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

Constantine Boussalis[1], Travis G. Coan[2], Mirya R. Holman[3], and Stefan Müller[4]

[1]Trinity College Dublin
[2]University of Exeter & Exeter Q-Step Centre
[3]Tulane University
[4]University College Dublin

Version: November 17, 2020

Abstract

Voters evaluate politicians not just by what they say, but also how they say it, via facial displays of emotions and vocal pitch. Candidate characteristics can shape how leaders use – and how voters react to – nonverbal cues. Drawing on role congruity expectations, we focus on how gender shapes the use of and reactions to facial, voice, and textual communication in political debates. Using full-length debate videos from four German national elections (2005–2017) and a minor debate in 2017, we employ computer vision, machine learning, and text analysis to extract facial displays of emotion, vocal pitch, and speech sentiment. Consistent with our expectations, Angela Merkel expresses less anger and is less emotive than her male opponents. We combine second-by-second candidate emotions data with continuous responses recorded by live audiences. We find that voters punish Merkel for anger displays and reward her happiness and general emotional displays.

Constantine Boussalis is Assistant Professor, Trinity College Dublin (boussalc@tcd.ie). Travis G. Coan is Senior Lecturer, University of Exeter and Exeter Q-Step Centre (t.coan@exeter.ac.uk). Mirya R. Holman is Associate Professor, Tulane University (mholman@tulane.edu). Stefan Müller is Assistant Professor and Ad Astra Fellow, University College Dublin (stefan.mueller@ucd.ie).

Acknowledgments: Thanks to Helen Retzlaff, Molly McClure, and Caitlin Sharma for excellent research assistance. We also thank Friederike Nagel, Marcus Maurer, and Carsten Reinemann for sharing the content analysis, surveys and RTR data from the 2005 debate, and the team of the German Longitudinal Election Study for making the data for the debates in 2009, 2013, and 2017 publicly available. This research was funded through generous support from the Trinity College Dublin Arts and Social Sciences Benefactions Fund 2019–20. Paper previously presented at the 2020 European Political Science Association virtual conference, the UCL Political Science Departmental Research Seminar, and the Digital Democracy Workshop at the University of Zurich.

1 Introduction

In forming attitudes about political leaders, voters evaluate not just what leaders say, but how they say it. Nonverbal displays, including facial emotional expressions, voice pitch, and general the sentiment of speech all provide key pieces of information for voters about the suitability of individuals for leadership positions (Boussalis and Coan 2020, Carpinella and Bauer 2019, Sülflow and Maurer 2019). One place where these expressions are particularly important is in political debates. Debates between political leaders seeking support from voters are a central component to candidate selection in many democratic systems. Through analyses of these debates, scholars have transformed our understanding of the role of images and emotions in political contexts (Druckman 2003, Ridout and Searles 2011, Nagel et al. 2012, Boydstun et al. 2014, Fridkin et al. 2019). Yet, to date, this research largely focuses on single debates or single electoral cycles and laborious methodological techniques that rely on hand-coding images of candidates. Innovations in multimodal data collection and analysis offer new opportunities to study how candidates communicate and how voters respond to this communication in real time (Bakker et al. 2020, Masch 2020, Dietrich et al. 2019, Carpinella 2016).

This article posits that emotions matter in politics, leaders strategically express emotions via nonverbal cues in debates, and voters make judgements about leaders from the nonverbal aspects of political speech. Critically, we argue that not all candidates are equally able to use the full range of nonverbal cues because voters apply differing expectations based on the socially-meaningful identities of candidates. Gender is one such identity (Bauer and Carpinella 2018, Bauer 2019, Masch 2020). Applying gender role theory (Schneider and Bos 2019, Oliver and Conroy 2020) and research on emotions in nonverbal communications, we argue that men and women running for political office will attempt to strategically employ (Ridout and Searles 2011) specific emotions that are associated with political power (Carpinella and Johnson 2013, Carpinella et al.

2016, Everitt et al. 2016). Voters, moreover, will respond to these displays, supporting candidates who engage in gender- and role-congruent emotional expression. Given the social expectation that women should be communal and caring (and not agentic and aggressive) (Cassese and Holman 2018), we argue that voters will reward women seeking political office who are able to manage their expressions of anger and increase expressions of happiness.

We develop and test our expectations using four debates that feature Angela Merkel—arguably the world's most powerful woman—versus her male opponents.[1] The televised leader's debates are by far the most important event during an election campaign in Germany, with one-in-five people watching the debates. Restrictions on campaign finances mean that there are few ads on television, accelerating the importance of nonverbal expression in these debates (Nagel et al. 2012). By examining four debates over time, our approach provides the opportunity to understand candidate and voter behavior in a broad setting. In the first debate in 2005, Merkel competed for the Christian Democratic Union (CDU) against chancellor Gerhard Schröder from the Social Democratic Party (SPD). In the 2009, 2013, and 2017 debates, Merkel (the incumbent chancellor) debated three men representing SPD. We argue that Merkel's performance in these debates (and voters reactions) represents a critical 'least likely' case, where we expect to see very few gender differences across the candidates. We then extend our research to the 2017 minor debate in Germany, which featured (for the first time) two women candidates.

To assess our expectations about nonverbal communication and voter response, we combine computer coding of images and sound from 596,000+ frames across the four debates. We extract expressions of anger and happiness and overall levels of facial emotive engagement, using tools from computer vision (Torres and Cantú 2020). We combine this with measures of emotional intensity from vocal pitch (measured by fun-

[1]For example, in the Forbes list of The World's 100 Most Powerful Women, Angela Merkel took the No. 1 spot for nine consecutive years (2011–2019).

damental frequency) and the sentiment of words spoken via text analysis (Schoonvelde et al. 2019). Our research builds on the growing body of political science research that studies these nonverbal cues (Dietrich et al. 2019, Torres 2018). Our research suggests that Merkel expresses less anger than her male opponents and is less emotive (measured through vocal pitch) than her opponents.

Having shown that candidates emote in specific ways, we examine how voters respond to women's and men's emotions. We combine second-by-second emotions from the candidates with real-time responses (RTR) (Boydstun et al. 2014) from representative samples of voters watching the debate and providing continuous evaluations. Using these data, we are able to evaluate how voters respond in the moment that a candidate expresses emotion. Consistent with our expectations, we find that viewers reward Merkel, compared to her opponents, for expressing happiness and punish her for expressing anger. To examine the carrying capacity of these findings beyond Merkel, we also replicate this analysis for a debate between candidates from five smaller German parties. In 2017, two female and three male candidates participated in this debate. The results largely correspond to our analysis of the debates between Merkel and her male competitors, highlighting the role that gender plays in candidate behavior and voters' assessments of politicians.

We argue that combining computer vision, machine learning, and text analyses allows one to gain a more complete understanding of candidate behavior and voter decision-making. Candidates are fundamentally interested in presenting their best self to the public (Bystrom et al. 2005, Dittmar 2015). By capturing not just what candidates say, but how they say it and what they look like when they say it, we offer a comprehensive evaluation of candidate self-presentation. Moreover, the ability to leverage continuous responses from voters in a live audience offers an additional advantage for understanding political behavior. Taken together, the German case, in combination with very fine-grained data on multiple modes of communication and

voter reactions, provide a new and unique view of the role of emotions in politics.

2 Nonverbal and Emotional Communication in Politics

Political leaders seek to garner favor among voters through their words, voices, and facial expression; these 'hearts and minds' appeals shape voter evaluations (Carpinella and Johnson 2013, Carpinella et al. 2016, Everitt et al. 2016, Fridkin et al. 2019). Visual and nonverbal forms of communication like gestures, postures, and vocal pitch provide key information about candidates, including their electoral suitability (Bucy and Grabe 2007, Bucy and Stewart 2018). While voters broadly use personal characteristics like perceived competence as a tool in selecting candidates, images and videos accelerate this focus on personal qualities even more as those watching focus on mannerisms and infer personality characteristics (Druckman 2005, Brydon et al. 1992).

Nonverbal communications, including facial displays and vocal pitch, are a key mechanism by which candidate convey emotions and voters assess the emotional acceptability of candidates. For example, citizens infer candidate traits like competence and trustworthiness from vocal pitch (Klofstad et al. 2015, Anderson and Klofstad 2012) and the combination of "competent faces" and "competent voices" shapes vote outcomes of experimental elections (Klofstad 2015). Candidates need to fit a set of 'Goldilocks' expectations in both the overall level of their emotional expression and which specific emotions they express. In short, candidates do not want to appear as too emotional, but they also do not want to be perceived as apathetic. Candidates must also express emotions that are *congruent with the role* they seek. The acceptability of both the overall level of emotions and specific emotions is deeply rooted in evolutionary biology, where humans interpret facial displays of emotions as "ritualized signals" that dictate and maintain relationships (Eibl-Eibesfeldt 1979). The human desire to select

leaders who can "dominate others, and thus show how he or she is able to neutralize external as well as internal threats to the group" means prioritizing candidates who express anger and other agonistic emotions (Boussalis and Coan 2020 7). Yet, the appearance of domination also needs to be controlled and situationally appropriate, as voters shy away from leaders who would exert too much control over the group.[2] Thus, people want leaders to express happiness and hedonic emotions, which represent the ability to affiliate with others. These emotional expressions include facial cues, vocal pitch, and the sentiment of language.

2.1 Gender, Emotional Expression, and Voter Reactions

Not all individuals seeking leadership positions are equally able to leverage emotional expressions to gain support because voters do not respond to every candidate's behavior in the same way. Indeed, "political candidates differ widely in the effectiveness of their nonverbal behavior" (Grabe and Bucy 2009 148). These divergent reactions can be because of charisma, attractiveness, political party, age, and, importantly for us, gender.

Gender shapes which emotions people express, the levels of those emotions, and how others react to those expressions (Meeks 2012, Bauer and Carpinella 2018, Masch 2020). But gender does not just shape the emotional expression and reactions in the general population. Gender functions in the "processes, practices, images and ideologies, and distribution of power" in society and especially in politics (Acker 1992 567). We use gender role theory (Eagly and Karau 2002) to guide our understanding of candidate behavior and voter reactions. Gender role theory posits that men and women are socialized into particular roles in society. Women are expected to hold communal

[2]Individuals seeking political office are well aware of the role congruity expectations that voters have, and try to express appropriate emotions (Boussalis and Coan 2020, Dittmar 2015). For example, it is quite clear that voters would respond negatively to a candidate expressing fear, nervousness, or evasion (Bucy and Grabe 2008). Evaluations of nonverbal displays in political debates in the United States show that candidates know this and rarely display fear or evasion (Bucy and Grabe 2008, Boussalis and Coan 2020).

characteristics, including being "affectionate, helpful, kind, sympathetic, interpersonally sensitive, nurturant, and gentle" (Eagly and Karau 2002 574). In comparison, men are expected to present with agentic traits, which include being decisive, assertive, and strong leaders (Eagly and Johannesen-Schmidt 2001). These gendered expectations constrain both verbal and nonverbal behavior (Everitt et al. 2016, Bucy and Grabe 2007).

Gender role socialization leads to gender differences in the *type* of emotional expression that individuals engage in as well as the overall *level* of emotions. Women are socialized to feel and express a greater intensity of emotions overall (Kring and Gordon 1998) and the emotions like happiness that facilitate communal skills (Brody 2009). [3] Men, alternatively, are socialized to express less emotions generally, but when they do express emotions, they are consistent with the male gender roles of assertiveness and leadership, such as anger (Schneider and Bos 2019).

These gender roles produce congruency expectations, such that women are expected to act "like women" and men are expected to act "like men" (Eagly and Karau 2001, Schneider and Bos 2019). If individuals engage in gender congruent behavior, they receive internal and external rewards; similarly, gender incongruent behavior is punished (Eagly and Karau 2001, Bauer 2017, Cassese and Holman 2018). These expectations spill over to emotional and nonverbal behavior, where people believe women to be more emotional generally and to express a broader range of emotions, except anger and pride (Plant et al. 2000). As such, a woman can be punished for expressing anger and rewarded for happiness and sadness, while a man may experience the opposite (Meeks 2012, Fischbach et al. 2015, Cassese and Holman 2018).

There thus emerges a challenge for women seeking leadership roles: because of leadership role expectations, voters want leaders who express anger and happiness.

[3] While people generally think that women are more emotionally expressive than are men, daily diaries suggest that men and women actually feel the same types and levels of emotion (Van Boven and Robinson 2012).

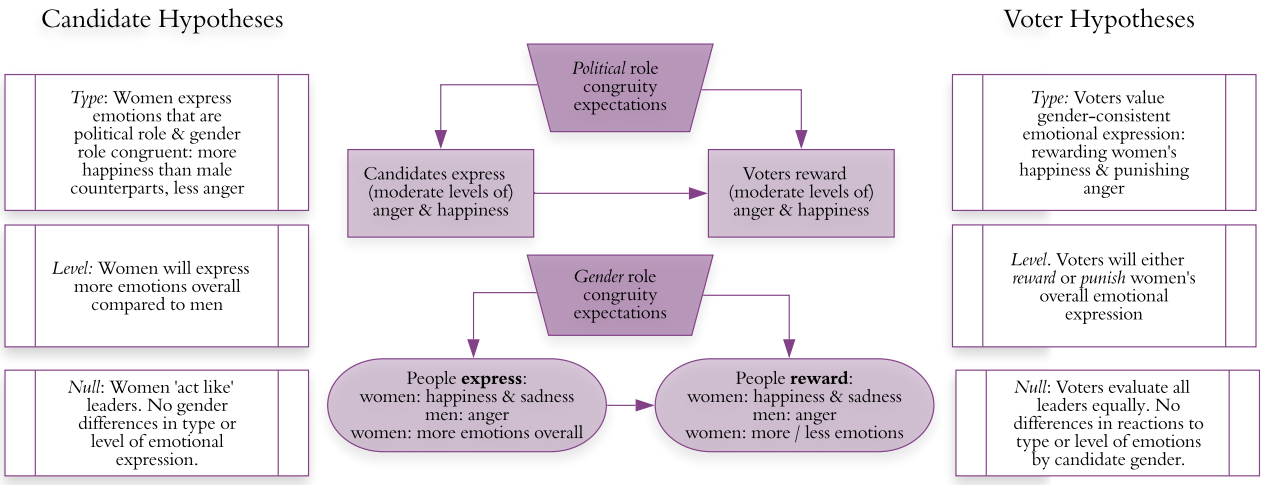


Figure 1: Theoretical expectations.

But gender role expectations mean that women should express happiness and sadness. Women seeking political office are highly aware of the potential of gendered expectations about their behavior from voters (Dittmar 2015, Lazarus and Steigerwalt 2018). The natural solution, then, for women and men seeking positions of power, is to express the emotions that are both political role and gender role consistent, such that:

> **Candidate-H1**: *Women seeking office will express more happiness than will men and men will express more anger than will women*

As we previously noted, voters want leaders who express role congruent emotions (Klofstad et al. 2015). But voters also apply varying standards to how women and men in public office look and sound (Bauer 2018, Bauer and Carpinella 2018, Carpinella and Bauer 2019) and may want women and men who express gender role congruent emotions (Fischbach et al. 2015). In Germany, Masch and colleagues find voters react positively when leaders express happiness (Gabriel and Masch 2017, Masch 2020). Research also suggests that voters are particularly unlikely to accept masculine behavior from women. For example, research on nonverbal displays and gender find that voters

do not react to men's agentic nonverbal displays, but see women as less likeable when they engage in displays of dominance (Copeland et al. 1995, Everitt et al. 2016). If voters want gender- and leader-consistent emotional expression, we would expect that:

> **Voter-H1**: *Voters will reward women's happiness and punish their anger, relatively to men's expression of happiness and anger.*

It is also possible that voters evaluate men and women by not just the specific emotions that they express but also by their overall level of emotional expression. Recall that gender role socialization suggests that women are granted a broader leeway for general emotional expression and are assumed to feel and express a broader set of emotions (Plant et al. 2000). Thus, if men and women in political office behave in a gender-role congruent manner, we would expect:

> **Candidate-H2**: *Women will express more emotions overall compared to men.*

If voters want leaders who conform to gender roles, they may reward women's higher levels of emotional expression, even in political settings where emotions are expected to be controlled (Gleason 2020, Masch 2020). Yet, while people generally believe that women express more emotions than do men (Durik et al. 2006), women in leadership roles can be punished for expressing excess emotions (Bauer 2015, Heilman et al. 1995) and people generally expect men in leadership roles to be more successful at managing their emotional expression (Fischbach et al. 2015). This also applies to debate performances: in their analysis of the fourth Republican primary debate in the 2016 US Presidential election, (Boussalis and Coan 2020) found that the sole woman candidate on stage was penalized by viewers when expressing any emotion through facial displays. These findings lead to mixed expectations:

> **Voter-H2a**: *Voters will react* ***positively*** *to any emotional expression by women compared to men.*

Voter-H2b: *Voters will react* ***negatively*** *to any emotional expression by women compared to men.*

While gender role expectations may shape the behavior of men and women in the general population, it is possible that political leaders are a different emotional animal. Political leaders learn how to manage their self-presentation during their time in office. Research suggests women are particularly cognizant of the need to adapt to voter expectations (Dittmar et al. 2018). From developing expertise in areas where voters might perceive a weakness (Swers 2013) to controlling their appearance (Dittmar 2015), to responding at higher levels to constituent concerns (Bauer 2020), women work to limit the potential effect of gender biases in elections. Scholars have documented these efforts in the types of campaign advertisements that candidates develop (Bauer and Carpinella 2018, Bauer 2018; 2020), their verbal behavior (Dittmar 2015, Dittmar et al. 2018), and even their choice of where and when to run for political office (Ondercin 2020, Silva and Skulley 2019). Differences may be particularly hard to identify among men and women seeking high level political office, as candidates who have reached this stage are well-trained and experienced with controlling their emotions in public.

Voters may also see women in political office as "leaders not ladies" (Brooks 2013) and evaluate their behavior within the lens of acceptable actions from politician. Consistent with the idea that the public has less well formed views of women in political office than they do of women in the general population (after all, the public often does not encounter many women in political office!) (Schneider and Bos 2014). As such, we may find no differences between men and women's emotional expression and voters will not react to the type of overall level of emotional expression by candidate gender.

3 Political Debates and the German Case

We test our expectations using a novel set of data across four national debates in Germany: 2005, 2009, 2013, and 2017. These debates provide an opportunity to understand within-case variation over time and evaluate the behavior and voter responses to the most powerful woman in the world: Angela Merkel. These debates allow us to examine candidates' nonverbal behavior and voter reactions within a context where a) debates provide political information to a broad set of voters; b) leaders' performances in the debate are important components of the political selection process; and c) voters and media coverage often focus on the importance of emotional displays by candidates.

Political debates offer an opportunity for voters to assess not just how candidates present themselves in isolation, but also how to compare directly to each other. Studies of debates demonstrate that voters obtain information about candidate traits, policy platforms, and electability from on-stage exchanges, and debate performance can ultimately influence vote choice (Lanoue and Schrott 1989, Benoit et al. 2003). From examining the importance of television over radio in the Kennedy-Nixon debates (Druckman 2005) to showing how Trump and Clinton's on-stage interactions shaped how men and women evaluated 2016 debates (Fridkin et al. 2019), scholars demonstrate that *seeing and hearing* debates shifts how people view the participants. In debates, viewers focus on person traits, emotions, and electoral suitability and other information conveyed by nonverbal cues over issue topics and substance.

Our approach also provides an opportunity that is rarely available to researchers: we can compare emotional expression *and* voter reactions across time while holding the institutional setting constant. While the issues and political contexts of each debate shift, much does *not* vary across the years (see SI Section A for a more detailed discussion). In many ways, these data are an embarrassment of riches: few scholars have access to multiple iterations of debates that hold the setting constant, nor is it common to have voter reactions, obtained through a consistent method, across multiple years

of debates. That Angela Merkel appears in each of the major debate is an additional benefit, as we can compare her behavior over time. The supplement of the minor debate, which features two other women candidates, provides us with an opportunity to examine the ways that gender shapes emotional expression and voter reactions.

Leadership debates are important events in German politics (Maier and Faas 2011). The televised leader's debates are the most important event during an election campaign, with more than 20 percent of the German electorate watching each debate. The structure of the finances of German election campaigns further elevate the debates, particularly for assessing nonverbal expression. Candidates and parties have few opportunities to present themselves to the public. Parties have strict spending limits, can only air few ads on TV, and mainly rely on posters, face-to-face campaigning, advertisements in the media, and (increasingly) social media. The TV debate is the only opportunity to address a very large share of the electorate. Emotional displays during these 90 minutes could potentially convince or deter voters (Maier and Faas 2019). While previous findings underscore that emotions occur in German debates and talk shows (e.g., Maier and Jansen 2017, Masch 2020), we lack a comprehensive assessment of how visual displays of emotions influences candidate assessment, and whether candidates 'learn' to control their emotions with more experience in office.

To test the set of candidate- and voter-specific hypotheses outlined in Section 2, our study focuses on four televised leaders' debates between the candidates of the two largest parties in Germany between 2005 and 2017.[4] Angela Merkel participated in all of these four debates. In 2005, she competed against the incumbent chancellor Gerhard Schröder. After the election in 2005, Merkel led a grand coalition between the Christian Democrats (CDU/CSU) and the Social Democrats (SPD). In the three subsequent debates Merkel was the incumbent chancellor and faced three male candidates from the SPD: Frank-Walter Steinmeier, Peer Steinbrück, and Martin Schulz. We describe

[4]There are no data on reactions by the audience for the first televised debates between Gerhard Schröder and Edmund Stoiber in 2002.

the context of the four elections and the perceptions of the candidates' performances during the debates in Appendix A.

Since 2002, candidates from the smaller German parties also compete in a 'minor' debate, which usually takes place shortly after the main debate. In 2017, for the first time, these minor debates featured women candidates, allowing us to assess candidate expression and respondents' reactions to facial, verbal, and vocal emotions by men *and* women. The TV debate involved the candidates of the five smaller parties with a promising chance of entering the German Bundestag. Sahra Wagenknecht, the candidate of the ideologically most 'left' party (Die Linke) and Alice Weidel, the candidate of the right-wing populist Alternative for Germany (AfD) competed against three male competitors from the Green Party (Cem Özdemir), Christian Lindner (FDP), and Joachim Herrmann (CSU). [5] The minor debate provides an opportunity to assess if the same patterns of expression are found across women running for office and whether gender shapes voter reactions in the same way.

4 Data

We employ a set of automated methods to extract granular visual, vocal, and verbal information of debate participants and combine these data with second-by-second real-time response measurements from focus group subjects who watched the debates live. This section describes in detail the steps taken to measure these multimodal candidate signals. We also discuss other candidate- and voter-level data which are used as controls in the analysis. Figure 2 shows the variation of the four most important variables across each debate. The x-axis shows the time in each debate (each debate lasts 1.5 hours). The y-axis shows the values of emotions (row 1), voice pitch (row 3), sentiment (row 3) and candidate evaluations (row 4); we describe each of these measures below.

[5] The Free Democratic Party (FDP) and the Christian Social Union (CSU) are centre-right parties; CSU operates as the Bavarian (regional) counterpart to the CDU.

4.1 Automated Classification of Candidate Facial Displays of Emotion

We build upon burgeoning scholarship that uses computational methods to study images-as-data (Joo and Steinert-Threlkeld 2018, Anastasopoulos et al. 2016, Torres 2018, Cantú 2019, Casas and Williams 2019, Zhang and Pan 2019), and in particular to capture and analyze facial expressions of political actors (e.g., Joo et al. 2019, Boussalis and Coan 2020). While there is a strong interest in the nonverbal communication literature to increase the granularity of facial measurements (Bucy and Stewart 2018), the field continues to be hampered by the methodological challenges involved with manually content analyzing images of faces at large scales, i.e., every frame of a set of hours-long debate videos. It takes an average of 10 minutes to apply the widely used Facial Action Coding System (FACS) (Ekman and Friesen 2003) to identify the emotional expression from a face in an image (Stewart et al. 2011). At this rate, a debate with more than 100,000 images could take over 160,000 hours to code. Given that our study seeks to classify candidate facial displays of emotion at each frame of four debates (more than 500,000 frames), the time and resource costs needed to manually approach this measurement task exceed prohibitive levels.

Luckily, innovations from the fields of machine learning and computer vision allow us to to extract these facial nonverbal signals in a much quicker and more systematic process. To do so, we follow this protocol: we downloaded the debate videos from either YouTube (2009, 2013, 2017) or C-SPAN (2005) and extracted their frames ($n = 595,169$).[6] After obtaining the frame-level images, we relied on the Face API from Microsoft Azure Cognitive Services to identify the faces in each frame and to extract the emotive display of each face. The Face API recognizes human faces and predicts the level of eight emotions (anger, contempt, disgust, fear, happiness, neutral, sadness

[6]The total runtime of the debate videos were as follows: 2005 (01:34:48 @ 30 fps), 2009 (01:32:16 @ 25 fps), 2013 (01:33:18 @ 25 fps), 2017 (01:37:33 @ 25 fps). We downloaded the 2009 debate in 10 parts from YouTube and stitched it together prior to analysis.

and surprise). This software relies on deep convolutional neural network architectures (LeCun et al. 1998; 2015, Krizhevsky et al. 2012) and has been trained largely on the Ekman and Friesen (2003) model of discrete facial expressions (Bargal et al. 2016). After passing an image to the Face API, the service returns the identity of each face and a confidence score of the eight emotions mentioned above, ranging over the interval $[0, 1]$, with all emotion confidence scores for a given image summing to one.[7] We collapse the frame data to the second-by-second level for each debate. This resulted in average per second facial emotion confidence scores.

The first row in Figure 2 shows the non-neutral emotional displays for both candidates (Merkel: black, solid line; her opponents: red, dotted line), measured through our automated classification. The descriptive overview of facial expressions suggests that that candidates express high levels of emotion at the beginning and end of debates with more variation in emotions from Merkel's opponents than from her. Consistent with the theoretical expectations described in Section 2, we focus our analysis on facial displays of *any emotion* (as depicted in Figure 2), as well as displays of either *happiness* or *anger* (depicted in 3). For an extensive validation of the Face API for each these key measures, see Section B of the appendix, where we compare the manual coding of smiles in the 2005 debate (Nagel et al. 2012, Sülflow and Maurer 2019) with our automated measure. The very high degree of correspondence between both measures strongly suggests that we pick up the most frequent emotional display very reliably.

4.2 Measuring Emotional Intensity via Candidate Vocal Pitch

We next capture the emotional content of a candidate's vocal characteristics. Following the work of Dietrich et al. (2019), we operationalize emotional intensity by measuring

[7]The face recognition model used by the Face API relies on user-provided images of persons as input to the model. We uploaded 9 to 15 images of the four political candidates (Angela Merkel, Gerhard Schröder, Frank-Walter Steinmeier, Peer Steinbrück, and Martin Schulz) and 18 journalists who fielded questions to the candidates over the four debates. The German debates occur without a live audience, so there was no need to account for faces in the background.

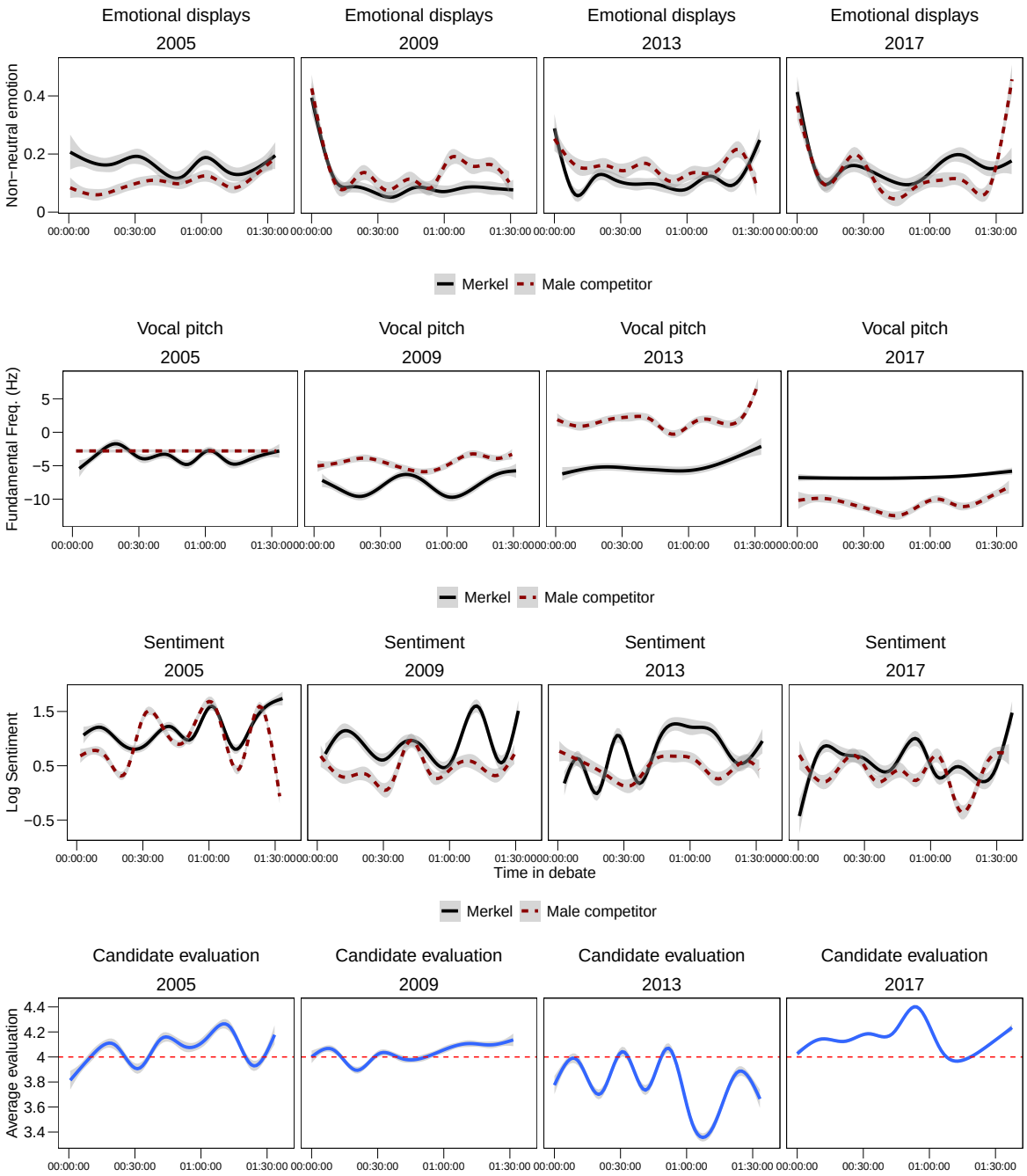


Figure 2: Variation and developments of the measures of interest across the four debates. Lines are generalized additive models with integrated smoothness. In rows 1-3, black lines show the smoothed lines for Merkel, red dotted lines display the values for her male opponent. In row 4, the candidate evaluations range from 1 to 7 where higher values imply more support for Merkel/disapproval of her opponent and 4 (red, dotted line) a neutral evaluation.

the fundamental frequency (F0) of the voice of a candidate while speaking during a debate. We extracted the audio from the debate videos using `ffmpeg` and then passed the files to the `parselmouth` library in Python (Jadoul et al. 2018) which builds directly upon the source code of `Praat` (Boersma and Weenink 2018). Specifically, the audio from each debate video was extracted and written as a WAV file, which was then passed to `parselmouth` and converted into a `Praat` sound object. This sound object contains 100 "frames" per second, each of which includes at least one "candidate" estimate of F0. That is, given that pitch estimation varies in accuracy, `Praat` provides a varying number of "candidate" pitch estimates, which are ranked from best to worst. We kept the highest ranked candidate for each "frame" and then compute the average F0 for each second of a given debate. We, therefore, measure the *average per second fundamental frequency* of the debate audio. The second row of Figure 2 reports the voice pitch, measured as the fundamental frequency standardized for each candidate. We see variation in vocal pitch across the candidates and debates, with Merkel's vocal pitch becoming more controlled with time.

4.3 Sentiment of Candidate Utterances

We measure statement-level sentiment with a dictionary approach. We use the German translation of the Lexicoder Sentiment Dictionary, which has recently been validated extensively for political speech (Proksch et al. 2019). The dictionary consists of 3,998 positive and 5,849 negative terms.[8] We identified the words spoken by each politician and passed them through the same sentiment dictionary using the `quanteda` R package (Benoit et al. 2018). Afterwards, we count the number of positive and negative words in each statement by a politician or moderator and apply the aggregation formula recommended by Proksch et al. (2019), which estimates sentiment as the logged ratio

[8]We have statement-level debate transcripts for the debates in 2009, 2013, and 2017. The 2005 debate comes from a different source (Nagel et al. 2012) and do not contain the debate transcripts. To create the 2005 transcript, we automatically transcribed the audio using Google's transcription service, which creates timestamps at the word level.

of the sum of positive ($\sum Pos$) and negative terms ($\sum Neg$):

$$Sentiment = log\left(\frac{\sum Pos + 0.5}{\sum Neg + 0.5}\right) \quad (1)$$

A value of 0 indicates that a document contains the same number of positive and negative terms (or does not contain any of the terms included in the dictionary), a value above 0 implies a larger number of positive words, relative to the sum of negative words. This score is generated for each debate participant. The third row of Figure 2 reports the statement-level sentiment, measured through the transcripts of the debates. Three patterns stand out from the sentiment: all candidates shift their sentiment repeatedly during the debate, the general tone of these debates is positive, and sentiment appears to track somewhat with facial emotions (row 1) and voter reactions (row 4).

4.4 Real-time Reactions of Debate Audience Members

Our study relies on continuous response measures of debate audience members to observe how voters react to candidates' visual, vocal, and verbal signals in real-time. For the debate in 2005, we use real-time response (RTR) data from Nagel et al. (2012).[9] RTR data on the debates from 2009, 2013 and 2017 are included in the German Longitudinal Election Study (Rattinger et al. 2010; 2011a;b; 2014; 2015; 2018, Roßteutscher et al. 2019a;b;c). All respondents are eligible voters and were recruited by press releases, leaflets and posters advertising participation in a study on media reception based on a quota plan drawn up in advance. The number of respondents ranges from 46 (2017) to 154 (2009), with an average of 90 respondents across the four debates. The unit of analysis is the respondent-second level, resulting in a range of 169,510 (2017) and 571,648 (2009) observations per debate.

To test the hypotheses relating to voter reactions to emotional displays, we next

[9]The authors of this study generously shared all their data, extensive coding, and design information of the first RTR study in Germany.

construct a dataset of the *real-time response measures* at the individual respondent-second level, meaning that the unit of analysis is the evaluation of candidates in a given second by a respondent.[10] The scale of this measure ranges from 1 to 7. Participants were asked to move the dial to the left (values 1 to 3) if they had a good (bad) impression of the challenger (current chancellor). The stronger this impression was, the further the knob should be turned. If a person had a good (bad) impression of the chancellor (other candidate), they were to move the dial to 5 to 7. The scale value 4 implies a neutral impression or that positive and negative impressions of both candidates cancelled each other out. To compare across the debates, we inverted the values of the measure for observations where the challenger is speaking—that is, higher values of the re-coded variable indicate more agreement with the current speaker.

The final row reports of Figure 2 shows a GAM smoother of the average evaluation of the candidates for each second of the debate. The scale ranges from 1 to 7. Values exceeding 4 (red dotted horizontal line) imply (on average across respondents) a more positive evaluation of Merkel. Across the four debates, on the aggregate level, neither candidate appears to have a clear advantage in approval.

4.5 Other Candidate and Voter Data

We combine the response data with *individual-level data on each respondent* based on a survey conducted prior to each debate. These variables include the age, gender, party identification, self-reported political interest, and political knowledge.[11]

Finally, we merge in a *manual content analysis* (provided by Nagel et al. (2012)

[10] Two different approaches are available in the dataset: in 2013, some respondents evaluated the candidates through a push-button system and in 2017, some respondents were asked to turn the knob to the right if they really like something during the debate. The rule was that whenever a respondent had a good impression and no matter what, turn the knob to the right. The better your impression, the further you turn to the right. We limit the analysis to respondents who used a "dial" button with standard instructions to evaluate candidates as this process was consistent across the debates.

[11] We measure political knowledge by re-coding the answers to three to four (depending on the study) factual questions on political developments and the economic situations. The scale of knowledge ranges from 0 (no correct response) to 3 or 4 (all questions answered correctly).

and the GLES team) of each second of the debate, which includes an indicator of who is speaking at a given second (or if no one is speaking), the text of their speech, and the issue substance of the speech. We generated 14 substantive policy areas (such as taxes, education, or foreign policy) that were present across the four debates. We assigned one of topics to each segment of speech. Drawing on scholarship on gender issue stereotypes (Bauer and Carpinella 2018, Bauer 2018, Cassese and Holman 2018), we also classify these topics into "feminine", "masculine", and "neutral" topics. Table A2 provides an overview of the policy areas and the coding of gendered policy areas.

5 Statistical Methods and Measures

5.1 Candidate-level Methods

We examine the candidate-level hypotheses described in Section 2.1 using a number of different measures. First, when considering facial displays of emotion, we conduct an analysis at the *second-by-second level*, combining data for all four political debates. As described above, we have expectations both for the extent to which Merkel's facial displays express any emotion relative to her male counterpart, as well as for the specific emotions (especially anger) expressed. Specifically, we calculate the per second average confidence score estimated by the Microsoft Face API for anger, happiness, and non-neutral displays by taking the mean value of the frame-level confidence scores for these emotions within a given second. We use these average confidence scores as as dependent variables to explain the variation in the emotional displays of candidates.

Next, in order to examine candidate-level expressions of emotional intensity, we focus on second-by-second measures of vocal pitch. Specifically, we gauge emotional intensity in a candidate's pitch as binary measure for whether, in a given second, the vocal pitch is 2 standard deviations above the candidate's mean pitch. Merkel and her competitors' pitch is standardized within each debate.

The statistical analysis of the candidates' facial displays of emotion rely on random-effects panel data regression models with AR(1) disturbances, while also including a lagged dependent variable. The analysis of vocal pitch is carried out with a panel data probit regression model which includes Huber-White standard errors. All models include a fixed effect for the debate year, the sentiment of a candidate's utterance (see Section 4.3, and whether the topic under discussion is considered feminine, masculine, or neutral (see Table A2). Reference categories in all models are the 2005 debate and "neutral" gender topic.

5.2 Voter-level Methods

To examine our voter-level hypotheses, we draw on the RTR data described in Section 4.4. Past scholarship highlights a number of challenges associated with determining a suitable estimation strategy for studies using RTR data (Schill et al. 2016). One immediate challenge is that the relationship between candidate behavior (e.g., facial expressions, pitch, etc.) and participant response is inherently dynamic, and the lag time between an expression and response is not known in advance. To capture these dynamics when estimating the influence of a candidate's emotional expressions, we build on previous approaches (Boussalis and Coan 2020). Based on information criteria, we determine that roughly 4 seconds suitably captures the dynamics of our key facial, vocal and verbal measures, which is largely consistent with past scholarship (Nagel et al. 2012, Boussalis and Coan 2020). While it is standard practice to place constraints on the lag structure in autoregressive distributed lag models (ADLs) to avoid multicollinearity issues—particularly when using small to medium sized datasets—we leverage a massive sample size to estimate the lag structure directly by including four lags of these key variables. In doing so, we offer a flexible parameterization of the salient dynamics without making—perhaps inappropriate—assumptions on the underlying lag distribution.

We employ an ordinary least squares regression model to test the voter-level hypotheses, with the 7-point dial score as the dependent variable. The main explanatory variables are a binary variable of whether Angela Merkel (1) or her opponent (0) is the speaker, and the standardized per second average confidence scores of facial displays of emotion across four lags. These models also include as controls for respondent gender, political identification, political knowledge, political interest, as well as the narrow gendered topic (see Section 4.5) being discussed at any given second. We cluster standard errors at the participant level.

To evaluate how voters react to Merkel's emotions relative to her male opponents, we estimate a single model for each debate, a departure from past scholarship (see Nagel et al. 2012, Boussalis and Coan 2020). Given how individual responses are encoded in our data (i.e., higher values mean greater support for a candidate *when they are speaking*), we estimate a fully conditional model, interacting whether Merkel is the speaker with all covariates in the model. This approach allows us to estimate our main comparison of interest and ensure that key control variables have a substantively meaningful interpretation.

6 Results

We first present descriptive data and analysis of gender and nonverbal cues, testing our candidate-side hypotheses. Here, we describe the prevalence of different emotions and predict Merkel's and her opponents' facial displays, and vocal pitch. Then we turn to the voter-side hypotheses. In doing so, we analyze whether and what types of multimodal expressions change real-time responses by the live audiences. We conclude with the analysis of the 2017 debate between candidates of the five smaller parties.

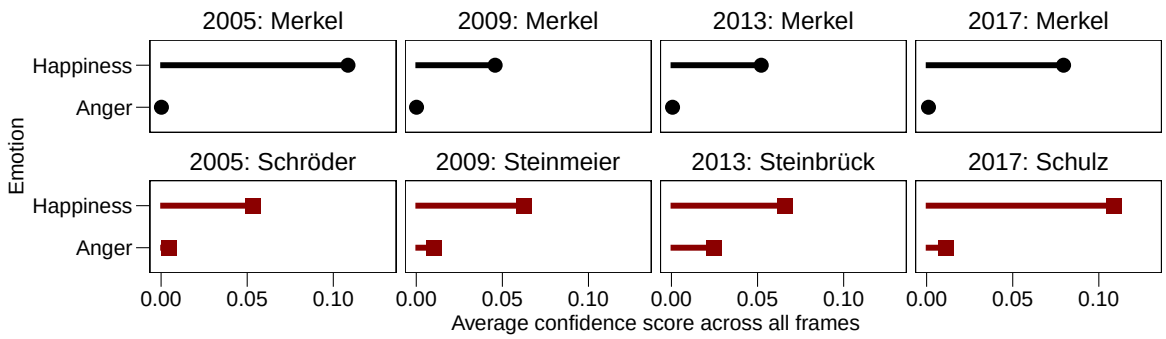


Figure 3: Average confidence scores for emotional displays aggregated for all frames per speaker and debate.

6.1 Gender and Nonverbal Cues

We first examine the nonverbal emotional expression from the candidates in the main debates. Figure 3 plots the average confidence scores for visual emotional displays across all frames for a given speaker in a debate. The plot shows which emotional displays are more or less frequently expressed by the candidates.

First, all candidates display a high level of happiness in the debates. All three men (Steinmeier, Steinbrück and Schulz) display more anger than Merkel, with values ranging between 0.01 and 0.03. The descriptive findings are consistent with our candidate hypothesis 1: that men will express more anger, but Merkel only expresses more happiness than her male opponent in 2005.

We next test our expectations about the *type* of emotions (Candidate Hypothesis 1) using per-second averages of confidence scores of anger and happiness in Models 1 and 2 of Figure 4 and the *level* of emotions (Candidate Hypothesis 2) using non-neutral facial displays and vocal pitch, reported in Models 3 and 4. Note that the x-axes vary to better display the coefficients of each model.

Models 1 and 2 test our expectation about specific emotions: that female candidates are less likely to express anger in televised debates, but would also express more happiness, an emotion that is both congruent with women's gender role expectations and acceptable for political leaders. We find mixed evidence in support of our expec-

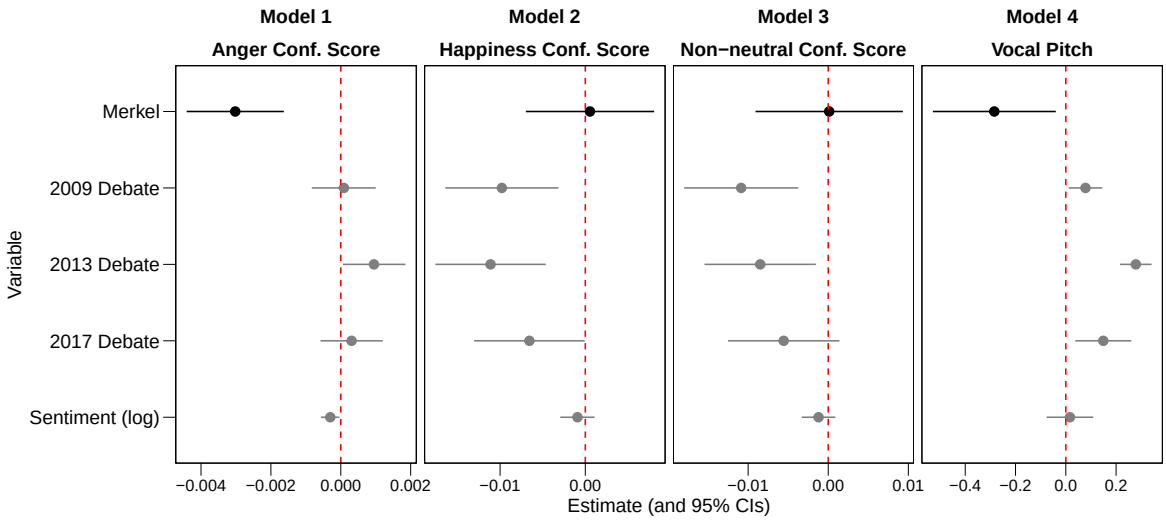


Figure 4: Random-effects panel data linear regression (Models 1–3) and random-effects panel data probit regression (Model 4) results of per-second average confidence scores of anger, happiness and non-neutral facial displays, and per-second candidate heightened vocal pitch. All models include the lagged dependent variable and controls for masculine, feminine, and "none" debate topics. Reference categories in all models are the 2005 debate and neutral debate topic. The x-axes are re-scaled for each model. Coefficients are displayed in Table A1.

tations: the results suggest that *ceteris paribus*, Merkel is less likely than her male counterparts to express anger (1% error level). With controls, we still do not find evidence that Merkel expresses more happiness: there is no a statistically significant difference in average happiness displays between Merkel and her opponents (Model 2).

We next estimate the propensity to emote more generally by candidate gender, using non-neutral facial displays and much higher than average per-second vocal pitch. As show in the third model in Figure 4, we find results opposite to our expectations: Merkel is just as likely as her opponents to express nonverbal emotional cues. In the final model presented in Figure 4, however, the results suggest a more nuanced picture: Merkel is less likely to express emotional intensity through increased vocal pitch (5% error level), which is the opposite directly from our expectations.

6.2 Voter Responses to Candidate Emotions

Do these differences in emotional expression matter for how voters perceive the candidates? To assess our expectations, we turn to the real-time response data and examine the positive and negative responses from the voters to these nonverbal displays from candidates. To refresh, our dependent variable is the reaction (on a 7-point scale) to the candidate that is shown on the screen with a four-second lag. We estimate a separate model for each debate. Like the candidate analyses, we control for the topic of the debate and we also control for respondent gender, political knowledge, and political party affiliation. Given that we are principally interested in the difference in reactions to Merkel's emotions as compared to her opponent's emotions, we present the effect of a 1 standard deviation increase in the nonverbal display of the emotion between Merkel and her opponent.

We find some evidence of our first expectation for voter reactions: voters punish Merkel for her expression of anger and reward her expression of happiness. Moving from the top down in Figure 5, we see negative coefficients for Merkel's anger in two of the four debates. The 2013 response is particularly interesting, given that the Eurozone crisis and foreign policy were dominant themes in the debate. In comparison, Merkel's expression of happiness is rewarded by voters (with an exception for 2005) with positive and significant effects in the 2009, 2013, and 2017 debates. These reactions are consistent with our expectations in the voter hypothesis 1.

Recall that leader role expectations and gender role expectations led us to agnostic expectations about the responses to the overall amount of emotional expression from candidates. While women in the general population are granted more leeway in emotional expression, women as leaders may need to avoid the appearance of lacking emotional control (Brescoll 2016). To examine how voters react to the overall level of emotional expression, we look at the total of non-neutral emotional expression as well as vocal pitch and text sentiment. Across our three measures of emotional intensity,

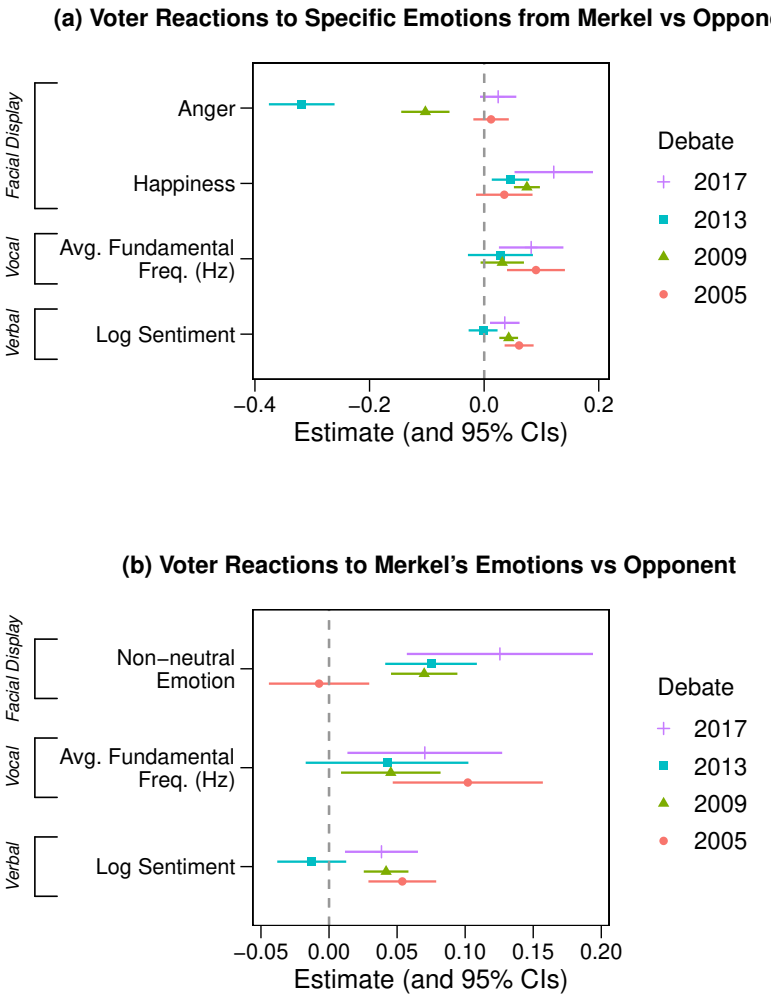


Figure 5: Voter reactions to candidate emotions. Figures (a) and (b) provide an estimate of the cumulative effect (across 4 lags) of the key textual, vocal, and facial variables of interest as outlined in Section 2.1. Note that while (a) only presents facial expressions of emotion for the two emotions of interest (happiness and anger), the model includes all of the non-neutral emotions returned by the Face API. The models include control variables for the gender, party identification, political knowledge, and political interest of respondents. The full estimates are provided in Table A3 in the Supporting Information.

voters generally reward Merkel for her emotional expression, consistent with Voter Hypothesis 2a. Most coefficients for increased emotional intensity measured through voice pitch, and textual sentiment are positive and statistically significant. The exception in the facial non-neutral model is the 2005 debate, when she was a challenger and was the most expressive out of all four debates. In short, while voters respond negatively to Merkel's expression of anger (an emotion incongruent with her gender), they do appear to like her happiness and her general emotional expression.

6.3 Robustness: The 2017 Debate Between Candidates from Smaller Parties

We test the robustness of our findings and conclusions by applying the same set of analyses to the 2017 'minor debate.' We followed the exact same procedure as for the debates involving Merkel: we retrieved facial emotional displays by candidates, analyze the voice pitch, code the statement-level sentiment, and match these values to the second-level RTR data.

Figure 6 shows the variation in facial emotions across the five candidates using the confidence scores. Looking at the two female candidates, Wagenknecht and Weidel, we observe similar patterns to the main leadership debates: both female candidates express high levels of happiness (in particular Weidel with a value of 0.2). Herrmann's happiness expressions are also very high. Özdemir and Lindner express lower expressions of happiness than the female candidates, but their values do correspond closely to happiness detected for Merkel and her male opponents. Comparing this plot to Figure 3 reveals that most male and female candidates do not regularly engage in displays of anger.

Figure 7 displays a similar set of tests to those carried out in Section 6.1. Here we test whether the female candidates in the 2017 minor party debate differed from their male counterparts in terms of how frequently they displayed anger, happiness, and

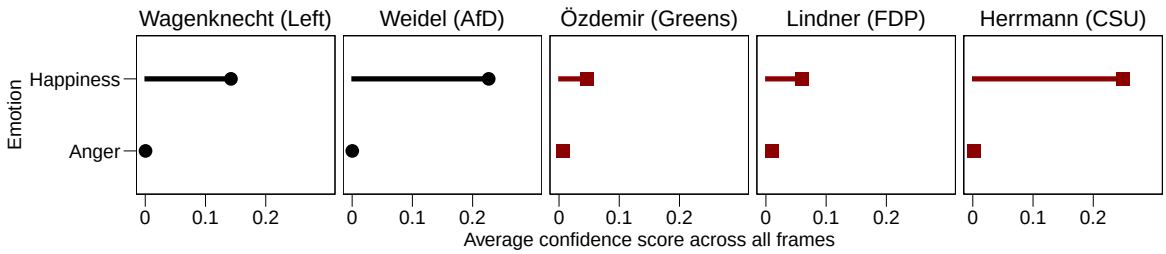


Figure 6: Average confidence scores for emotional displays aggregated for all frames for the five speakers involved in the 2017 minor debate.

general emotive facial displays as well as vocal intensity, with the same methodological approach as with the major debates.

The results are strikingly similar to those of the debates with Angela Merkel. The female candidates display less anger (5% error level) and less likely to elevate their vocal pitch (1% error level).

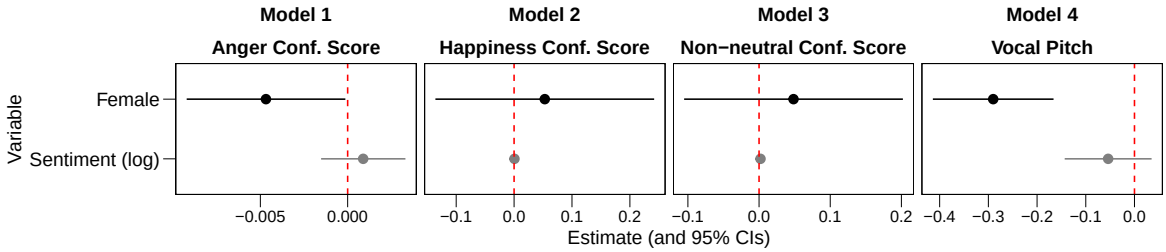


Figure 7: Random-effects panel data linear regression (Models 1–3) and random-effects panel data probit regression (Model 4) results for the 2017 minor debate of per-second average confidence scores of anger, happiness and non-neutral facial displays, and per-second candidate heightened vocal pitch. All models include the lagged dependent variable and controls for masculine, feminine, and "none" debate topics, with neutral topics as the reference category. The x-axes are re-scaled for each model. Coefficients are displayed in Table A4.

On the voter reaction side, the five candidates mean a slight change in the rating procedure. Instead of turning a dial 'for' or 'against' a particular candidate, voters indicated whether they have a bad impression (lower values) or good impression (higher values) on a 1–7 scale (Roßteutscher et al. 2019d;e). 36 eligible voters—a considerable smaller sample of voters than in debates involving Merkel—provided real-time responses during the minor debate. Given that the candidates of the far-left (Wagenknecht) and the far-right (Weidel) parties were female, our results based on

differences in gender should not be confounded by ideological positions of parties or candidates in this debate.

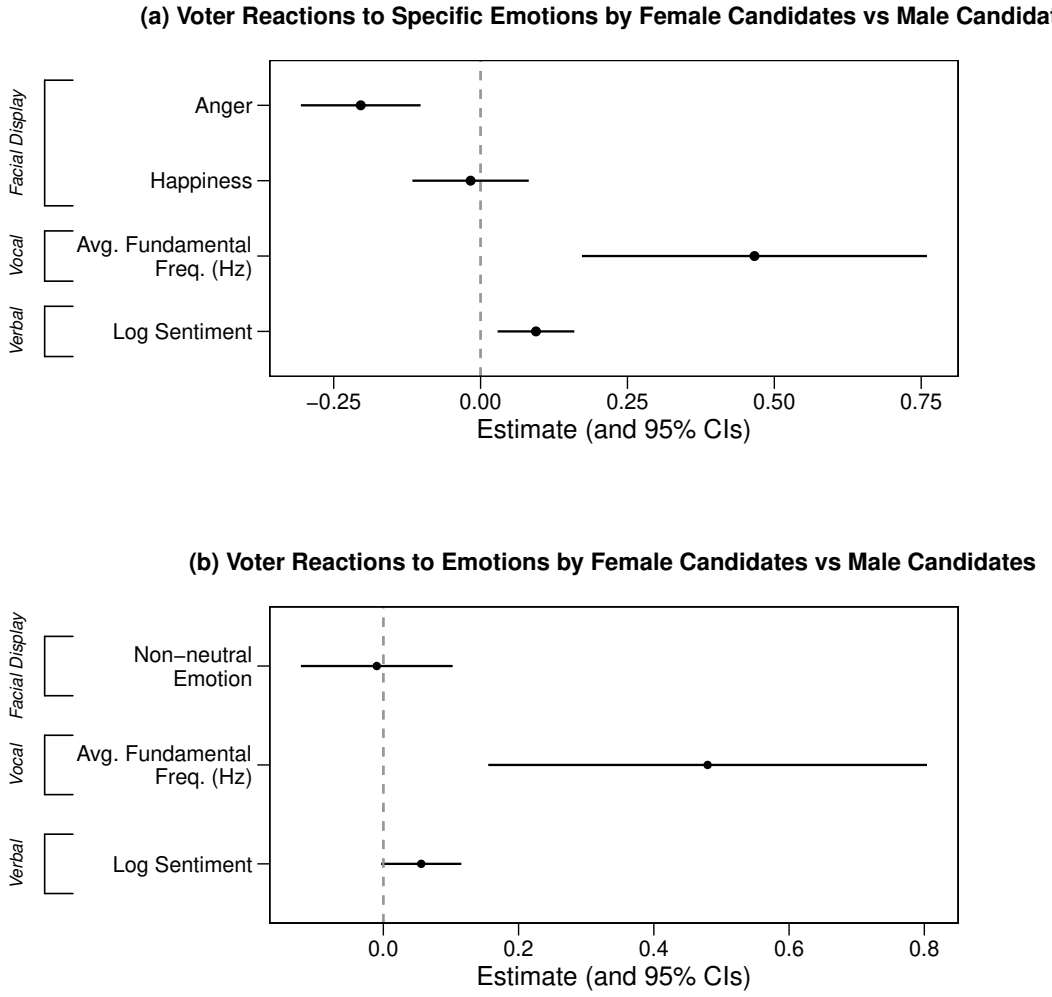


Figure 8: Voter reactions to candidate emotions in the 2017 minor debate. Figures (a) and (b) provide an estimate of the cumulative effect (across 4 lags) of the key textual, vocal, and facial variables of interest as outlined in Section 2.1. Note that while (a) only presents facial expressions of emotion for the two emotions of interest (happiness and anger), the model includes all of the non-neutral emotions returned by the Face API. Positive coefficients indicate that respondents tend to react positively to emotional expressions by one of the two candidates.

We again find that voters react negatively to women's expression of anger and some evidence of positive reactions to women's overall emotional expression. Unlike in the Merkel debates, however, voters do not reward the women in the minor debate for happiness or general facial expressions; they do, however, reward women's emotive

vocal pitch and positive sentiment.

7 Discussion

Despite the importance of political debates and nonverbal cues to electoral outcomes and voter behavior, candidate emotions during debates have received little attention from political scientists. Some of this is due to the methodologically taxing process of manually coding debate images. As a result, the scholarship has often, understandably, relied on snippets of debates, on the text of the debate, or on candidate rhetoric. The integration of real-time-responses with nonverbal cues from candidates is thus a major methodological improvement on understanding how voters perceive politicians in modern political debates.

Drawing on work from psychology, communications, and gender studies, we bring a robust evaluation of candidate gender into dialogue with scholarship on political debates and nonverbal communication. Relying on theories of role congruity and, particularly, gender role congruity, we argue that candidates express nonverbal cues strategically and that voters respond to these. Critically, however, not male and female candidates are equally able to express all emotions because voters assess nonverbal behavior by whether it meets gendered expectations.

Using more than 500,000 frames of candidate facial expressions from four German national debates, we find evidence consistent with our expectations: Merkel is less likely to express anger than her male opponents and she emotes less over the course of the four debates. Examining millions of real-time responses from voters reveals that this strategy for Merkel is successful: Merkel expresses happiness much more frequently than anger, and voters reward Merkel for her presentation of happiness. Indeed, voters reward Merkel generally for her emotional expressions. The results suggest that Merkel's use of emotions is largely successful.

The possibility remains that reactions to the emotions in these debates are fleeting. Yet, analyses of these elections in Germany suggest that the debates were pivotal points in the campaigns (Maier and Faas 2019). As a method of assessing the longer term impacts of the emotional expressions in the debates, we engaged in an analysis of the media coverage of the debates. We retrieved all newspaper articles from six national German outlets across the entire ideological spectrum that mention the TV debate or one of the candidates in the week after each debate. A manual content analysis of over 400 articles reveals that newspapers report extensively about emotions. Across all major debates, between 13% and 18% of the published articles on the debate mentioned at least one emotional display by at least one of the candidates. SI Section D describes the content analysis and results in detail. In short, not only do emotional expressions shape how voters react in real-time to the debates, but also how the media covers the candidates' performances in the debates.

These analyses are just a small piece of what could be learned from nonverbal behavior, particularly in an environment where emotional displays can be obtained at scale through computational methods. Understanding, for example, how voters react when verbal sentiment and nonverbal emotions align or conflict could provide a key to understanding the full context by which voters interpret candidate speech and images during debates. Moving beyond a single measure and evaluating multimodal expression concurrently represents a significant step forward in the scholarship on political communication.

Our results demonstrate the importance of considering the ways that candidates constrain themselves to fit what they think voters want. Angela Merkel, like other women seeking positions of power that have been denied to them for centuries (Dittmar et al. 2018), is well aware that her gender shapes how voters react to her. That Merkel – and women in the minor party debate – expresses little anger during these debates suggest that she adjusts her behavior to better fit voter expectations. Yet, this may

also constrain women's ability to lead in different contexts. Research might examine whether this means that women are less likely to be selected for positions of leadership during times of foreign policy crisis, when voters might want an angry leader who will defend them. Future studies might also consider the ways that powerful women express anger in alterate ways; by expressing surprise or disgust, for example. Our research also speaks to the experiences and judgement of women outside politics. We would expect that women's anger would be constrained in business and philanthropy settings, just as in politics.

References

Acker, J. (1992). From sex roles to gendered institutions. *Contemporary Sociology*, 21(5):565–569.

Anastasopoulos, L. J., Badani, D., Lee, C., Ginosar, S., and Williams, J. (2016). Photographic home styles in congress: a computer vision approach. *arXiv preprint arXiv:1611.09942*.

Anderson, R. C. and Klofstad, C. A. (2012). Preference for leaders with masculine voices holds in the case of feminine leadership roles. *PLoS One*, 7(12).

Bakker, B. N., Schumacher, G., and Rooduijn, M. (2020). Hot politics? affective responses to political rhetoric. *American Political Science Review*, FristView.

Bargal, S., Barsoum, E., Ferrer, C., and Zhang, C. (2016). Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, page 433–436.

Bauer, N. M. (2015). Emotional, sensitive, and unfit for office? Gender stereotype activation and support female candidates. *Political Psychology*, 36(6):691–708.

Bauer, N. M. (2017). The effects of counterstereotypic gender strategies on candidate evaluations. *Political Psychology*, 38(2):279–295.

Bauer, N. M. (2018). Untangling the relationship between partisanship, gender stereotypes, and support for female candidates. *Journal of Women, Politics & Policy*, 39(1):1–25.

Bauer, N. M. (2019). The effects of partisan trespassing strategies across candidate sex. *Political Behavior*, 41(4):897–915.

Bauer, N. M. (2020). Shifting standards: How voters evaluate the qualifications of female and male candidates. *The Journal of Politics*, 82(1):1–12.

Bauer, N. M. and Carpinella, C. M. (2018). Visual information and candidate evaluations: The influence of feminine and masculine images on support for female candidates. *Political Research Quarterly*, 71(2):395–407.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *The Journal of Open Source Software*, 3(30):774.

Benoit, W., Hansen, G., and Verser, R. (2003). A meta-analysis of the effects of viewing us presidential debates. *Communication Monographs*, 70(4):335–350.

Boersma, P. and Weenink, D. (2018). Praat: Doing phonetics by computer [computer program]. version 6.0.37.

Boussalis, C. and Coan, T. G. (2020). Facing the electorate: Computational approaches to the study of nonverbal communication and voter impression formation. *Political Communication*.

Boydstun, A. E., Glazier, R. A., Pietryka, M. T., and Resnik, P. (2014). Real-time reactions to a 2012 presidential debate: A method for understanding which messages matter. *Public Opinion Quarterly*, 78(S1):330–343.

Brescoll, V. L. (2016). Leading with their hearts? how gender stereotypes of emotion lead to biased evaluations of female leaders. *The Leadership Quarterly*, 27(3):415–428.

Brody, L. (2009). *Gender, Emotion, and the Family*. Harvard University Press, Boston.

Brooks, D. J. (2013). *He Runs, She Runs: Why Gender Stereotypes Do Not Harm Women Candidates*. Princeton University Press, Princeton.

Brydon, S., Hellweg, S. A., and Pfau, M. (1992). *Televised Presidential Debates: Advocacy in Contemporary America*. Praeger, Westport.

Bucy, E. P. and Grabe, M. E. (2007). Taking television seriously: A sound and image bite analysis of presidential campaign coverage, 1992–2004. *Journal of Communication*, 57(4):652–675.

Bucy, E. P. and Grabe, M. E. (2008). Happy warriors' revisited: Hedonic and agonic display repertoires of presidential candidates on the evening news. *Politics and the Life Sciences*, 27(1):78–98.

Bucy, E. P. and Stewart, P. (2018). The personalization of campaigns: Nonverbal cues in presidential debates. In *Oxford Research Encyclopedia of Politics*.

Bystrom, D. G., Robertson, T., Banwart, M. C., and Kaid, L. L. (2005). *Gender and Candidate Communication: VideoStyle, WebStyle, NewStyle*. Routledge, New York.

Cantú, F. (2019). The fingerprints of fraud: Evidence from Mexico's 1988 presidential election. *American Political Science Review*, 113(3):710–726.

Carpinella, C. M. (2016). Visual political communication: The impact of facial cues from social constituencies to personal pocketbooks. *Social Personality Psychology Compass*, 10(5):281–297.

Carpinella, C. M. and Bauer, N. M. (2019). A visual analysis of gender stereotypes in campaign advertising. *Politics, Groups, and Identities*, published ahead of print.

Carpinella, C. M., Hehman, E., Freeman, J. B., and Johnson, K. L. (2016). The gendered face of partisan politics: Consequences of facial sex typicality for vote choice. *Political Communication*, 33(1):21–38.

Carpinella, C. M. and Johnson, K. L. (2013). Appearance-based politics: Sex-typed facial cues communicate political party affiliation. *Journal of Experimental Social Psychology*, 49(1):156–160.

Casas, A. and Williams, N. W. (2019). Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly*, 72(2):360–375.

Cassese, E. C. and Holman, M. R. (2018). Party and gender stereotypes in campaign attacks. *Political Behavior*, 40(3):785–807.

Copeland, C. L., Driskell, J. E., and Salas, E. (1995). Gender and reactions to dominance. *Journal of Social Behavior and Personality*, 10(4):53–68.

Dietrich, B., Hayes, M., and O'Brien, D. Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech on Women. *American Political Science Review*, 113(4):941–962.

Dittmar, K. (2015). *Navigating Gendered Terrain: Stereotypes and Strategy in Political Campaigns*. Temple University Press, Philadelphia.

Dittmar, K., Sanbonmatsu, K., and Carroll, S. J. (2018). *A Seat at the Table: Congresswomen's Perspectives on Why Their Presence Matters*. Oxford University Press, New York.

Druckman, J. N. (2003). The power of television images: The first Kennedy-Nixon debate revisited. *The Journal of Politics*, 65(2):559–571.

Druckman, J. N. (2005). Media matter: How newspapers and television news cover campaigns and influence voters. *Political Communication*, 22(4):463–481.

Durik, A. M., Hyde, J. S., Marks, A. C., Roy, A. L., Anaya, D., and Schultz, G. (2006). Ethnicity and gender stereotypes of emotion. *Sex Roles*, 54(7-8):429–445.

Eagly, A. H. and Johannesen-Schmidt, M. (2001). The leadership styles of men and women. *Journal of Social Issues*, 57(4):781–797.

Eagly, A. H. and Karau, S. J. (2001). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3):573–598.

Eagly, A. H. and Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological review*, 109(3):573.

Eibl-Eibesfeldt, I. (1979). Ritual and ritualization from a biological perspective. In

Human Ethology: Claims and Limits of a New Discipline. Cambridge University Press, Cambridge.

Ekman, P. and Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.

Everitt, J., Best, L. A., and Gaudet, D. (2016). Candidate gender, behavioral style, and willingness to vote: Support for female candidates depends on conformity to gender norms. *American Behavioral Scientist*, 60(14):1737–1755.

Fischbach, A., Lichtenthaler, P. W., and Horstmann, N. (2015). Leadership and gender stereotyping of emotions. *Journal of Personnel Psychology*.

Fridkin, K. L., Gershon, S. A., Courey, J., and LaPlant, K. (2019). Gender differences in negative affect and well-being: The case for emotional intensity. *Political Behavior*, published ahead of print.

Gabriel, O. W. and Masch, L. (2017). Displays of emotion and citizen support for merkel and gysi. *Politics and the Life Sciences*, 36(2):80–103.

Gleason, S. A. (2020). Beyond mere presence: Gender norms in oral arguments at the us supreme court. *Political Research Quarterly*, 73(3):596–608.

Grabe, M. E. and Bucy, E. P. (2009). *Image Bite Politics: News and the Visual Framing of Elections*. Oxford University Press, Oxford.

Heilman, M. E., Block, C. J., and Martell, R. F. (1995). Sex stereotypes: Do they influence perceptions of managers? *Journal of Social behavior and Personality*, 10(4):237.

Jadoul, Y., Thompson, B., and De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Joo, J., Bucy, E. P., and Seidel, C. (2019). Automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision and deep learning. *International Journal of Communication*, 13:23.

Joo, J. and Steinert-Threlkeld, Z. (2018). Image as data: Automated visual content

analysis for political science. *arXiv preprint arXiv:1810.01544*.

Klofstad, C. A. (2015). Looks and sounds like a winner: Perceptions of competence in candidates' faces and voices influences vote choice. *Journal of Experimental Political Science*, 4(3):229–240.

Klofstad, C. A., Anderson, R. C., and Nowicki, S. (2015). Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. *PLoS One*, 10(8):1–7.

Kring, A. M. and Gordon, A. H. (1998). Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality and Social Psychology*, 74(3):686–703.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, page 1097–1105.

Lanoue, D. and Schrott, P. (1989). *Voters' reactions to televised presidential debates: Measurement of the source and magnitude of opinion change*. Political Psychology.

Lazarus, J. and Steigerwalt, A. (2018). *Gendered Vulnerability*. University of Michigan Press, Ann Arbor.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, page 2278–2324.

Maier, J. and Faas, T. (2011). 'Miniature campaigns' in comparison: The German televised debates, 2009–09. *German Politics*, 20(1):75–91.

Maier, J. and Faas, T. (2019). *TV-Duelle*. Springer VS, Wiesbaden.

Maier, J. and Jansen, C. (2017). When do candidates attack in election campaigns? exploring the determinants of negative candidate messages in German televised debates. *Party Politics*, 34(5):549–559.

Masch, L. (2020). *Politicians' Expressions of Anger and Leadership Evaluations: Empirical Evidence from Germany*. Nomos, Baden-Baden.

Meeks, L. (2012). Is she 'man enough'? Women candidates, executive political offices, and news coverage. *Journal of Communication*, 62(1):175–193.

Nagel, F., Maurer, M., and Reinemann, C. (2012). How verbal, visual, and vocal communication shape viewers' impressions of political candidates. *Journal of Communication*, 65(5):833–850.

Oliver, S. and Conroy, M. (2020). *Who Runs? The Masculine Advantage in Candidate Emergence*. University of Michigan Press, Ann Arbor.

Ondercin, H. L. (2020). Why strategic behavior doesn't always lead to increased women's representation. Working Paper.

Plant, E. A., Hyde, J. S., Keltner, D., and Devine, P. G. (2000). The gender stereotyping of emotions. *Psychology of Women Quarterly*, 24(1):81–92.

Proksch, S.-O., Lowe, W., Wäckerle, J., and Soroka, S. N. (2019). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1):97–131.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2010). TV-Duell-Analyse, Inhaltsanalyse TV-Duell (GLES 2009). GESIS Datenarchiv, Köln. ZA5311 Datenfile Version 1.0.0.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2011a). TV-Duell-Analyse, Befragung (GLES 2009). GESIS Datenarchiv, Köln. ZA5309 Datenfile Version 2.0.0.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2011b). TV-Duell-Analyse, Real-Time-Response-Daten (GLES 2009). GESIS Datenarchiv, Köln. ZA5310 Datenfile Version 1.1.0.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2014). TV-Duell-Analyse Real-Time-

Response-Messung (Dial) (GLES 2013). GESIS Datenarchiv, Köln. ZA5711 Datenfile Version 1.0.0.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2015). TV-Duell-Analyse Befragung (GLES 2013). GESIS Datenarchiv, Köln. ZA5709 Datenfile Version 3.0.0.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2018). TV-Duell-Analyse Inhaltsanalyse (GLES 2013). GESIS Datenarchiv, Köln. ZA5710 Datenfile Version 2.2.0.

Ridout, T. N. and Searles, K. (2011). It's my campaign i'll cry if i want to: How and when campaigns use emotional appeals. *Political Psychology*, 32(3):439–458.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2019a). TV-Duell-Analyse, Befragung (GLES 2017). GESIS Datenarchiv, Köln. ZA6810 Datenfile Version 1.0.0.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2019b). TV-Duell-Analyse, Inhaltsanalyse TV-Duell (GLES 2017). GESIS Datenarchiv, Köln. ZA6811 Datenfile Version 1.0.0.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2019c). TV-Duell-Analyse, Real-Time-Response-Daten TV-Duell (dial) (GLES 2017). GESIS Datenarchiv, Köln. ZA6812 Datenfile Version 1.0.0.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Faas, T., Maier, J., and Maier, M. (2019d). TV-Duell-Analyse, Inhaltsanalyse Fünfkampf (GLES 2017). GESIS Datenarchiv, Köln. ZA6829 Datenfile Version 1.0.0.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Faas, T., Maier, J., and Maier, M. (2019e). TV-Duell-Analyse, Real-Time-Response Daten Fünfkampf (dial). GESIS Datenarchiv, Köln. ZA6830 Datenfile Version 1.0.0.

Schill, D., Kirk, R., and Jasperson, A. E. (2016). *Political Communication in Real*

Time: Theoretical and Applied Research Approaches. Routledge, New York.

Schneider, M. C. and Bos, A. L. (2014). Measuring stereotypes of female politicians. *Political psychology*, 35(2):245–266.

Schneider, M. C. and Bos, A. L. (2019). The application of social role theory to the study of gender in politics. *Political Psychology*, 40(S1):173–213.

Schoonvelde, M., Schumacher, G., and Bakker, B. N. (2019). Friends with text as data benefits: Assessing and extending the use of automated text analysis in political science and political psychology. *Journal of Social and Political Psychology*, 7(1):124–143.

Silva, A. and Skulley, C. (2019). Always running: Candidate emergence among women of color over time. *Political Research Quarterly*, 72(2):342–359.

Stewart, P., Salter, F., and Mehu, M. (2011). *The face as a focus of political communication: Evolutionary perspectives and the ethological methods. The Sourcebook for Political Communication Research*. Routledge, New York.

Swers, M. L. (2013). *Women in the Club: Gender and Policy Making in the Senate*. University of Chicago Press, Chicago.

Sülflow, M. and Maurer, M. (2019). The power of smiling. how politicians' displays of happiness affect viewers' gaze behavior and political judgments. In *Visual political communication*, page 1097–1105. Palgrave Macmillan.

Torres, M. (2018). Give me the full picture: Using computer vision to understand visual frames and political communication. *URL: http://qssi. psu. edu/new-faces-papers-2018/torres-computer-vision-and-politicalcommunication*.

Torres, M. and Cantú, F. (2020). Learning to see: Visual analysis for social science data. *Political Analysis*, 74(Forthcoming).

Van Boven, L. and Robinson, M. D. (2012). Boys don't cry: Cognitive load and priming increase stereotypic sex differences in emotion memory. *Journal of Experimental Social Psychology*, 48(1):303–309.

Zhang, H. and Pan, J. (2019). Casm: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, 49(1):1–57.

Supporting Information

Contents

Appendix A The German Debates

A.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 as the strongest party, but only 6,000 votes separated the SPD and the CDU/CSU (Roberts 2006). Under Schröder's tenure, Germany suffered a severe economic crisis. The Economist famously called Germany the "sick man of the Euro".[12] Because of rising unemployment, the government implemented a set of massive labour market reforms (the so-called Hartz IV reforms) during the second "red-green" coalition. These reforms led to the exit of high-level social democrats from the party in 2004. Their new party (WASG) later merged with the successor of the PDS, the former Communist party in the German Democratic Republic (GDR) and was remade 'The Left-Party PDS.'

In May 2005, Chancellor Schröder announced that he aimed for an early election to strengthen his position of further reforms of the economy and labour-market. This announcement happened shortly after the SPD lost the State election in North-Rhine Westphalia, which was regarded as the stronghold of social democracy in Germany. 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, and opinion surveys suggested that Schröder outperformed Merkel. 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.

[12] The Economist, 3 June 1999: https://www.economist.com/special/1999/06/03/the-sick-man-of-the-euro

A.2 Merkel v Steinmeier (2009)

The "grand coalition" under the leadership of Angela Merkel coalition worked pragmatically and smoothly,[13] 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.

The SPD selected Frank-Walter Steinmeier, Secretary of State and vice-chancellor, as their candidate for the election in 2009. 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), resulting in a loss of 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.

A.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. However, Angela Merkel's popularity was not affected by this performance. "Almost miraculously, however, Chancellor Merkel [...] remained largely unaffected by these disputes. Instead, she reigned in an almost presidential style, well above everyday business" (Faas 2015 239).

The first central event happened in autumn of 2012, around one year before the election. Three SPD politicians were regarded as potential contenders for becoming the party's main candidate for the election in 2013: the party chairman Sigmar Gabriel, Frank-Walter Steinmeier (again), or Peer Streinbrück. 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. The SPD presented Steinbrück 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. Another issue was that the left-wing manifesto of the 2013 election did not fit with the more moderate candidate Steinbrück (Faas 2015).

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

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). With 4.7% of list votes, the Eurosceptic – and not yet right-wing populist party – Alternative für Deutschland (AfD), which was founded only a few months prior to the election, narrowly failed to pass the five percent threshold of list votes to obtain seats in the Bundestag.

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 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.

A.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, instead. Shortly after this announcement, the support for the SPD increased drastically. Party support increased by over 10 percentage points within weeks, and the party gained over 10,000 new members. 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. For instance, in the summer of 2017, in an on-stage interview Merkel changed her opinion from openly opposing same-sex marriage, and instead noted that it was an issue of conscience. Shortly afterwards, same-sex marriage was introduced. The SPD claimed that the party delivered on a lot of their central promises during the 2005–2009 and 2013–2017 "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, but also portrayed her style of governing as an 'attack on democracy"' (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 B Comparing the Automated Detection of Emotions with Human Coding of Emotional Displays

While both vocal pitch and sentiment have been extensively validated elsewhere (see Dietrich et al. 2019 and 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 leadership debates. We examined the validity of the Face API predictions by comparing them to a large sample of human coded clips across the four debates ($N = 1,341$). To generate the validation set, we recruited two research assistants 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 A1 for an illustration of the annotation tool). Following Boussalis and Coan (2020), the coders were asked to rate the level of emotion expressed on a five point scale ranging from "not at all" to "extremely". After completing a short training session, both annotators coded a sample of 75 clips. The inter-coder reliability between the annotators was reasonably high based on this sample (Krippendorff's $\alpha = 0.81$). The coders then annotated an additional 1,134 clips which are used to establish the correspondence between the model and human judgements.

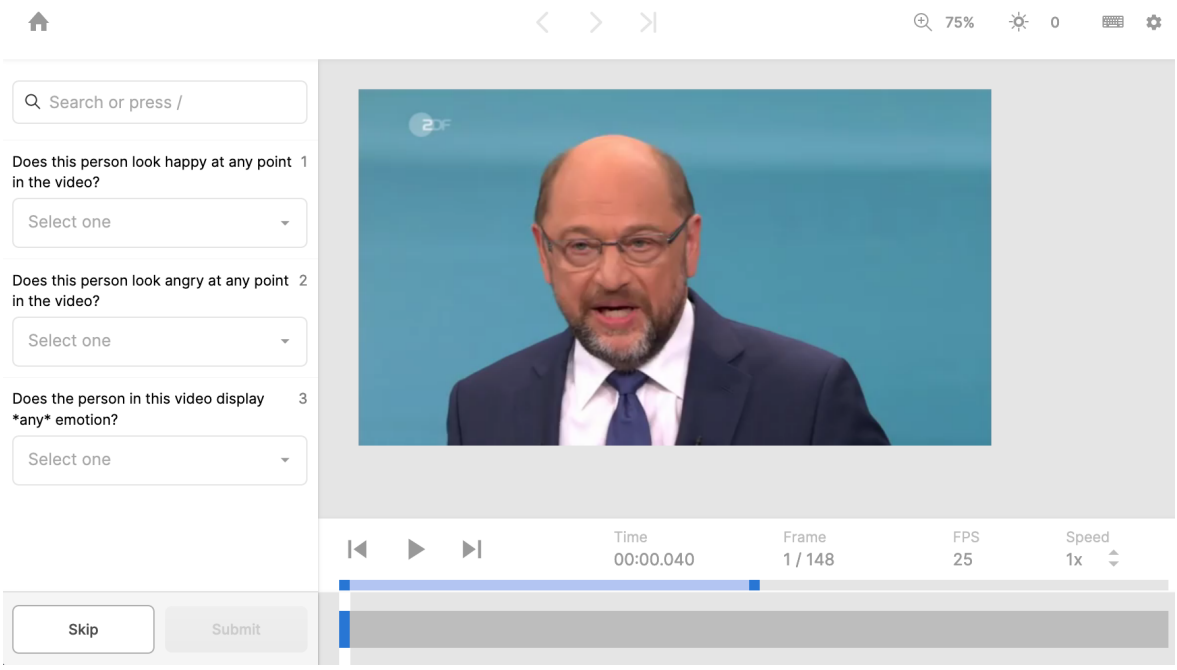


Figure A1: Example of the annotation tool. The coding of debate clips was carried out using software from Labelbox (see `https://labelbox.com`).

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 an-

notations by assessing the root mean squared error (RMSE), where zero error is represented by an RMSE value of 0. To assess out-of-sample performance, we employ five-fold, repeated cross-validation. The RSME for the expression of "any emotion" is 0.83, suggesting that predictions based on the model are within less than a point on the scale of 1-5. Moreover, consistent with Boussalis and Coan (2020), we find that the model does a better job at predicting happiness ($RMSE = 0.82$) than it does for anger ($RMSE = 1.03$). While the correspondence between the computer and human coding does not align perfectly, the relatively low RMSE as compared to the scale suggests accuracy in coding emotions.

In addition to examining performance via the RMSE, Boussalis and Coan (2020) examine *classification* performance by transforming the Likert scale measure of emotions into a binary measure. We perform a similar analysis here. First, we recode each emotion measure (anger, happiness, and any emotion) to equal 1 for clips coded as "very much" or "extremely" and 0 otherwise. Second, 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 and with the results presented in Boussalis and Coan (2020). The F1 score for happiness 0.95 (precision = 0.94, recall = 0.97), the score for anger is 0.77 (precision = 0.64, recall = 0.96), and the F1 score for any emotion 0.58 (precision = 0.57, recall = 0.58).

The data produced in Nagel et al. (2012) offer an additional opportunity to examine the validity of our happiness measure. Nagel et al. (2012) hand coded the smiles the candidates during the 2005 debate. Their codebook distinguishes between no smile, light smile, and strong smile. This variable is coded for every second in the debate. 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 A2 confirms 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 (1.61 and 0.66) of happiness, but the mean is considerable lower. For seconds coded as 'no smile' the averages are lowest (-0.22 for both candidates). Given that both measures clearly contain measurement error, these results are very encouraging: the human-coded assessment and automated classification of happiness show a high levels of similarity.

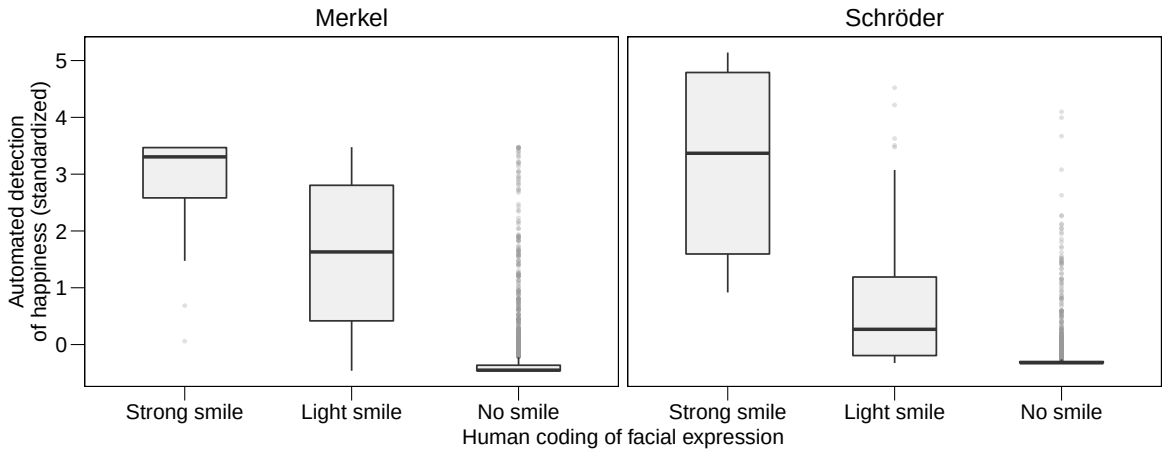


Figure A2: Comparing the coding of smiling in the 2005 debate, with the automated detection of emotional displays (y-axis), standardized by speaker. The manual codings are retrieved from the replication materials Nagel et al. (2012).

Appendix C Supplementary Tables and Plots

	(1) Anger conf. score	(2) Happiness conf. score	(3) Non-neutral conf. score	(4) Vocal pitch
Merkel	-0.00302***	0.000549	0.000121	-0.284**
	(0.000709)	(0.00384)	(0.00469)	(0.125)
2009 Debate	0.0000859	-0.00979***	-0.0109***	0.0783**
	(0.000466)	(0.00339)	(0.00363)	(0.0340)
2013 Debate	0.000954**	-0.0111***	-0.00849**	0.278***
	(0.000456)	(0.00331)	(0.00355)	(0.0320)
2017 Debate	0.000310	-0.00656**	-0.00556	0.149***
	(0.000455)	(0.00331)	(0.00355)	(0.0568)
Topic: Feminine	-0.0000405	-0.0110***	-0.0110***	-0.0230
	(0.000380)	(0.00292)	(0.00304)	(0.112)
Topic: Masculine	0.00121***	-0.0125***	-0.0118***	0.0265
	(0.000359)	(0.00275)	(0.00287)	(0.101)
Topic: None	-0.0000705	0.00849***	0.00873***	0.159***
	(0.000386)	(0.00292)	(0.00305)	(0.0102)
Sentiment (log)	-0.000301**	-0.000922	-0.00121	0.0164
	(0.000135)	(0.00102)	(0.00107)	(0.0473)
Anger conf. score (t-1)	0.725***			
	(0.00464)			
Happiness conf. score (t-1)		0.775***		
		(0.00476)		
Non-neutral conf. score (t-1)			0.757***	
			(0.00489)	
Constant	0.00272***	0.0292***	0.0427***	-1.958***
	(0.000551)	(0.00381)	(0.00419)	(0.151)
Observations	19,034	19,034	19,034	20,136
Baltagi-Wu LBI	2.272	1.966	2.018	

Table A1: Random-effects panel data linear regression (models 1–3) and random-effects panel data probit regression (model 4) results of per-second average confidence scores of anger, happiness and non-neutral facial displays, and per-second candidate heightened vocal pitch. Models 1–3 report the Baltagi-Wu LBI test statistic. Model 4 includes Huber-White standard errors. Reference categories in all models are the 2005 debate and "neutral" gender topic. $***p < 0.01, **p < 0.05, *p < 0.1$.

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 A3: This table presents the cumulative effects across 4 lags for all seven emotions provided by the Face API.

Debate year	Emotion	Coef.	SE	t-statistic	p-value	95% CI	
2005	Anger	0.012	0.016	0.770	0.446	-0.019	0.044
2005	Happiness	0.035	0.025	1.400	0.166	-0.015	0.085
2005	Fear	0.045	0.013	3.560	0.001	0.020	0.070
2005	Disgust	-0.005	0.016	-0.290	0.776	-0.036	0.027
2005	Contempt	-0.155	0.025	-6.140	0.000	-0.205	-0.104
2005	Sadness	-0.055	0.035	-1.590	0.117	-0.125	0.014
2005	Surprise	0.095	0.015	6.150	0.000	0.064	0.126
2005	Vocal Freq.	0.090	0.026	3.510	0.001	0.039	0.142
2005	Sentiment	0.061	0.013	4.720	0.000	0.035	0.087
2009	Anger	-0.103	0.021	-4.770	0.000	-0.145	-0.060
2009	Happiness	0.075	0.012	6.450	0.000	0.052	0.097
2009	Fear	0.104	0.026	3.970	0.000	0.052	0.156
2009	Disgust	0.064	0.012	5.480	0.000	0.041	0.087
2009	Contempt	-0.048	0.014	-3.540	0.001	-0.075	-0.021
2009	Sadness	-0.365	0.081	-4.530	0.000	-0.525	-0.206
2009	Surprise	0.045	0.019	2.350	0.020	0.007	0.083
2009	Vocal Freq.	0.032	0.019	1.640	0.104	-0.007	0.070
2009	Sentiment	0.043	0.008	5.140	0.000	0.026	0.059
2013	Anger	-0.318	0.029	-10.890	0.000	-0.376	-0.260
2013	Happiness	0.046	0.017	2.760	0.007	0.013	0.079
2013	Fear	-0.269	0.044	-6.110	0.000	-0.356	-0.181
2013	Disgust	0.146	0.025	5.830	0.000	0.096	0.195
2013	Contempt	-0.007	0.033	-0.220	0.829	-0.074	0.059
2013	Sadness	0.313	0.058	5.410	0.000	0.198	0.428
2013	Surprise	-0.019	0.018	-1.060	0.294	-0.056	0.017
2013	Vocal Freq.	0.028	0.029	0.980	0.329	-0.029	0.086
2013	Sentiment	-0.002	0.013	-0.140	0.887	-0.027	0.024
2017	Anger	0.024	0.016	1.510	0.137	-0.008	0.057
2017	Happiness	0.122	0.035	3.480	0.001	0.051	0.192
2017	Fear	0.038	0.027	1.400	0.169	-0.017	0.093
2017	Disgust	-0.068	0.025	-2.710	0.009	-0.119	-0.018
2017	Contempt	-0.290	0.099	-2.920	0.005	-0.490	-0.090
2017	Sadness	0.048	0.022	2.190	0.034	0.004	0.093
2017	Surprise	0.036	0.021	1.710	0.094	-0.006	0.078
2017	Vocal Freq.	0.082	0.029	2.860	0.006	0.024	0.140
2017	Sentiment	0.036	0.013	2.740	0.009	0.010	0.063

Figure A3 reports the voter reactions to all specific emotions from Merkel vs her opponent. It reproduces the upper panel of Figure 5 but does not only report anger and happiness, but also contempt, disgust, surprise, fear, and sadness. Given that our main expectations related only to happiness and anger, we do not report the other coefficients in the main paper. Nevertheless, the results are substantively relevant. Before interpreting these findings, it is important to keep in mind that many of these emotional displays do not appear often in the sample.

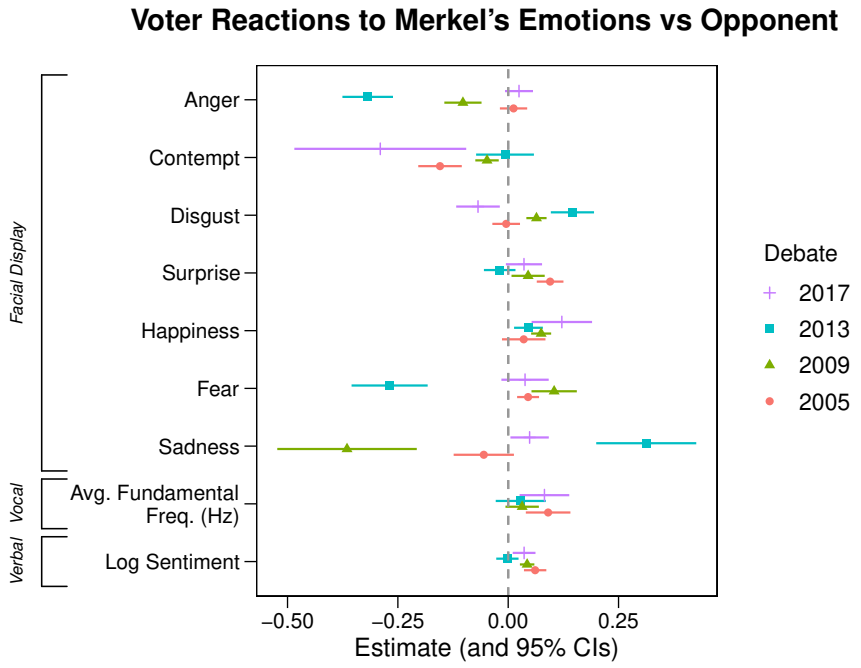
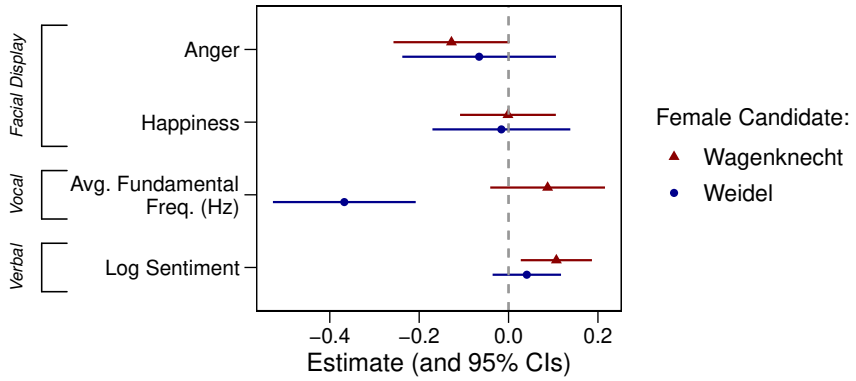


Figure A3: Voter reactions to the cumulative effect (across 4 lags) of the key textual, vocal, and facial variables of interest as outlined in Section 2.1. The models include control variables for the gender, party identification, political knowledge, and political interest of respondents.

	(1) Anger conf. score	(2) Happiness conf. score	(3) Non-neutral conf. score	(4) Vocal pitch
Female	-0.00468**	0.0530	0.0484	-0.290***
	(0.00164)	(0.0680)	(0.0554)	(0.0632)
Topic: Feminine	0.000457	-0.0241	-0.0166	0.350
	(0.000284)	(0.0122)	(0.0143)	(0.218)
Topic: Masculine	0.000442	0.00473	0.00902	0.424**
	(0.000485)	(0.0116)	(0.0159)	(0.193)
Topic: None	4.18e-05	0.00136	0.0114	0.557**
	(0.000148)	(0.00589)	(0.00710)	(0.223)
Sentiment (log)	0.000894	0.000626	0.00202	-0.0541
	(0.000870)	(0.00114)	(0.00174)	(0.0455)
Constant	0.00442*	0.126	0.170**	-2.142***
	(0.00186)	(0.0593)	(0.0476)	(0.177)
Observations	5,021	5,021	5,021	5,021
R-squared	0.006	0.015	0.016	
Pseudo-R2				0.036
Durbin-Watson statistic	1.886	1.868	1.926	

Table A4: Prais-Winsten AR(1) regression (models 1–3) and probit regression (model 4) results of per-second average confidence scores of anger, happiness and non-neutral facial displays, and per-second candidate heightened vocal pitch for 2017 minor party debate participants. Models 1–3 report the transformed Durbin-Watson statistic for the Prais-Winsten regression models. All models include standard errors clustered by candidate. $***p < 0.01, **p < 0.05, *p < 0.1$.

(a) Voter Reactions to Specific Emotions by Female Candidates vs Male Candidates



(b) Voter Reactions to Emotions by Female Candidates vs Male Candidates

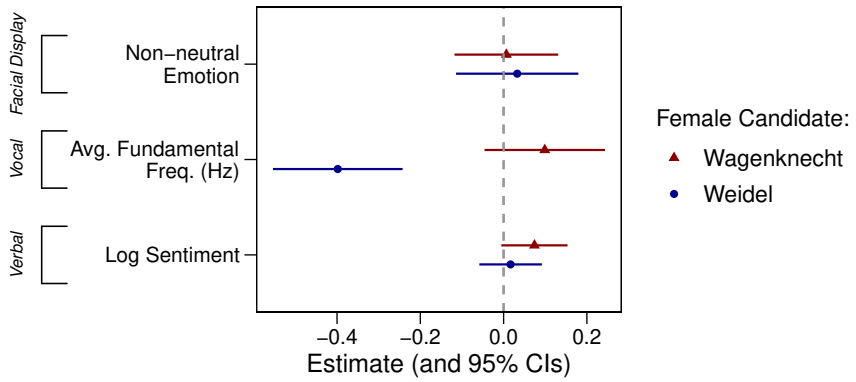


Figure A4: Voter reactions to candidate emotions in the 2017 minor debate. Figures (a) and (b) provide an estimate of the cumulative effect (across 4 lags) of the key textual, vocal, and facial variables of interest as outlined in Section 2.1. Note that while (a) only presents facial expressions of emotion for the two emotions of interest (happiness and anger), the model includes all of the non-neutral emotions returned by the Face API. Positive coefficients indicate that respondents tend to react positively to emotional expressions by Weidel or Wagenknecht (the baseline group are the three male candidates). The models include control variables for the gender, party identification, political knowledge, and political interest of respondents.

Appendix D Media Coverage of Emotions During German TV Debates

Televised leaders' debates in Germany are the most important campaign event during election campaigns (Maier and Faas 2019). Previous research, however, has not analyzed the degree to which emotions expressed by candidates are covered in the media. Given that media coverage of campaign events is highly relevant, we examine how and when news outlets report on emotional displays.

We downloaded all newspaper articles from the online database *LexisNexis* that mentioned the TV debate or one of the candidates and were published within a window of seven days after each debate.[14] Overall, this results in a sample of 426 articles. We selected six newspapers from the entire ideological spectrum in Germany (taz, die Tageszeitung; Der Spiegel; Der Tagesspiegel; Die ZEIT; Süddeutsche Zeitung; Die Welt/Welt am Sonntag).[15]

A research assistant read the full text of all 426 articles. Afterwards, the coder assessed (1) whether an article reported on any emotion of any of the candidates, (2) whose candidate emotions was mentioned, and (3) whether the reported emotional display was agonistic, hedonic, or something else. The instructions asked the coder to indicate agonistic emotions if the media coverage discussed "anger, threat, enraged, feisty, bold, aggressive, or eager and willing to do political battle." Hedonic emotions included any coverage of "happiness, reassurance, optimistic, cheery, full of hope, and channeling a positive feeling about what is likely to happen." Other emotions included "fear, evasion, timid, unsure, equivocal, uncertainty, indecision, weakness, anxiety, uneasiness, apprehension, or agitation in response to a difficult situation." Coverage might reference a candidate's voice trembling or stuttering, misspeaking, or being reluctant to answer a question." These coding categories are adapted from the extant scholarship (Grabe and Bucy 2009).

First, Figure A5 shows the number of newspaper articles reporting on a candidate or the TV debate for each year, along with the number of articles that explicitly mention an emotion. Almost all news outlets published articles that covered one of the candidates' emotions. This alone suggests that the emotional displays by candidates are considered important by the media and that the media coverage may reinforce voter reactions to the emotional displays.

Second, Figure A6 reports the proportions of news articles that cover an emotional display. Between 13% and 18% of the articles mention at least some emotional display. Given that we retrieved *all* articles mentioning a candidate, this proportion suggests a centering of emotional displays in coverage of the candidates.

Third, Figure A7 shows the different types of emotions an article focuses on.[16] In

[14] For instance, the search terms for the debate in 2005 are `TV-Duell OR Fernsehduell AND Schröder OR Merkel`.

[15] For the debate in 2005, we had to limit the sample to three outlets because only taz, Die Tageszeitung, Süddeutsche Zeitung, and Die Welt/Welt am Sonntag were available at *Nexis Lexis*. Note that the newspaper landscape in Germany is generally less polarized than outlets in other countries, such as the United States or the United Kingdom.

[16] If an article mentions more than one emotional display, it is included repeatedly (one observation

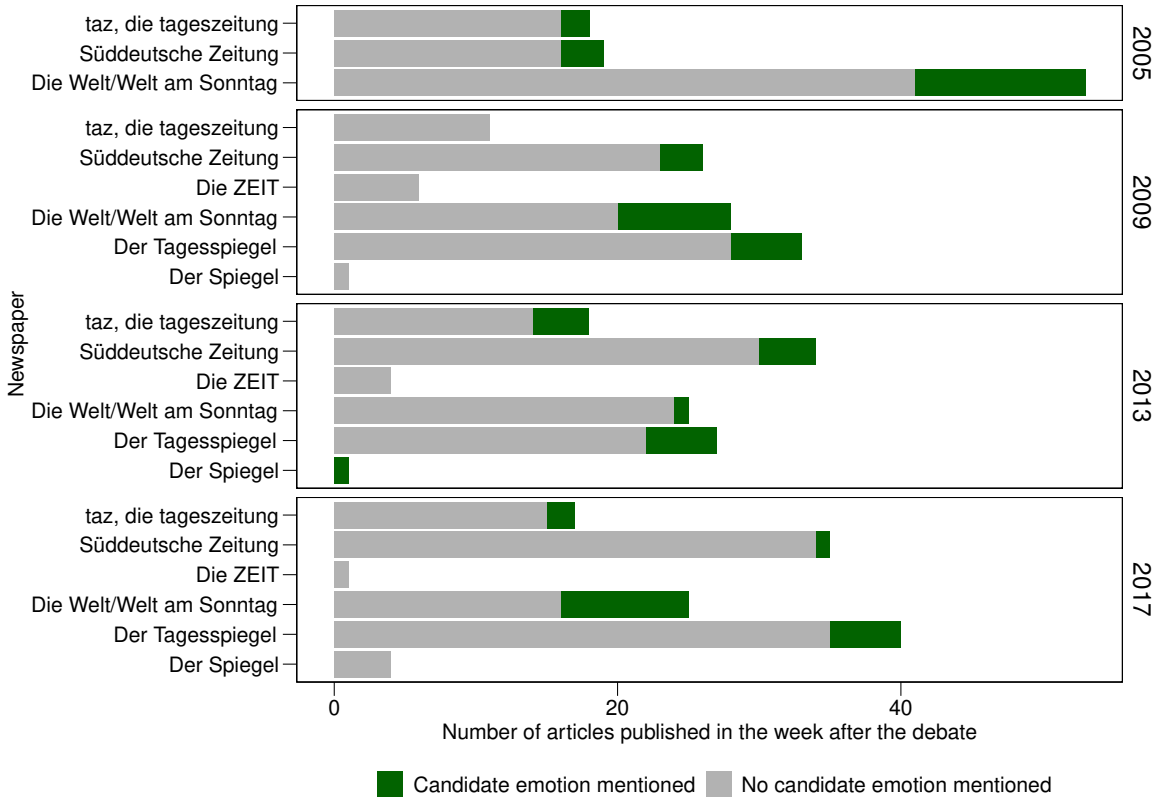


Figure A5: Number of articles per newspapers that reported on the candidates or the TV debate in the week following the debate.

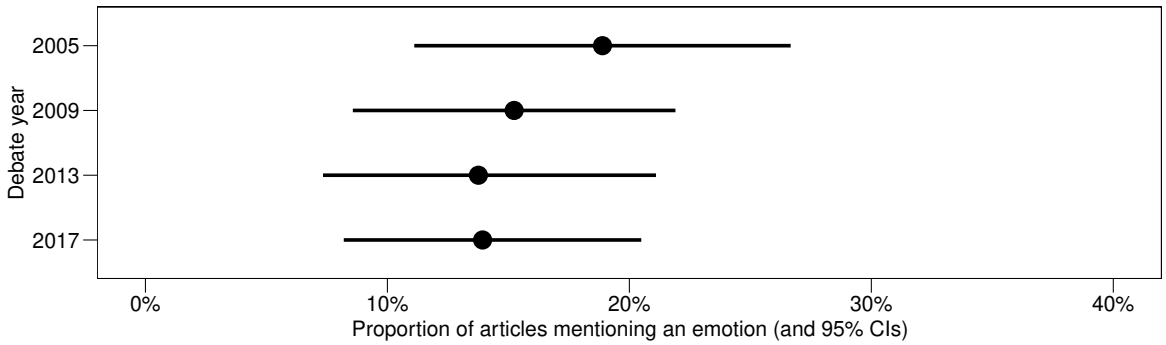


Figure A6: The proportions of news articles about candidate or the TV debate that mention at least one emotional expression by a candidate

all four debates, news outlets reported on the three types of emotions expressed by Angela Merkel. Importantly, Merkel's emotional displays are covered most often in her first debate. In 2005, the three newspapers mentioned 18 emotional displays by Merkel, which is considerably lower than in the emotions covered by six outlets in the three following debates (2009: 11; 2013: 9; 2017: 8). This finding is even more worthy of note given that we included fewer outlets for the debate in 2005.

Fourth, we also created a text corpus of the sentences that have been manually classified as descriptions of emotions. Figure A8 plots the 50 most frequent terms and multi-word expressions, along with their English translations after removing German stopwords and punctuation characters. The plot shows that the sentences clearly relate to the debate, given that the candidates are mentioned most often, along with words such as duel, chancellor, moderators, competitors, or questions. Moreover, many terms that describe emotional displays appear among the most frequent terms. Examples include appear, nervous, authentic, face, powerful, or serenity. This additional validation test underscores that the media indeed reported on emotional displays in articles about the TV debates.

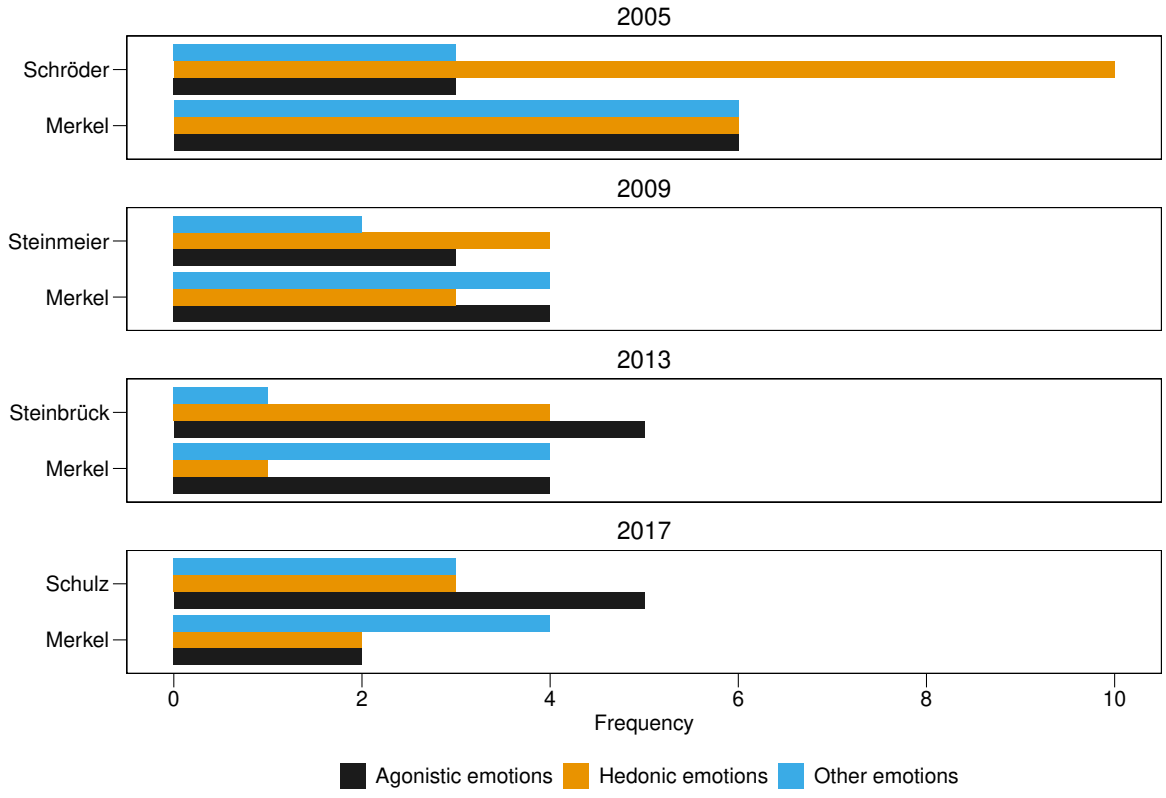


Figure A7: The frequency of specific emotions by each candidate covered in newspaper articles.

per emotional display).

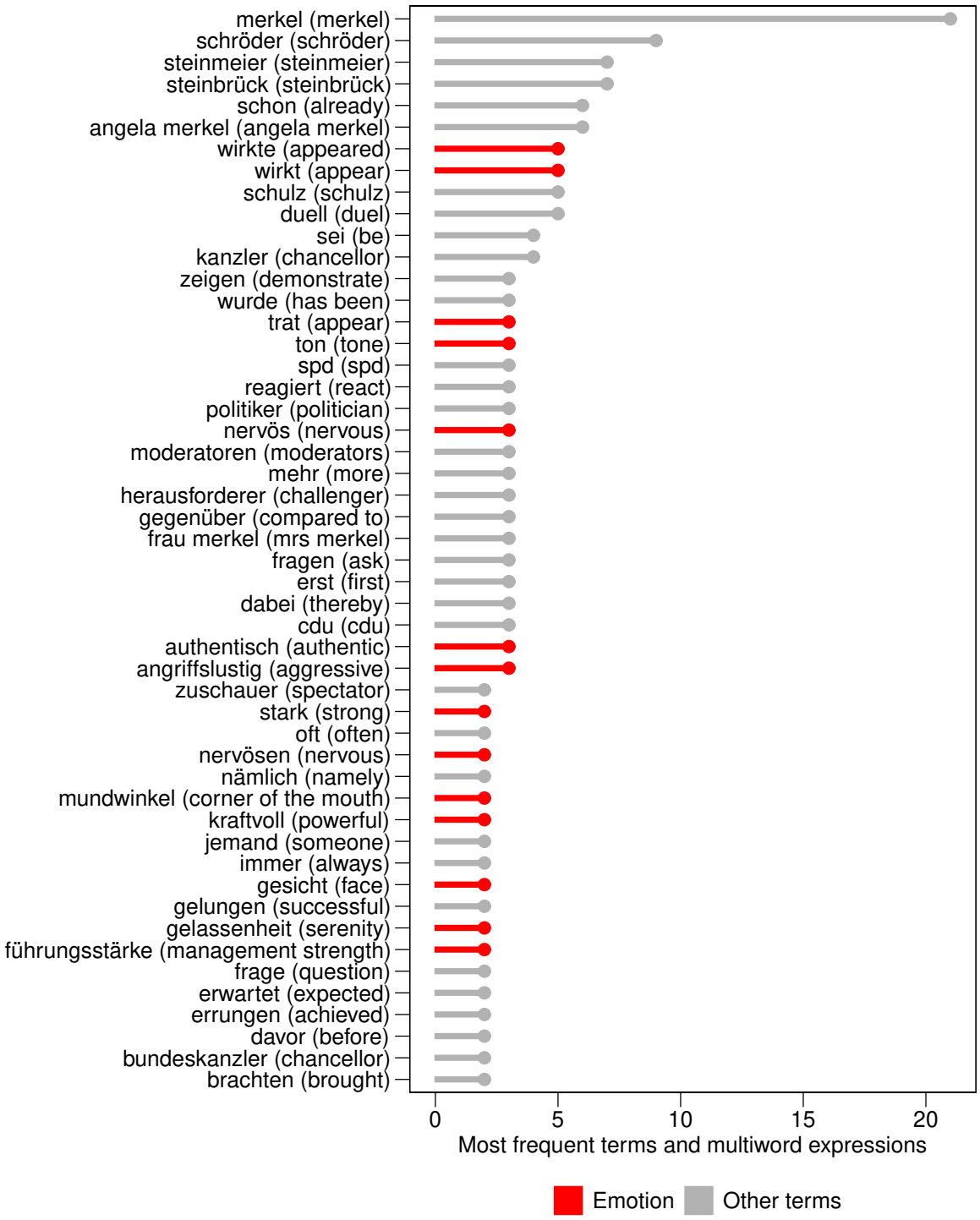


Figure A8: The most frequent terms and multi-word expressions in sentences from newspaper articles that contain descriptions of emotions.

Appendix E Ethics and Transparency

In this section, we summarize the procedures for collecting our data.

We collected labels on emotional displays by German candidates in five debates using the Face API from Microsoft Azure Cognitive Services.[17] The images of frames that we uploaded come from publicly available sources (YouTube: 2009, 2013, 2017) or C-SPAN (2005). 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. For the debate in 2005, "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 et al. 2012 838). The authors of this study shared the anonymized replication data of their paper (Nagel et al. 2012) with us in April 2020 for our study.

The data of the debates in 2009, 2013, and 2017 were collected and administered by the German Longitudinal Election Study.All datasets are freely available online at the GESIS homepage.[18] 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."[19] The experimental group was offered an allowance of 25 EUR (2013) 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.

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 (a short induction plus debate which lasted 1.5 hours). The data collection procedures are summarised in the codebooks of the following studies: (Rattinger et al. 2010; 2011a;b; 2014; 2015; 2018, Roßteutscher et al. 2019a;b;c;d;e).

[17] https://azure.microsoft.com/en-us/services/cognitive-services/face/.
[18] See https://search.gesis.org/ and https://gles-en.eu/download-data/.
[19] https://www.gesis.org/en/institute.

References

Boussalis, C. and Coan, T. G. (2020). Facing the electorate: Computational approaches to the study of nonverbal communication and voter impression formation. *Political Communication*.

Dietrich, B., Hayes, M., and O'Brien, D. Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech on Women. *American Political Science Review*, 113(4):941–962.

Faas, T. (2010). The German federal election of 2009: Sprouting coalitions, dropping social democrats. *West European Politics*, 33(4):894–903.

Faas, T. (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, T. and Klingelhöfer, T. (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.

Grabe, M. E. and Bucy, E. P. (2009). *Image Bite Politics: News and the Visual Framing of Elections*. Oxford University Press, Oxford.

Maier, J. and Faas, T. (2019). *TV-Duelle*. Springer VS, Wiesbaden.

Nagel, F., Maurer, M., and Reinemann, C. (2012). How verbal, visual, and vocal communication shape viewers' impressions of political candidates. *Journal of Communication*, 65(5):833–850.

Proksch, S.-O., Lowe, W., Wäckerle, J., and Soroka, S. N. (2019). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1):97–131.

Proksch, S.-O. and Slapin, J. B. (2006). Institutions and coalition formation: The German election of 2005. *West European Politics*, 29(3):540–559.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2010). TV-Duell-Analyse, Inhaltsanalyse TV-Duell (GLES 2009). GESIS Datenarchiv, Köln. ZA5311 Datenfile Version 1.0.0.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2011a). TV-Duell-Analyse, Befragung (GLES 2009). GESIS Datenarchiv, Köln. ZA5309 Datenfile Version 2.0.0.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2011b). TV-Duell-Analyse, Real-Time-Response-Daten (GLES 2009). GESIS Datenarchiv, Köln. ZA5310 Datenfile Version 1.1.0.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2014). TV-Duell-Analyse Real-Time-Response-Messung (Dial) (GLES 2013). GESIS Datenarchiv, Köln. ZA5711 Datenfile Version 1.0.0.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2015). TV-Duell-Analyse Befragung (GLES 2013). GESIS Datenarchiv, Köln. ZA5709 Datenfile Version 3.0.0.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2018). TV-Duell-Analyse Inhaltsanalyse

(GLES 2013). GESIS Datenarchiv, Köln. ZA5710 Datenfile Version 2.2.0.

Roberts, G. K. (2006). The German bundestag election 2005. *Parliamentary Affairs*, 59(4):668–681.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2019a). TV-Duell-Analyse, Befragung (GLES 2017). GESIS Datenarchiv, Köln. ZA6810 Datenfile Version 1.0.0.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2019b). TV-Duell-Analyse, Inhaltsanalyse TV-Duell (GLES 2017). GESIS Datenarchiv, Köln. ZA6811 Datenfile Version 1.0.0.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Brettschneider, F., Faas, T., Maier, J., and Maier, M. (2019c). TV-Duell-Analyse, Real-Time-Response-Daten TV-Duell (dial) (GLES 2017). GESIS Datenarchiv, Köln. ZA6812 Datenfile Version 1.0.0.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Faas, T., Maier, J., and Maier, M. (2019d). TV-Duell-Analyse, Inhaltsanalyse Fünfkampf (GLES 2017). GESIS Datenarchiv, Köln. ZA6829 Datenfile Version 1.0.0.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Faas, T., Maier, J., and Maier, M. (2019e). TV-Duell-Analyse, Real-Time-Response Daten Fünfkampf (dial). GESIS Datenarchiv, Köln. ZA6830 Datenfile Version 1.0.0.