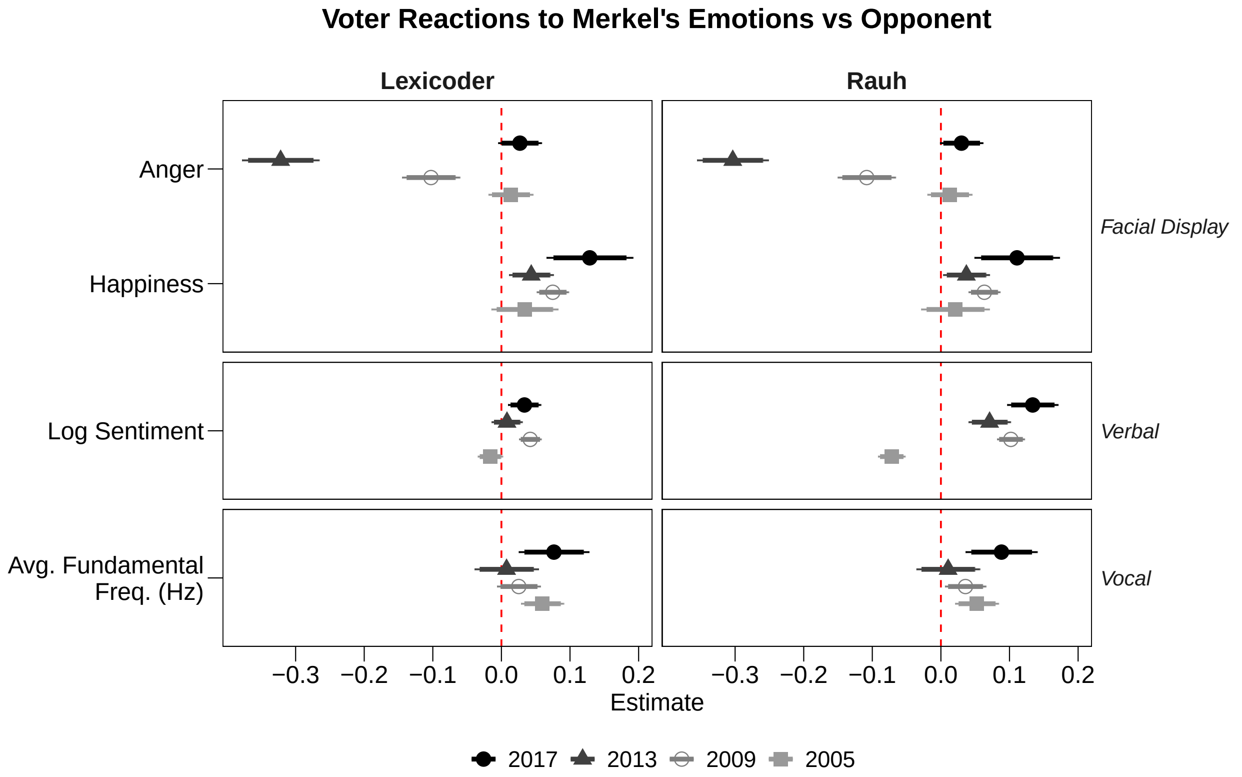


Figure A15: Voter reactions to the cumulative effect (across four lags) of the key textual, vocal, and facial variables of interest. The models include control variables for the gender, party identification, political knowledge, and political interest of respondents. The left-hand panel reports results using the Lexicoder Sentiment Dictionary. The right-hand panel reruns the analysis using Rauh's sentiment dictionary. Horizontal bars show 90% and 95% confidence intervals.

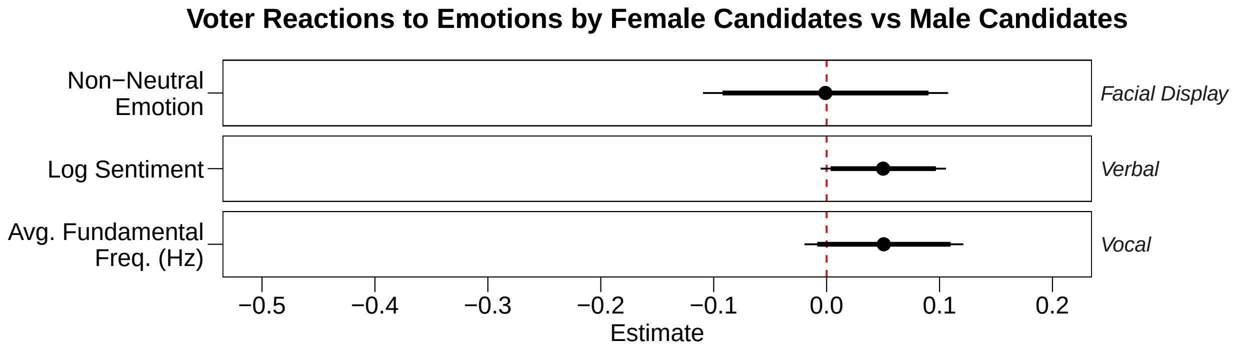


Figure A16: Voter reactions to candidate emotions in the 2017 debate of minor parties. The figure provides an estimate of the cumulative effect (across four lags) of the key textual, vocal, and facial variables of interest. This figure presents the estimates for non-neutral facial expressions of emotion. Positive coefficients indicate that respondents tend to react positively to emotional expressions by the two female candidates. Horizontal bars show 90% and 95% confidence intervals.

Appendix G Ethics and Transparency

In this section, we summarize the procedures for collecting our data. All practices correspond to the transparency obligations described in "A Guide to Professional Ethics in Political Science" and in "Principles and Guidance for Human Subjects Research" (2020).

We collected labels on emotional displays by German candidates in five debates using the Face API from Microsoft Azure Cognitive Services.[4] The images of frames that we uploaded come from publicly available sources. This part of the research design does not involve any human participants.

We match the data retrieved from the facial and emotion recognition systems with real-time-response data of German voters who participated voluntarily in the experiments. We did not collect this data, but rely on data collected and generously shared by other researchers. 2005 Debate: "72 participants were recruited using newspaper articles in the local press. Subjects were offered 25 EUR for their participation. As more subjects applied than seats were available, they were selected using quota sampling (political predispositions, educational levels, gender, and age)" (Nagel, Maurer and Reinemann 2012 838). The authors of this study shared the anonymized replication data of their paper (Nagel, Maurer and Reinemann 2012) with us in April 2020 for our study.

2009, 2013, and 2017 debates: samples were collected and administered by the German Longitudinal Election Study. All datasets are freely available online at the GESIS homepage.[5] According to their website "With more than 300 employees at two locations – Mannheim and Cologne – GESIS provides essential and internationally relevant research-based services for the social sciences. As the largest European infrastructure institute for the social sciences GESIS offers advice, expertise and services at all stages of scientists' research projects. With this support socially relevant questions can be answered based on the latest scientific methods, and with high quality research data."[6] The experimental group was offered an allowance of 25 EURO (2013) or 40 EURO (in 2009 and 2017). Respondents were recruited through press releases and ads and were informed about the design of the study. Respondents also received extensive information on how the survey instruments (the dial buttons) work and that the position of their dials would be saved at every second during the debate. The data collection procedures are summarised in the codebooks of the following studies: Rattinger et al. (2010; 2011*a*;*b*; 2014; 2015; 2018), Roßteutscher, Schmitt-Beck, Schoen, Weßels, Wolf, Brettschneider, Faas, Maier and Maier (2019*a*;*b*), Roßteutscher, Schmitt-Beck, Schoen, Weßels, Wolf, Faas, Maier and Maier (2019*c*;*a*;*b*).

Given that we were not involved in collecting the original data, we had no influence in the compensation that was paid to the respondents. Yet, the allowance of EUR 25 or EUR 40 seems fair and justified given that respondents spent approximately two hours at the location where their responses to the debates were stored.

For our validation exercises, we hired undergraduate research assistants who work for an hourly wage for one of the PIs as our trained coders. The research assistants were compensated for their time both in training and in completing the data tasks. We then recruited workers through the CID survey platform to complete the additional validation exercise; (approved via Tulane University Institutional Review Board; Study Number 2021-298). Respondents were compensated by LUCID directly for this work.

[4] https://azure.microsoft.com/en-us/services/cognitive-services/face/.

[5] See https://search.gesis.org/ and https://gles-en.eu/download-data/.

[6] https://www.gesis.org/en/institute.

References

ACE Electoral Knowledge Network. 2021. “Media and Elections: Television Debates.”.
URL: *https://aceproject.org/epic-en/CDTable?view=country&question=ME059*

Boussalis, Constantine and Travis G. Coan. 2021. “Facing the Electorate: Computational Approaches to the Study of Nonverbal Communication and Voter Impression Formation.” *Political Communication* 38(1–2):75–97.

Cassese, Erin C. and Mirya R. Holman. 2018. “Party and Gender Stereotypes in Campaign Attacks.” *Political Behavior* 40(3):785–807.

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

Faas, Thorsten. 2010. “The German Federal Election of 2009: Sprouting Coalitions, Dropping Social Democrats.” *West European Politics* 33(4):894–903.

Faas, Thorsten. 2015. “The German Federal Election of 2013: Merkel’s Triumph, the Disappearance of the Liberal Party, and Yet Another Grand Coalition.” *West European Politics* 38(1):238–247.

Faas, Thorsten and Tristan Klingelhöfer. 2019. “The More Things Change, the More They Stay the Same? The German Federal Election of 2017 and Its Consequences.” *West European Politics* 42(4):914–926.

Giannakopoulos, Theodoros and Aggelos Pikrakis. 2014. *Introduction to Audio Analysis: a MATLAB Approach.* Amsterdam: Elsevier Science.

GLES. 2019*a*. “Pre-Election Cross Section (GLES 2009).” GESIS Data Archive, Cologne. ZA5300 Data file Version 5.0.2.

GLES. 2019*b*. “Pre-Election Cross Section (GLES 2013).” GESIS Data Archive, Cologne. ZA5700 Data file Version 2.0.2.

GLES. 2019*c*. “Pre-Election Cross Section (GLES 2017).” GESIS Data Archive, Cologne. ZA6800 Data file Version 5.0.1.

Hess, Ursula, Reginald B Adams, Karl Grammer and Robert E Kleck. 2009. “Face Gender and Emotion Expression: Are Angry Women More like Men?” *Journal of Vision* 9(12):19.

Kühnel, Steffen, Oskar Niedermayer and Bettina Westle. 2012. “Bundestagswahl 2005 – Bürger und Parteien in einer veränderten Welt.” GESIS Data Archive, Cologne. ZA4332 Data file Version 2.0.0.

Maier, Jürgen and Thorsten Faas. 2019. *TV-Duelle.* Wiesbaden: Springer VS.

Nagel, Friederike, Marcus Maurer and Carsten Reinemann. 2012. “How Verbal, Visual, and Vocal Communication Shape Viewers’ Impressions of Political Candidates.” *Journal of Communication* 65(5):833–850.

Proksch, Sven-Oliver and Jonathan B. Slapin. 2006. “Institutions and Coalition Formation: The German Election of 2005.” *West European Politics* 29(3):540–559.

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

Rattinger, Hans, Sigrid Roßteutscher, Rüdiger Schmitt-Beck, Bernhard Weßels, Christof Wolf, Frank Brettschneider, Thorsten Faas, Jürgen Maier and Michaela Maier. 2014. “TV-Duell-Analyse Real-Time-Response-Messung (Dial) (GLES 2013).” GESIS Data Archive, Cologne. ZA5711 Data file Version 1.0.0.

Rattinger, Hans, Sigrid Roßteutscher, Rüdiger Schmitt-Beck, Bernhard Weßels, Christof Wolf, Frank Brettschneider, Thorsten Faas, Jürgen Maier and Michaela Maier. 2015. “TV-Duell-Analyse Befragung (GLES 2013).” GESIS Data Archive, Cologne. ZA5709 Data file Version 3.0.0.

Rattinger, Hans, Sigrid Roßteutscher, Rüdiger Schmitt-Beck, Bernhard Weßels, Christof Wolf, Frank Brettschneider, Thorsten Faas, Jürgen Maier and Michaela Maier. 2018. “TV-Duell-Analyse Inhaltsanalyse (GLES 2013).” GESIS Data Archive, Cologne. ZA5710 Data file Version 2.2.0.

Rattinger, Hans, Sigrid Roßteutscher, Rüdiger Schmitt-Beck, Bernhard Weßels, Frank Brettschneider, Thorsten Faas, Jürgen Maier and Michaela Maier. 2010. “TV-Duell-Analyse, Inhaltsanalyse TV-Duell (GLES 2009).” GESIS Data Archive, Cologne. ZA5311 Data file Version 1.0.0.

Rattinger, Hans, Sigrid Roßteutscher, Rüdiger Schmitt-Beck, Bernhard Weßels, Frank Brettschneider, Thorsten Faas, Jürgen Maier and Michaela Maier. 2011*a*. “TV-Duell-Analyse, Befragung (GLES 2009).” GESIS Data Archive, Cologne. ZA5309 Data file Version 2.0.0.

Rattinger, Hans, Sigrid Roßteutscher, Rüdiger Schmitt-Beck, Bernhard Weßels, Frank Brettschneider, Thorsten Faas, Jürgen Maier and Michaela Maier. 2011*b*. “TV-Duell-Analyse, Real-Time-Response-Daten

(GLES 2009)." GESIS Data Archive, Cologne. ZA5310 Data file Version 1.1.0.

Rauh, Christian. 2018. "Validating a Sentiment Dictionary for German Political Language: A Workbench Note." *Journal of Information Technology & Politics* 15(4):319–343.

Roberts, Damon C. and Stephen M. Utych. 2020. "Linking Gender, Language, and Partisanship." *Political Research Quarterly* 73(1):40–50.

Roberts, Geoffrey K. 2006. "The German Bundestag Election 2005." *Parliamentary Affairs* 59(4):668–681.

Roßteutscher, Sigrid, Rüdiger Schmitt-Beck, Harald Schoen, Bernhard Weßels, Christof Wolf, Frank Brettschneider, Thorsten Faas, Jürgen Maier and Michaela Maier. 2019*a*. "TV-Duell-Analyse, Befragung (GLES 2017)." GESIS Data Archive, Cologne. ZA6810 Data file Version 1.0.0.

Roßteutscher, Sigrid, Rüdiger Schmitt-Beck, Harald Schoen, Bernhard Weßels, Christof Wolf, Frank Brettschneider, Thorsten Faas, Jürgen Maier and Michaela Maier. 2019*b*. "TV-Duell-Analyse, Inhaltsanalyse TV-Duell (GLES 2017)." GESIS Data Archive, Cologne. ZA6811 Data file Version 1.0.0.

Roßteutscher, Sigrid, Rüdiger Schmitt-Beck, Harald Schoen, Bernhard Weßels, Christof Wolf, Thorsten Faas, Jürgen Maier and Michaela Maier. 2019*a*. "TV-Duell-Analyse, Inhaltsanalyse Fünfkampf (GLES 2017)." GESIS Data Archive, Cologne. ZA6829 Data file Version 1.0.0.

Roßteutscher, Sigrid, Rüdiger Schmitt-Beck, Harald Schoen, Bernhard Weßels, Christof Wolf, Thorsten Faas, Jürgen Maier and Michaela Maier. 2019*b*. "TV-Duell-Analyse, Real-Time-Response Daten Fünfkampf (dial)." GESIS Data Archive, Cologne. ZA6830 Data file Version 1.0.0.

Roßteutscher, Sigrid, Rüdiger Schmitt-Beck, Harald Schoen, Bernhard Weßels, Christof Wolf, Thorsten Faas, Jürgen Maier and Michaela Maier. 2019*c*. "TV-Duell-Analyse, Real-Time-Response-Daten TV-Duell (dial) (GLES 2017)." GESIS Data Archive, Cologne. ZA6812 Data file Version 1.0.0.

Schneider, Monica C. and Angela L. Bos. 2019. "The Application of Social Role Theory to the Study of Gender in Politics." *Political Psychology* 40(S1):173–213.