Electoral Studies 86 (2023) 102700

Contents lists available at ScienceDirect

Electoral Studies

journal homepage: www.elsevier.com/locate/electstud

Candidate awareness in mixed-member electoral systems: A data-driven approach

Oliver Rittmann [a,*], Marie-Lou Sohnius [b], Thomas Gschwend [a]

[a] University of Mannheim School of Social Sciences A5, 6, 68159 Mannheim, Germany
[b] University of Oxford Nuffield College, New Road, Oxford, OX1 1NF, United Kingdom

ARTICLE INFO

Dataset link: https://doi.org/10.7910/DVN/T2XFXG

ABSTRACT

Voters need to be at least aware of candidates to hold them accountable. How does this work in mixed-member electoral systems where nominal votes often play a subordinate role and voters could entirely rely on party heuristics to choose between candidates? In lieu of existing causal explanations, we compile data on many factors contributing to candidate awareness and use a data-driven approach to identify variables that strongly predict voters' awareness of district candidates in the run-up of the German Federal Elections 2009, 2013 and 2017. We find factors that describe candidate-, voter-, and district characteristics politically to be important out-of-sample predictors in contrast to factors that describe them socio-demographically. Interestingly, we find that incumbency predicts candidate awareness, but it does not matter whether incumbents were elected nominally or via a party list. These findings can be a starting point for developing causal theories and have implications for our understanding of how voters perceive the different types of MPs a mixed-member electoral system generates.

1. Introduction

Mixed-member electoral systems create ambiguities regarding the relative importance of individual candidates and political parties in the electoral process. On one side, the majoritarian component of mixed-member electoral systems puts individual candidates in a prominent role. After all, their reelection depends on a nominal vote. To hold nominally elected candidates accountable and to compare them with other candidates, voters need to be aware of the candidates they can choose from and, in the best case, obtain sufficient information to cast an informed vote that is best in line with their interest. On the other side, given the importance of party-list votes, individual candidates often only play a subordinate role in mixed-member systems. Rather than focusing on individual candidates, voters in mixed-member systems develop partisan lenses that help them understand the political process. Yet, if voters entirely rely on party heuristics to choose between candidates, then they have little incentive to become aware of who those candidates are—party affiliation would be the only necessary information they need. This, in turn, raises questions on how voters can still hold individual candidates accountable: Is their reelection a function of their personal performance, or only a function of their party's performance?

In this study, we set out to better understand the circumstances under which voters in mixed-member systems become aware of the candidates who run in their electoral district. This is relevant for at least three reasons. First, almost all electoral systems include some nominal component. Many require voters to explicitly cast a nominal vote for one or more particular candidates. This is true for mixed-member systems, which are our interest in this study. Yet, even in some list PR systems, voters can cast a preference vote for candidates or at least can see a printed list of candidates' names next to the respective party label on a party-list ballot. Nominal components create informational demands for voters (Shugart et al., 2005). If voters want to rely on something other than the candidate's party brand, they need to possess at least some minimal information about the candidates to be able to employ additional criteria in their decision-making process. Nominal votes also create incentives to provide party-independent information for candidates. In times of partisan dealignment around the globe, party brands become less important to structure the interaction between voters and elites (Gschwend and Zittel, 2015). Supply and demand of party-independent information about candidates can, therefore, be helpful to counterbalance this development, especially during election campaigns. This might increase citizens' feelings of being represented even though the impact of partisanship is weakened.

[*] Corresponding author.
E-mail addresses: orittman@mail.uni-mannheim.de (O. Rittmann), marie-lou.sohnius@nuffield.ox.ac.uk (M.-L. Sohnius), gschwend@uni-mannheim.de (T. Gschwend).

https://doi.org/10.1016/j.electstud.2023.102700
Received 9 November 2022; Received in revised form 25 September 2023; Accepted 29 September 2023

Second, voters' knowledge about political candidates is essential for holding them accountable. According to political science's textbook understanding of democratic accountability, voters act as "rational god of vengeance and reward" (Key, 1964, p. 568). They hold politicians accountable by either re-electing a satisfactory incumbent or punishing them by voting for an opponent instead (Fiorina, 1981; Kramer, 1971). At least some minimal information about political candidates is required to assign credit and blame, especially if party brands are a less diagnostic tool for voters. Third, voters' knowledge about political candidates is important beyond normative representational concerns. Research has repeatedly shown that knowledge about candidates can shape the outcome of the district races, especially prominent in favor of incumbents rather than challengers (e.g., Elms and Sniderman, 2006; Mann and Wolfinger, 1980; Prinz, 1995).

Here, we study voters' awareness of district candidates in the run-up of federal elections in 2009, 2013, and 2017 in Germany. In the German electoral system at that time, voters cast two votes: a nominal vote for a district candidate and a party-list vote for a closed party list. Each of the 299 electoral districts sends one representative, elected via plurality rule based on the nominal vote, to the Bundestag. However, the overall composition of the Bundestag is determined by the outcome of the party-list vote. The seat share of a party in the Bundestag is proportional to its vote share based on the party-list votes. Thus, on two ballots, voters have the possibility to vote for a district candidate and a party list. Consequently, candidates have two modes of getting elected. There are district MPs who get into parliament because they win their district race. There are also list MPs who get into parliament because they ranked high enough on the party list, even though they potentially run unsuccessfully in a district as well. For the nominal vote, voters can either rely on a party heuristic, which does not require them to gather any information about the local candidates. Or, they can incorporate information about the local candidates in their decision-making process. This presupposes that they become aware of the candidates who run in their district. Without minimal candidate awareness, nothing local campaigns do should matter to citizens.

While candidate awareness has been thoroughly studied in the context of majoritarian elections of the US Congress (e.g., Cain et al., 1984; Elms and Sniderman, 2006; Parker, 1981; Prinz, 1995), we know little about candidate awareness in mixed-member electoral systems. Our goal is to understand better the circumstances under which voters are aware of the candidates who run in their electoral district in mixed-member electoral systems from a holistic perspective.

Following the comparative behavior literature (e.g., Holmberg, 2009; Pattie and Johnston, 2004), we measure candidate awareness using a free name recall item in pre-election surveys. Voters are asked whether they can name one or more candidates who run in their electoral district and the party they run for. We perceive voters' ability to freely recall candidates' names as useful to better understand the relationship between voters and candidates during the electoral process within mixed-member systems. Candidate awareness is a latent concept that helps voters to recognize candidates from a list of names, for instance, as written on a ballot, to assign a rating score to the candidates, or to recall the candidates' names. Free recall of candidate names requires a higher level of information attainment than recognizing them on a list or being able to assign them feeling thermometer scores, which is why it can be considered a rather conservative test of candidate awareness.

To correctly assess the scope conditions of our study, it is important to differentiate between being aware of candidates and having knowledge about candidates. Being aware of candidates' names does not imply that voters also possess information about the candidate: we do not necessarily expect that all voters who correctly recall candidates' names (and party affiliations) have a profound understanding of the candidates' personality and what they stand for politically nor can they easily retrieve this information from long-term memory. Still, we conceive candidate awareness as a first step towards gathering relevant information about the candidates. If voters are unaware of a candidate's name before an election, they are unlikely to have engaged with their campaign. Consequently, they cannot attribute the issues, themes, or appeals made during a campaign to any candidates. Thus, unaware voters are unlikely to hold any information about local candidates they can rely on in their decision-making process.

To learn about the factors contributing to candidate awareness, we pursue a data-driven approach and identify variables that strongly predict candidate knowledge in out-of-sample predictions. We consider three sets of explanatory variables: Candidate-level characteristics, voter-level characteristics, and district-level characteristics. To assess the predictive power of different predictors, we compile a new dataset of voter-candidate dyads based on pre-election surveys from three recent federal elections in Germany (2009, 2013, 2017). We match these voter surveys with detailed information about respondents' electoral districts, the district candidates on the respondent's ballot, and the candidates' prior political careers. We then divide this data into training and test data and use the training data to train a random forest ensemble predicting whether survey respondents can recall the names and parties of the candidates who compete in their electoral district (Breiman, 2001b). Finally, we evaluate the performance of the trained model out-of-sample on the hold-out test data.

Our contribution is to identify which conceivable explanatory factors matter empirically so that scholars can start developing parsimonious causal theories based on them that future research can test. One theme that runs through our results is that variables that describe candidates, voters, and districts *politically* hold valuable information for predicting candidate awareness. In contrast, variables that describe candidates, voters, and districts *socio-demographically* seem to have little value for predicting candidate awareness. Our findings have important implications not only for future research on the determinants of candidate knowledge and the development of causal explanations but also for our understanding of MMP systems and the two types of MPs they generate (Klingemann and Wessels, 2001; Manow, 2015; Stratmann and Baur, 2002; Zittel and Gschwend, 2008). It seems that, based on the German data we analyze in this study, district MPs and list MPs are perceived less differently than often assumed.

2. What does potentially explain candidate awareness?

We are interested in factors that explain voter awareness of local district candidates. Factors contributing to candidate knowledge are as manifold as research on the topic. Much work has been devoted to individual factors explaining candidate awareness, such as the effects of campaign spending (Coleman and Manna, 2000), the type of campaign (Gschwend and Zittel, 2015), online advertisement (Broockman and Green, 2014), or the type of election (Parker, 1981). Yet, we know little about the relative importance of causes determining whether voters recall the candidates' names on their ballot.

Giebler and Weßels (2017) were among the first to analyze determinants of candidate awareness simultaneously. In their work, they differentiate between three explanatory blocks: Candidate-related factors, voter-related factors, and context-related factors. Even though we consider a different set of explanatory variables, we consider this a useful theoretical frame for studying candidate awareness. Specifically, the frame allows us to speak to a broader debate about the importance of district candidates for local representation.

The focus on candidate-level explanations allows us to study whether candidates who are district incumbents (i.e., district MPs) get recalled easier among their district electorate than other candidates and, specifically, than list incumbents, i.e., candidates who gained their seat through the party list after they lost the district race the last time (i.e., list MPs). Much research argues (e.g., Klingemann and Wessels, 2001; Manow, 2015; Sieberer, 2010; Stratmann and Baur, 2002; Zittel and Gschwend, 2008) that there are two classes of incumbent MPs in the German Bundestag: the more local district MPs, and list MPs.

The former are usually seen as strong representatives of local district interests, while the latter are argued to represent the party interest. Thus, if this apparent division of labor also exists in the eyes of the voters, we should expect that district MPs are better known than list MPs. Both types of incumbents should be better known than non-incumbents. If the difference between these two types of incumbents has no explanatory power to predict whether they are known among voters, then the value of a local district mandate, that district MPs exclusively represent local interests, has to be questioned.

The theoretical frame also opens the door to analyzing whether specific contextual factors that pertain to the electoral district level, such as the geographic size of electoral districts, play a significant role in the personal link between voters and their local representatives. If, for example, the size of electoral districts helps explain candidate awareness in a way where voters in larger districts are less likely to know their candidate, then this would indicate that increasing the size of electoral districts could be harmful to the quality of local personal representation (Sohnius et al., 2022).

2.1. Candidate-level explanations

The first set of explanatory factors for candidate awareness focuses on the candidates themselves. This group of potential predictors is motivated by the idea that candidates can have specific attributes that make them more or less frequently recalled among the electorate. We consider six factors that we group into aspects related to candidates' political careers and personal characteristics. A first expectation is that candidates' publicity should grow with their success in the district. It seems plausible that a candidate who receives 40% of the votes is more well-known than a candidate who receives 5% of the votes. Closely connected to this idea is that there may be two front-runners in a district race who are viable to get the majority of all votes and that the public attention is focused on those front-running candidates.

Second, we consider party affiliation as a potential predictor of candidate awareness. The majority of candidate votes (*Erststimme*) in Germany are won by the Christian Democratic Party group (CDU/CSU) and the Social Democrats (SPD). Furthermore, there is variation in the emphasis that different party groups put on local representation. For example, the CSU is known to emphasize the importance of district candidates for local representation. While this is plausibly a function of the number of district MPs a party has in their parliamentary group, it is reasonable to expect that voters should more likely recall candidates of parties usually winning electoral district races and of parties emphasizing the role of district MPs.

A third predictor that naturally comes to mind is the incumbency status of candidates. Previous studies have found that incumbents are more well-known among voters than candidates who do not already hold office (Cain et al., 1984; Kam and Zechmeister, 2013; Parker, 1981; Prinz, 1995). However, the mixed-member system of the German Bundestag creates two different types of incumbents and, therefore, renders a binary classification of candidates into incumbents and non-incumbents insufficient. Following the two-vote principle, mixed-member systems open two ways of political representation. First, citizens can elect candidates to parliament via the candidate vote. They are nominally elected with a plurality of votes within their electoral district. Second, voters can help elect candidates into office with their party vote (*Zweitstimme*), with which they vote for a party list. If candidates are ranked high enough, they get elected as list MP even though they have potentially lost their district race. Importantly, candidacy in mixed-member systems is not always mutually exclusive. This creates a strategic incentive for legislators to pursue a dual candidacy and run concurrently in an electoral district and on a party list. A dual candidacy maximizes candidates' chances of getting elected. The list candidacy offers district candidates a fallback option if they lose the district vote. In consecutive elections, legislators who lost their district vote but gained a seat through the party list regularly rerun in the district race.

The German electoral system thus creates two types of incumbents in district races: Candidates who won the district race at the previous election, i.e., district MPs, and candidates who lost the district race at the last election but gained a seat through the party list, i.e., list MPs. Previous research considered only the first type as incumbent (Giebler and Weßels, 2017), but this is likely to be an oversimplified conceptualization. There may be different normative expectations for members of the Bundestag who won their seat through the candidate vote as opposed to those who won the seat through the party list. But despite the normative difference in their roles, the mandates do not differ. Thus, we are particularly interested in whether voters are even more aware of district incumbents than list incumbents.

2.2. Voter-level explanations

The second set of explanatory factors focuses on voter characteristics. This group of variables reflects the notion that voters differ from one another and that those differences may make them more or less aware of district candidates. We can roughly group this set of explanations into factors related to the political identity of voters, including their, for example, political interest and ideological leaning; and socioeconomic factors. One set of explanatory factors in the political domain describes the general relationship between a voter and the democratic system. These factors include voters' political interest, whether they are satisfied with the democratic system, their political knowledge about how the German electoral system works, whether they think that local representation is important in the German political system, and whether they intend to turn out to vote in the upcoming election. This set of predictors is motivated by the assumption that voters with a more positive relationship to the political system should also be more informed about the actors in that system (Grönlund and Milner, 2006). Therefore, variables related to citizens' relationship with the political system may help to predict their awareness of candidates. Beyond the general relationship between voters and the political system, we expect information about their ideological identity to help predict candidate awareness. Voters should be more likely to know the candidate they intend to vote for, even more so if it is the candidate of a party they identify with. The information conveyed in this set of explanations should help predict whether voters recall the name of local candidates in general and predict *which* candidate they recall.

We also include variables that indicate whether voters recall being in contact with the electoral campaign by either one party or a district candidate. Campaign contact, for example, includes exposure to television or newspaper advertisements, rallies, remembering campaign posters, or being targeted by phone and door-to-door canvassing. We include a variable that indicates contact with the campaign of a candidate's party and a variable indicating contact with the personal campaign of the local district candidate (Gschwend and Zittel, 2015). Including these variables speaks to the idea that campaigning efforts should help candidates increase voters' awareness (Broockman and Green, 2014; Pattie and Johnston, 2004).[1] We expect voters who recall contact with the campaign of a district candidate to be more likely to recall that candidate's name.

As the third set of voter-level explanations, we consider socioeconomic factors. These include age (different cohorts are more or less invested in politics, Frazer and Macdonald, 2003), gender (previous research found differences in political knowledge between men and

[1] Ideally, we would include not only a variable that relies on voters' self-reported contact with electoral campaigns but also a measure of campaigning activities, e.g., campaign spending of local candidates. Unfortunately, a respective measure is unavailable for all district candidates in German Federal Elections.

women—even though this is now debated, Kraft and Dolan, 2022), education (higher education may be correlated with higher political knowledge, Grönlund and Milner, 2006), and voters' economic situation (voters who experience financial hardship may have less capacity to engage with politics, Schaub, 2021).

2.3. District-level explanations

Finally, we consider a set of district-level explanations. Here we have a particular interest in district characteristics. One group of predictors in this domain focuses on the general features of the electoral district. The set of features includes the population size, geographic size, and population density in the electoral district. A larger electoral district, in population or area, may make it more difficult for candidates to make themselves prominent in the district. Population density reflects the notion that it may not be the number of voters in a district that makes it more difficult for candidates to become well known but the concentration of voters within the district. In the recent past, the German Bundestag discussed proposals for a reform of the electoral system that foresaw a reduction of the number of electoral districts, which would have led to larger electoral districts, on average. In light of this discussion, questions about the connection between the sizes of electoral districts and the quality of the link between local representatives and voters in Germany recently received increased public and scholarly attention (Sohnius et al., 2022; Gschwend et al., 2023). If the population size of electoral districts is predictive of voters' candidate awareness in a way that voters are less likely to be aware of district candidates in larger electoral districts, then this could be a negative unintended side-effect of an electoral reform that leads to on average, larger electoral districts.

The second group of district-level predictors concerns the political competition in the electoral district. Here, we consider the district race's competitiveness and the district's (effective) number of competitors. We also consider the potential consequences of a reform of the electoral districts in 1998 that led to a restructuring of multiple electoral districts. This disruption of the local political landscapes may have affected the relationship between voters and district candidates and made it harder for voters in these districts to recall the candidates' names.

3. Data & Methods

To gain insights about factors that matter for candidate awareness, we compile a dataset based on pre-election surveys from three federal elections in Germany in 2017, 2013, and 2009 (GLES, 2019a,b,c). Within these surveys, respondents were asked whether they could spontaneously recall the names and parties of candidates running in their local district at the federal election.[2] This is the traditional item of how candidate awareness is measured in the comparative political behavior literature (e.g., Holmberg, 2009; Pattie and Johnston, 2004). The item does not provide respondents with any assistance, they need to recall the names of their district candidates and their party affiliation solely from their memory. Using this item, we code our dichotomous dependent variable to indicate whether the voter could correctly recall the specific candidate's name and party (= 1) or not (= 0).[3] In our data, 54.9% of the respondents are aware of at least one local candidate. Candidate awareness differs vastly by the party. Historically, the vast majority of electoral districts were either won by the Christian Democratic party group (CDU/CSU) or by the Social Democrats (SPD). This is reflected in much higher recall rates for candidates of those parties than other parties' candidates. The most widely recalled candidates come from the CSU in Bavaria: Every second survey respondent in Bavaria (52.3%) correctly recalled the name of the CSU candidate in their district. This rate is lower for CDU and SPD candidates. About one out of three survey respondents correctly recall the local candidate of the CDU (36.2%) and the SPD (34.4%), respectively. These numbers drop for the district candidates of other parties. Only one out of four survey respondents (26.0%) correctly recall at least one candidate running for another party.

To measure all predictors, we compile data from various sources. We match information from these voter surveys with detailed information from the German Federal Elections Officer about respondents' electoral districts, including population size and geographic size. Finally, we add information about the respondents' district candidates, including information about the district candidates' demographic characteristics and their prior political careers. The resulting data set includes 33,868 unique voter-candidate dyads, with 6355 unique respondents and 3104 unique district candidates. As every electoral district in every election has a different number of candidates running, each voter has a differently sized choice set. The number of voter-candidate dyads thus depends on the number of running candidates within the respondent's electoral district and varies within and between districts across time.

All predictor variables are listed and summarized in Table 1. The first set of predictor variables describes the candidates themselves. To measure candidates' success, we use their vote share in the current district race. It is important to note that this measure is only realized *after* the election and thus cannot be used to predict candidate recall in the future. We still prefer it to pre-election measures because survey-based measures of voting intentions are not sufficiently accurate on the district level. To measure incumbency status, we differentiate between three incumbency types: First, non-incumbents are candidates who do not hold a mandate in the Bundestag; second, district incumbents are candidates who have won the district in the previous election (district MP); third, list incumbents are members of the Bundestag who gained their seat via the party list (list MP). Importantly, the vast majority (95.3%) of the list MPs ran unsuccessfully in the district before. All other candidate-level predictors are straightforward to measure: "Party" is a categorical predictor that denotes the party of a candidate, "Frontrunner" indicates whether a candidate was among the top 2 candidates in the current district race, "Age" measures the age of the candidate in the election year, and "Female" indicates whether the candidate identifies as a woman.

The second set of predictors describes voters. Here, we draw on a battery of survey items to measure political interest, whether a respondent identifies with the party of the district candidate, whether they reported voting for the candidate,[4] whether they recall contact with the campaign of the candidate's party or contact to personalized campaign of the candidate, whether they are satisfied with how democracy works in Germany, whether they intend to turn out to vote at the upcoming election, whether they think that local representation is

[2] Interviewer instructions specified that candidate names that were not completely correct had been nevertheless coded as correct. The exact wording of this question is documented in Appendix B of the supplementary material. Note, there are a few pure list candidates at each election, i.e., candidates that do not run in any electoral districts. Respondents are not asked about them.

[3] We code respondents as '0' who could recall a candidate's name but confuse the party or are unable to name a party at all. However, this does not happen very often and, thus, obviously requires only a slightly higher awareness level. In 2013 and 2017, conditional on correctly recalling a candidate name, respondents in our data were able to recall the correct party in about 92% of the cases. In 3.6% of the cases, those survey respondents recalled an incorrect party; in 4.4%, they did not recall any party. The survey data from the 2009 pre-election survey records the correct candidate recall binary, indicating only whether respondents recalled the candidate together with the correct party or not.

[4] To express the intent to vote for a candidate, respondents did not need to recall the name of the candidate. It was sufficient if they reported using their district vote to vote for the candidate of a specific party.

Table 1
Overview of all variables used to predict district candidate awareness during the 2009, ’13, and ’17 federal elections in Germany (before imputation).

Predictor variable	Range/Categories	Mean	Description
Candidate-level predictors			
Voteshare at t	$[0.01, 0.66]$	0.18	Realized district vote share of the candidate at the upcoming election.
Party	$\{$CDU, … , AfD$\}$	–	Party of the candidate.
Status at $t-1$	{New Candidate, District Incumbent, List Incumbent}	0.69 0.15 0.16	Incumbency Status in the previous legislative period. *New Candidate:* Did not run before; *District Incumbent:* District winner of previous election (district MP); *List Incumbent:* list MP, but (mostly) lost in district at the previous election.
Frontrunner	$\{0, 1\}$	0.38	Is the candidate among the top-2 candidates in the current district race?
Age	$[18, 78]$	47.83	Age of the candidate in the election year (Election year − Year of birth).
Female	$\{0, 1\}$	0.28	Does the candidate identify as a woman?
Voter-level predictors			
Political interest	$[1, 5]$	3.00	Self-reported level of political interest from high (1) to low (5).
Party identification	$\{0, 1\}$	0.13	Does the respondent identify with the party of the candidate?
Voted for candidate	$\{0, 1\}$	0.13	Did the respondent intend to vote for the candidate?
Party contact	$\{0, 1\}$	0.57	Does the respondent recall contact with the election campaign of the candidate’s party?
Candidate contact	$\{0, 1\}$	0.26	Does the respondent recall contact with the election campaign of the candidate?
Satisfaction with democracy	$[1, 5]$	2.74	Self-reported satisfaction with democracy in Germany from high (1) to low (5).
Turnout intention	$\{0, 1\}$	0.80	Self-reported intention to turn out to vote in the upcoming election.
Local representation important	$[1, 5]$	2.10	Agreement with “The MP should represent all citizens in the electoral district” from high (1) to low (5).
Political knowledge	$\{0, 1\}$	0.48	Respondent correctly answered “Which vote decides how many seats each party will have in parliament?”
Age	$[16, 99]$	52.36	Age of the respondent in election year (Election year − Year of birth).
Female	$\{0, 1\}$	0.50	Does the respondent identify as a woman?
High school	$\{0, 1\}$	0.29	Does the respondent hold a high school degree (*Abitur* or *Hochschulreife*)?
Subjective economic situation	$[1, 5]$	2.60	Self-reported satisfaction with own economic situation from high (1) to low (5).
District-level explanations			
Population size	$[197.6, 377.4]$	275.06	Population size of the electoral district in 1000.
Geographic size	$[26.9, 6250.3]$	1351.33	Geographic size of the electoral district in km^2.
Population density	$[0.04, 12.63]$	0.85	Population density of the electoral district ($\frac{\text{Population Size}}{\text{Geogr. Size}}$).
Effective number of candidates	$[2.17, 5.82]$	3.68	Effective number of candidates in the electoral district ($\frac{1}{N}\sum_{i=1}^{N} p_i^2$).
Winning margin	$[0.00, 0.51]$	0.14	Difference in vote shares between district winner and second-placed candidate.
Electorate change	$\{0, 1\}$	0.15	≥50% of the district’s electorate changed through 1998 electoral district reform.

important, whether they are knowledgeable about the electoral system in Germany, and their age, gender, education,[5] and subjective economic situation. The scales of all variables, as well as the question items for each variable, are summarized in Table 1.

The final set of predictors describes the respondents’ electoral districts. Here, we include population size, geographic size, and population density of the district. Two additional variables describe the electoral competition in the district. As a measure of the number of competitors, we include the effective number of candidates (Laakso and Taagepera, 1979). As a measure of the competitiveness of the district race, we include the winning margin of the district winner (e.g., Gschwend, 2007). Finally, we include a variable indicating whether at least 50% of the electorate changed due to an electoral district reform in 1998. This applies to 30 electoral districts that were most severely affected by redistricting (Eisel and Graf, 2002).

4. Predicting candidate awareness

In the previous section, we collected a wide range of variables that possibly predict voters’ candidate awareness. Our next goal is to study which variables contain information that helps us predict candidate awareness. The predominant approach to such a task involves statistical models that assume a specific stochastic process, e.g., a logistic regression model. In this framework, the probability of a voter’s candidate awareness would be modeled as a transformation of some linear combination of our independent variables. This comes down to assuming a functional form of the relationship between the independent variables and the outcome, even though this is often not true (Breiman, 2001b). Beyond that, because this classical approach treats the functional form of a regression model and the set of independent variables as known, it puts little emphasis on model evaluation (Athey and Imbens, 2019). That is, it relies on standard errors as measures of uncertainty for a specific statistical model’s parameters but rarely asks whether parameters estimated within one model enable us to predict the outcome based on new data. But if the specified model is incorrect, this uncertainty assessment has limited value. In cases with little theoretical guidance about how a set of predictor variables is related to the outcome, prespecifying a statistical model appears as a suboptimal idea: such a model requires us to make assumptions without theoretical backing and will likely lead to incorrect conclusions. Predictive modeling and machine learning offer a viable alternative. Instead of using theory to set up a statistical model, predictive modeling adopts a more inductive approach and treats the data-generating process as unknown. Instead, the data is used to determine the functional form of the relationship between the independent variables and the outcome (Molina and Garip, 2019; Grimmer et al., 2021). To assess modeling uncertainty and prevent overfitting, this approach uses a train-test set logic where a model’s predictive abilities are evaluated based on data not used during the model estimation.

In light of these arguments, scholars have increasingly turned towards machine learning models to study various contexts (see, for example, Lupu and Warner (2022) and Kim and Zilinsky (2022)). Given

[5] For education, we include a predictor indicating whether the respondent holds a high school degree (*Abitur* or *Hochschulreife*). Note that in the German context, this reflects a higher degree than a high school degree in the US.

the sparseness of the theory surrounding the explanation of candidate awareness, we adopt such a data-driven approach in this study. This approach includes splitting the data set into a train and test set, with 75% of the data going into the train set and 25% being reserved in the test set. We only use the training data to develop our prediction model and the test data to evaluate the trained model. Because the model did not see the test data during the training stage, the approach constitutes a more rigorous approach to assess the out-of-sample predictive power of the model. It ensures that drawn inferences are not a product of in-sample overfitting.

We start by imputing missing data among our predictor variables. Our data set comprises 33,868 unique voter-candidate dyads, but around 10% of the observation (3789) have missing values for at least one of the predictor variables. One particular concern is that the missingness of predictor variables is affected by factors that are of direct interest to us. For example, survey respondents with a low political interest may be less likely to answer specific survey questions. At the same time, it is plausible that our dependent variable, respondents' ability to recall the names of district candidates, is related to their willingness to answer all survey questions. If both propositions are true, listwise deletion would potentially bias our results, leading us to over or underestimate the predictive power of political interest for candidate awareness.

We use conditional multiple imputations, implemented in the R-package `mice` (van Buuren and Groothuis-Oudshoorn, 2011), to impute missing values among our predictor variables. Previous research has shown that this approach outperforms joint multivariate normal imputation when the data include missing values among categorical variables (Kropko et al., 2014). We generate ten data replicates with imputed values for all missing values among our predictor variables. Next, we split each of these replicates into train and test sets, with 75% of each replicate going into the train set and 25% being reserved in the test set. We perform all subsequent analysis steps on each of the ten data replicates.

We use random forest models to predict respondents' ability to recall a specific candidate's name (Breiman, 2001a; Montgomery and Olivella, 2018). Random forests are particularly useful for our goal. We want to highlight two advantages. First, random forests perform well even if there are only a few informative predictors among many uninformative predictors (Sandri and Zuccolotto, 2006). Second, there are well-established methods to assess the relative importance of individual predictors for the performance of random forest models, allowing researchers to learn which variables are important to predict the outcome and which are not.

The nested structure of our data gives a particular challenge for our task: Our unit of analysis are voter-candidate dyads, with the outcome variable indicating whether the respondent recalls a specific candidate name. Each respondent is paired with all relevant candidates in their electoral district. Thus, each respondent occurs usually five or six times in the dataset. The challenge arises from the fact that observations of the same respondent are not independent. Standard random forests do not account for such dependencies that arise in clustered data structures.

To address this issue, we implement a two-stage respondent-level bootstrap procedure that breaks the clustered structure of our data before fitting random forests.[6] In the first stage, we draw a bootstrapped sample of respondents from the training data. In the second stage, we sample one observation of each respondent sampled in the first stage. This results in a bootstrapped sample where the number of observations amounts to the number of respondents (rather than a sample where the number of observations amounts to the number of respondents times the number of candidates' observations). Given this random procedure, we can assume that the observations within each bootstrapped sample are independent. We use the resulting sample to fit a random forest model and repeat the procedure 50 times.[7] This results in 50 bootstrapped samples and an ensemble of 50 random forests for each imputation replicate. To predict candidate awareness for new observations, we average across the predicted probabilities of the 50 random forests that constitute one ensemble. Figure A1 in the supplementary material illustrates the procedure.

Table 2
Random forest ensemble evaluation. Note: To calculate evaluation scores, we applied each random forest ensemble to the hold-out test set of the imputation replicate it was trained on. This results in ten sets of evaluation statistics, one per imputation replicate. The evaluation scores in the table show the average of those statistics together with the standard deviation (in parentheses). The naive model predicts the modal category (voter does not know the candidate) and serves as a benchmark for the random forest model.

Measure	Naive model	Random forest
Percentage of correctly predicted	0.788	0.839 (0.001)
Sensitivity (true-positive rate)	0.000	0.715 (0.007)
Specificity (true-negative rate)	1.000	0.854 (0.001)

After training, we apply the ensembles of random forests to the hold-out test sets of each data replicate and calculate how well it predicts candidate awareness. Table 2 shows the out-of-sample performance scores and benchmarks them against a naive model that predicts the modal category for each observation, i.e., any voter does not know any candidate. The trained random forest ensembles perform adequately on the hold-out test sets: For more than eight out of ten respondent-candidate pairs (83.9%), the ensemble correctly predicts whether the respondent recalls the candidate name. The model is better able to correctly predict candidate awareness for voter-candidate pairs where the voter does not recall the candidate (85.4%) than when the voter *does* recall the candidate (71.5%). Overall, these results indicate that our set of predictor variables indeed stores information that is predictive of voters' awareness of district candidates in the run-up to the federal elections in Germany.

4.1. Which factors matter most for the prediction of candidate knowledge?

In the next step, we are interested in which of our predictive variables are most important for the model's ability to predict candidate awareness and which variables are least important. That is, we want to learn about the relevant factors for predicting candidate awareness. Fig. 1 presents helpful quantities of interest to infer this—variable importance measures for each predictive variable in the trained random forest ensemble. The scores represent the rate by which the percentage of incorrectly predicted voter-candidate pairs increases when the information stored in one variable is taken from the model.[8] If this number

[6] Accounting for clustered data structure within random forest is an active field of research (Karpievitch et al., 2009; Adler et al., 2011; Hajjem et al., 2014; Pellagatti et al., 2021). We provide an overview of this literature in appendix A of the supplementary material and lay out how our approach relates to the solutions proposed in this literature.

[7] Three hyperparameters of the random forest model are tuned using random search (Bergstra and Bengio, 2012). Specifically, the search space contains the number of trees ($\mathtt{n}_{\mathtt{trees}} \in [20, 100]$), the number of randomly sampled variables used as candidates at each split ($\mathtt{m}_{\mathtt{try}} \in [2, 10]$), and the minimum number of observations in terminal nodes of a tree ($\mathtt{nodesize} \in [10, 50]$). Random search is performed ten times using the R-package `mlr3` (Lang et al., 2019). We select the model with the lowest classification error based on tenfold cross-validation in the test set as the best-fitting model (Neunhoeffer and Sternberg, 2019).

[8] Precisely, we take the information from the model by shuffling the values of the predictor and recalculating the classification error of the trained model.

equals one, this means that withholding the variable's information from the model does not decrease the model's predictive performance. We calculate variable importance scores based on in-sample and out-of-sample predictions. We do this because a variable may have predictive value in the data that was used to train the model, but this may be a result of overfitting. To investigate the relevance of variables for predicting candidate knowledge, it is thus essential to check whether they help predict observations that were not used to train the model.

The first result is that the trained ensembles attribute at least some importance to the entire range of predictors based on the training data. However, once we apply the models to unseen data in the test set, the set of predictors contributing to the model's predictive ability shrinks considerably. This confirms our approach and underlines the importance of evaluating predictive models out of sample, as in-sample analysis may make some variables appear important even though they are not.[9]

One theme that runs through our results is that variables that describe candidates, voters, and districts *politically* appear to hold valuable information for predicting candidate awareness, while variables that describe candidates, voters, and districts *socio-demographically* contribute less. Starting with candidate-level predictors, we find that candidates' age and gender carry little to no value in predicting candidate awareness, even though the model emphasizes candidates' age, at least to some extent, in the training data. Variables that *are* important to predict whether a candidate is recalled among voters are related to their electoral and political success. This is reflected by the fact that the realized vote shares of district candidates are the second most important predictor among all variables, and their incumbency status is the fifth most important predictor.

The results on the voter level draw a similar picture: Variables that describe voters politically seem to be valuable to predict their awareness of district candidates, while socio-demographic characteristics do not play an important role: Whether a voter had contact with a candidate or a candidate's party, their political knowledge, their vote choice, their party identification, and their political interest all score higher in variable importance than socio-demographic predictors such as age, subjective economic situation, education, and gender.

The single most important variable to predict a voter's ability to recall a candidate's name is the variable indicating whether the voter had contact with the candidate during the electoral campaign. While this finding may suggest that campaigning may be an effective tool for candidates to increase public awareness, we do not interpret this finding as causal evidence for the effectiveness of campaigns for name recall. Instead, the finding is a plausible result of two endogenous processes: First, candidates who are well-known before a campaign may have more resources for personal campaigning. Second, whether a candidate manages to make direct contact with a specific voter likely depends on the characteristics of the voter. For example, politically interested voters may be easier to reach by a campaign than politically detached voters. Both mechanisms plausibly explain why the candidate contact holds valuable information to predict candidate awareness.

Political knowledge and whether a respondent intends to vote for the candidate's party are two further informative variables to predict whether a specific voter recalls a particular candidate, followed by party contact, political interest, and identification with the candidate's party. While the relevance of those political characteristics does not come as a surprise, it is remarkable how unimportant voters' socio-demographic characteristics are for predicting their ability to recall local district candidates. Neither voters' formal education nor their age and gender seem to carry information that helps predict voters' ability to recall local district candidates. Interestingly, whether a voter intends to cast a vote or to abstain does not seem to help predict their ability to recall candidate names. Potential explanations for this may be social desirability bias (respondents feel socially pressured to indicate that they plan to cast a vote) or willingness to participate in the survey in the first place (citizens who abstain from an election may be less willing to participate in an election survey).

On the district level, we again observe a superiority of variables that describe electoral districts politically over variables that describe electoral districts demographically—even though here, variables like geographic size, population size, and population density seem to improve the model's prediction to some extent. Yet, the most important variables to predict candidate awareness within an electoral district are closely connected to the political contest in the district. That is the effective number of candidates within that district, followed by the winning margin in the district.

Another interesting result on the district level is that major disruptions of electoral districts in 1998 (after an electoral reform reduced the number of districts from 328 to 299) do not help to predict the level of candidate awareness in the elections 2009–2017. We do not want to interpret this result in the sense that reforms of electoral districts do not affect local candidate awareness. Still, the result suggests that more than ten years after the reform, there are no dramatic differences in the level of candidate awareness between districts most highly disrupted by the reform and others.

Taken together, the results suggest that voters' ability to name candidates is not the result of either the candidates' or the voters' socio-demographic characteristics. Instead, what matters for candidate awareness is a candidate's ability to prevail in the political and electoral contest, the voter's political identity, and the nature of the electoral competition within a district. While these factors help us predict candidate awareness, we should not falsely interpret the results as causal. For example, our models indicate that a candidate's electoral success is the top predictor of their prominence among voters. Still, the models provide no answer to where a candidate's electoral success comes from. It is neither able to differentiate between causal directions (are candidates electorally successful because they are prominent among voters?; or are candidates prominent among voters because they are politically successful?), nor is it able to say anything about the roots of political and electoral success, which may, for example, be a function of their party's electoral success.

4.1.1. *Direction of effects*

Our primary goal is to learn about the predictive value of a wide set of variables for voters' candidate awareness. Yet, we are also interested in whether the most important variables of the trained model influence the predictions of the random forest model in a way that is in line with what we would expect theoretically. For this purpose, we select the three most important variables of each set of predictors and investigate

[9] It is important to point out that there are two theoretical reasons why a variable may seem important in the training data but unimportant in the test data. The first reason is that the model learned patterns in the training data that are unique to the training data and, therefore, do not generalize to the test data. This is what is called overfitting, and we consider it to be the driving force behind disparities between in-sample and out-of-sample measures of variable importance. The second potential reason is that the model learns patterns that generalize beyond the training data but are not present in the test set due to data sparseness. For example, the model may learn from the training data that women between 60 and 65 with a high school degree have a high knowledge of candidates. Suppose this is a general pattern, but by chance, there are simply no women between 60 and 65 with a high school degree in the test data. In this case, the model would have learned something meaningful from the predictors gender, age, and education. Still, it would not become visible in the out-of-sample variable importance assessment. Given our 75–25 train-test split, we consider this option less likely, but we cannot rule it out. We are also fairly convinced that if the second option has an impact on our out-of-sample importance estimates, those impacts should not be so large that they change any of our substantive conclusions.

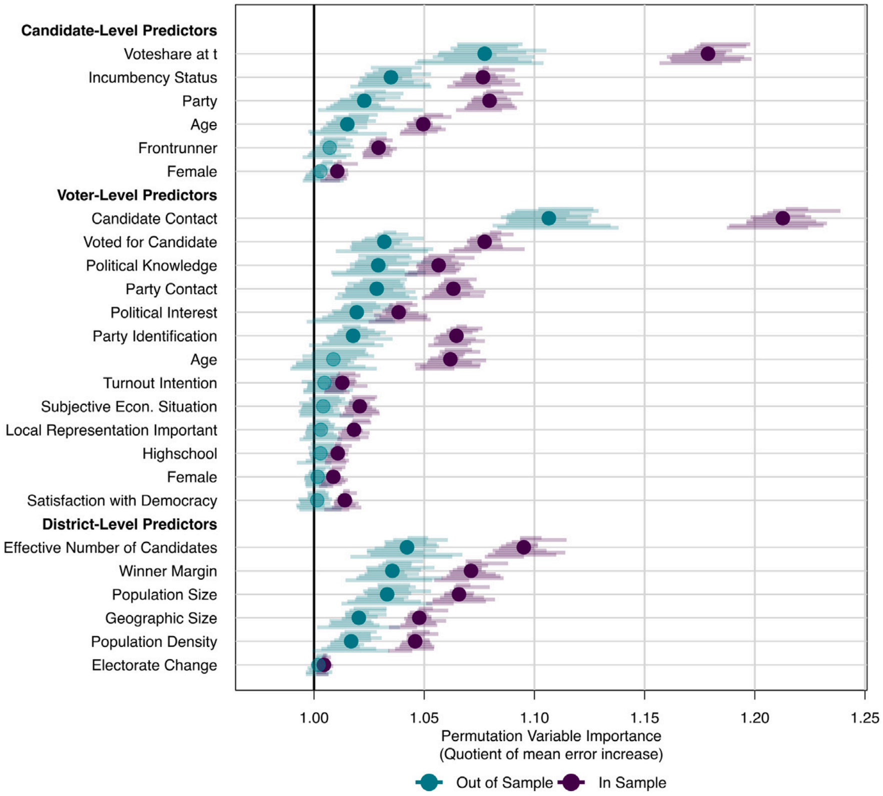

Fig. 1. Permutation variable importance of all predictors in the random forest ensembles. Note: For each imputation set, we calculate the importance score of one predictor variable by shuffling the values of the predictor and recalculating the classification error of the ensemble. This gives us an idea about how the ensemble performs if we withhold the information of the specific predictor. The more the classification error decreases, the more important the variable is for the predictive performance of the ensemble. We divide the classification error of the permutation data set by the classification error of the full data set. If this quotient is equal to one, then withholding the information of the predictor has no effect on the predictive performance of the ensemble, and the variable has no importance for the predictive power of the ensemble. Higher scores indicate higher variable importance. For each variable, we repeat the procedure 100 times and average across the results. For each imputation data set and each variable, the intervals in the figure show the center 95% of the 100 quotients. Points represent the mean variable importance across all imputation replicates. Points are depicted transparently if three or more (out of 10) of the accompanying intervals include the value one. We show variable importance scores based on the training data set (in-sample predictions) and on the test data set (out-of-sample predictions). The difference highlights the importance of keeping potential overfitting in mind when analyzing the model: A variable may seem important to predict outcomes in the training set, but this often does not generalize to the test set. To evaluate the informational value of specific variables, it is thus important to focus on their contribution to out-of-sample predictions.

how the model's average predicted probabilities change as a function of these variables. Fig. 2 shows the results of this exercise.[10]

The results are mostly consistent with what theoretical expectations would suggest. Starting with candidate-level predictors in the top row of Fig. 2, we observe that higher vote shares of a candidate and already being an incumbent in the election run-up are associated with a higher likelihood of being recalled among voters. Moreover, we observe higher predicted probabilities for candidates of the parties that traditionally win district races (CDU/CSU and SPD) compared to other parties. These results are hardly surprising but confirm that the model learned sensible relationships.

One observation on the candidate level stands out to us: While incumbency status matters in general, there seems to be almost no difference in candidate awareness with respect to the type of incumbency. Having won the district in the previous election seems to come with virtually no gain in prominence compared to candidates who lost the district and only entered the parliament via the party list. Recall that the vast majority of the list incumbents in our data ran in the district before but did not win their district race (95.3%). This implies that list incumbents were less electorally successful in the district before the election than district incumbents. Yet, by virtue of their list mandate, they seem to be almost as widely known in the district as district incumbents. In other words, district incumbents seem to enjoy no advantages over the list incumbents regarding their prominence in the electoral district.

Turning to the voter level, we find that respondents who indicated that a candidate contacted them during the campaign, and respondents who indicated that they intend to vote for a candidate are substantially more likely to be aware of this candidate. The same holds for respondents who know the German electoral system sufficiently well to realize which of their two votes is decisive for the overall composition of the parliament—but the magnitude of this effect is much smaller than the magnitude of candidate contact and vote choice.

Regarding the effective number of candidates within an electoral district, we find that a higher number of candidates is associated with lower probabilities of candidate awareness. To make sense of this result, it is helpful to remind ourselves about what unique information the effective number of candidates adds to the model that is not captured by other variables.[11] Since the model has access to a candidate's vote share, what the effective number of candidates adds is information about the number of other auspicious candidates that the candidate competes in the district race. Thus, the negative association between the effective number of candidates and candidate awareness suggests

[10] It is again important to emphasize that none of the graphs allow for a causal interpretation.

[11] After all, the effective number of candidates is a function of the vote shares of all candidates in the district, and the candidate's vote share entered the model as a separate variable.

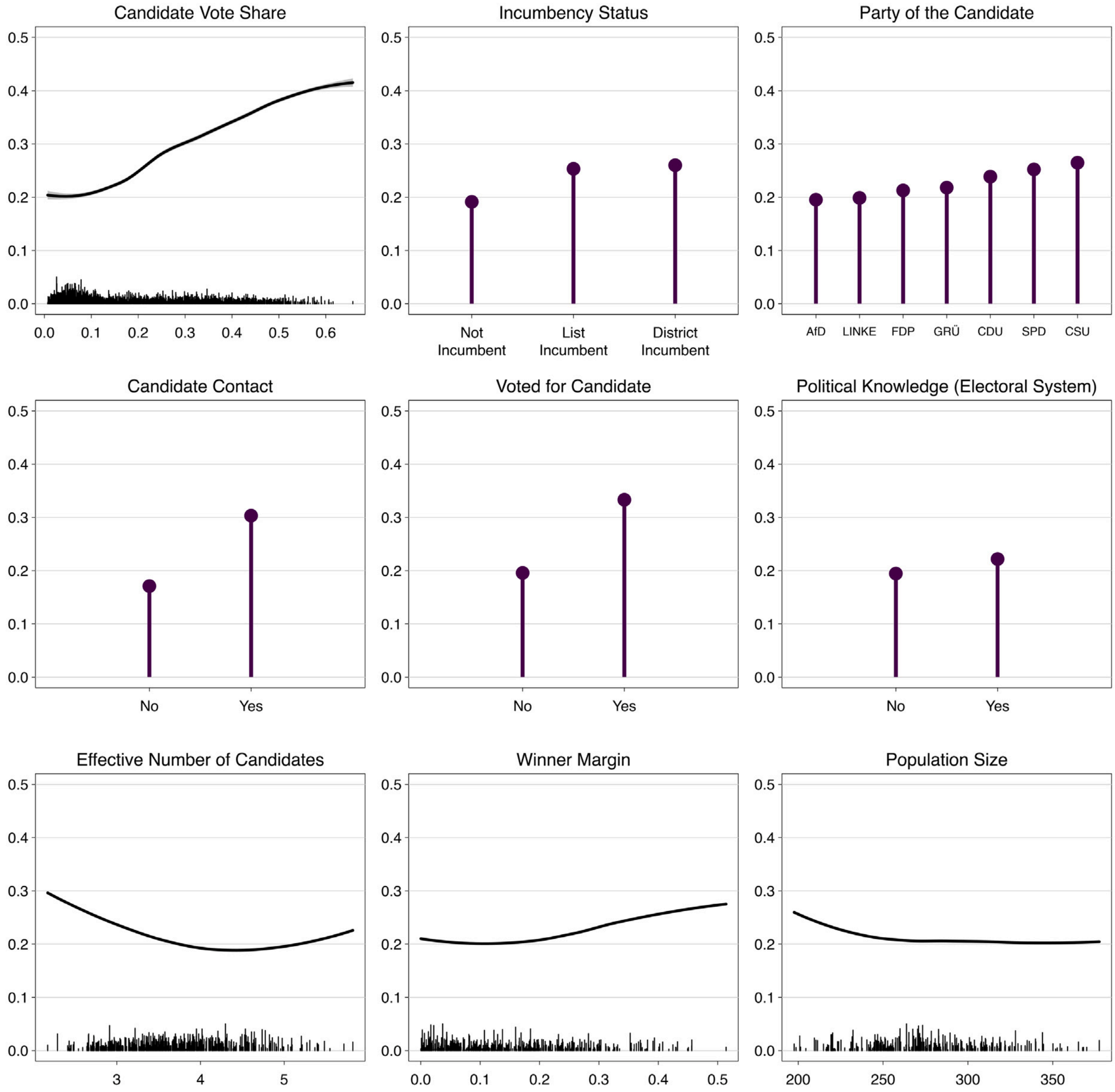

Fig. 2. Partial dependence plots of voters' candidate awareness. Note: The figure presents partial dependence plots for the top three most important variables on the candidate level (top row), voter level (center row), and district level (bottom row). x-axes represent the predictor variables, y-axes show predicted probabilities to recall the name (and party) of a district candidate by the random forest model. Predicted probabilities conditional on a specific value c of the covariate x_k are calculated by (1) creating a replicate of the predictor matrix X, replacing all observed covariate values x_k with c, (2) using the trained model to predict probabilities for each unit in the replicate matrix, and (3) averaging across all those predicted probabilities. Each panel shows how the ensembles' predicted probabilities change across the empirical range of the variables. The lines represent local polynomial regression lines fitted to the predicted probabilities based on all imputed data sets with 95%-Confidence Intervals. It is important to note that these confidence intervals *do not* quantify sampling uncertainty around the predicted probabilities. Panels with continuous predictors also show the empirical distribution of the variable in the full data set (train and test set) at the bottom of the panels.

that a candidate's chances of being recalled decrease the more (serious) competitors they face. This may indicate that voters have a limited capacity to recall candidates' names and are overwhelmed when there are four or five equally promising competitors for a seat in their district.

Finally, we gather little information from the partial dependencies on the remaining two district-level variables. Candidate awareness increases when there is a very high winning margin (30 percentage points), but this does not happen very often and is thus based on relatively few observations. The partial dependency plot for population size suggests that voters in districts with exceptionally small population sizes are more likely to name candidates. Beyond the districts with exceptionally small population sizes, there is no clear trend observable, but the variable's predictive value may stem from interactions with other variables that are not visible in the aggregate.

4.2. Awareness of candidates who are already MPs

The previous analysis investigated the predictability of the awareness of *all* candidates running for any of the parties that made it into parliament in the respective election year. Next, we now focus only on candidates who are already MPs. Notably, this does not only include district incumbents but also list incumbents.[12] We do this for

[12] Note that not all list incumbents of the *Bundestag* run in an electoral district, but most of them do. Our analysis does not consider candidates who exclusively run on a party list, as they lack a connection to a specific electoral district. District candidates who hold a list mandate are candidates who lost the district race in the previous election but entered the Bundestag via their party's list.

Table 3
Random forest ensemble evaluation. Note: To calculate evaluation scores, we applied each random forest ensemble to the hold-out test set of the imputation replicates it was trained on. This results in ten sets of evaluation statistics, one per imputation replicate. The evaluation scores in the table show the average of those statistics together with the standard deviation (in parentheses). The naive model predicts the modal category (voter does not know any incumbent) and serves as a benchmark for the random forest model.

Measure	Naive model	Random forest
Percentage of correctly predicted	0.633	0.757 (0.003)
Sensitivity (true-positive rate)	0.000	0.709 (0.006)
Specificity (true-negative rate)	1.000	0.776 (0.002)

two reasons. First, electoral competition at the district level always includes candidates without a real chance of winning the district seat. This might be an unfair comparison. Our results show that voters tend to be more aware of incumbents than of non-incumbents. There may be characteristics of incumbent candidates that may matter for voters' awareness but that do not become visible when analyzed together in a pool with many non-incumbent candidates of whom voters are rarely aware. Second, the focus on incumbents allows us to investigate further one of the most interesting results of the prior analysis: this is that incumbency helps predict candidate awareness, but it seems like there is no difference between list and district incumbents. This poses the question of whether the type of incumbency conveys any helpful information to predict candidate awareness in the subset of incumbency candidates.

Thus, our first goal of this analysis is to investigate candidate awareness among the group of comparable candidates, namely incumbents. Second, our focus on incumbents only provides a hard test of the remarkable finding from the previous section that voters are not more aware of district incumbents than list incumbents. Formally, only district incumbents are the elected representatives of an electoral district. Thus, one could argue that district incumbents should be more widely recalled in their district than list incumbents. Nevertheless, list incumbents may be connected to the district because they competed unsuccessfully in it before but now do similar service work in this district even though their mandate is not formally tied to the electoral district.

The subset of voter-incumbent candidate pairs comprises 10,329 observations and is thus substantially smaller than the full data set. Incumbents were known in about every third voter-candidate pair (36.7%). We keep the split between train and test data from before and train ensembles of forest models for imputation replicates following the same procedure as above. Table 3 shows the out-of-sample performance of the random forest ensembles trained on the subset of candidates who are already MPs. The trained random forest model is able to correctly predict incumbent awareness in about three out of four voter-incumbent pairs (75.7%), substantially improving predictive accuracy compared to the naive baseline model. Given the more balanced sample, it is no surprise that the Sensitivity-Specificity difference of the incumbent model is much less pronounced than in the full data model. Our incumbent model is only slightly better able to predict the lack of knowledge among those who do not recall a candidate (77.6%) than the knowledge of candidates among those who recall the candidate (70.9%).

Fig. 3 shows variable importance measures of the random forest model trained on the subset of candidates who are already incumbents. The variables that have been most important previously to predict candidate awareness in the full data set using all candidates (in Fig. 1) remain largely the same when focusing only on district and list incumbents as candidates. Among the most important predictors in the model are still candidate contact, candidate vote share, the winning margin of the district winner, and the effective number of candidates. At the same time, the model confirms the finding that the socioeconomic characteristics of both candidates and voters seem to carry little information that helps us predict candidate awareness, even though candidates' and voters' age seem to have at least some minimum level of information to predict voters' awareness in the subset of incumbent candidates.

Remarkably, the incumbency status that differentiates between list and district incumbents slipped down to the variables that have virtually no value for the predictive performance of the models. The average predicted probability of being aware of a candidate conditional on incumbency type increases by only 1.7 percentage points, from 36.1% to 37.8%, for district incumbents compared to list incumbents. Together with the previous results, this questions whether voters see a unique value of being a district MP for being known within the electoral district, compared to a list MP. Instead this result indicates that what matters for candidate awareness is whether candidates hold a mandate in the parliament, but not how they got there — through winning their district or through their party list.

Fig. 4 presents partial dependencies of the four most important predictors of our incumbent model, confirming the findings from the full model. Recalling being contacted by a candidate and a higher incumbent vote share are associated with a higher likelihood of being aware of the incumbent. More competitors within a district are associated with a lower likelihood of recalling an incumbent's name within that district. Finally, incumbents in districts in which one candidate is far more successful than all the other candidates seem to be easier recalled than incumbents in more competitive districts. Still, this association only takes effect above an exceptional winning margin of about 30 percentage points.

5. Conclusion

In mixed-member electoral systems, such as the electoral system of the German Bundestag, voters are required to cast two votes: one for a local district candidate and one for a party list. For many reasons, the nominal candidate vote makes it desirable that voters are aware of the candidates who run in their electoral district and, in the best case, have some knowledge about those candidates that allows them to cast an informed vote. At the same time, district candidates have party affiliations. Since voters need to make up their minds about which party list they vote for anyway, they can use party heuristics to also select the district candidate they vote for. This poses the question of whether voters do, in fact, become aware of their local candidates and, if so, under which circumstances. In this study, we pursued a data-driven approach to learn about factors that contribute to voters' candidate awareness given the German mixed-member electoral system.

Our findings show that candidate awareness, i.e., being able to recall the names of local candidates in the run-up of federal elections in Germany, is far from general knowledge among voters. Survey evidence suggests that about every other voter recalls the name of at least one candidate, one out of three voters can recall at least two candidates, and about 15% can recall three or more candidates. Our analyses of the predictors of candidate awareness reveal that political characteristics of either candidates, voters, or districts are more predictive of voters' candidate awareness than social-demographic characteristics. This is normatively reassuring as existing inequalities in the propensity to recall candidate names seem to get channeled only through politically charged characteristics such as vote intention, party identification, and political knowledge into actual candidate awareness of voters.

While we find that many politically loaded variables matter for voters' awareness, some variables stand out because they carry no information that helps predict candidate awareness. Most notably, our results suggest that the type of incumbency does not contribute to voters' awareness of candidates. Typically, mixed-member electoral systems allow for dual candidacies, i.e., candidates can compete in both tiers, the nominal tier as well as the party-list tier, at the same time.

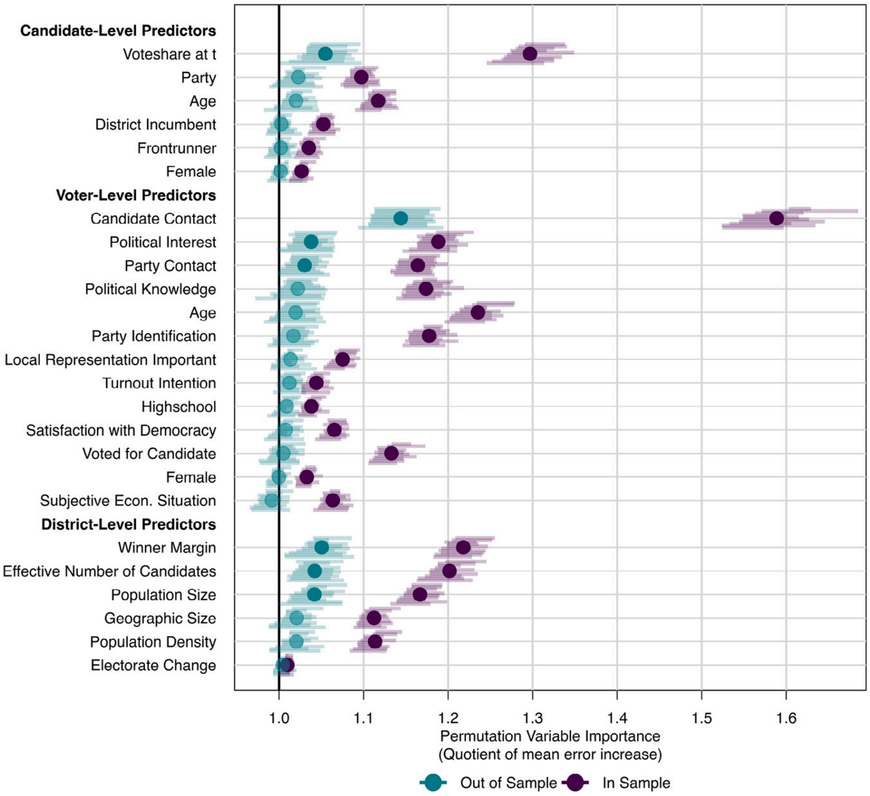

Fig. 3. Permutation variable importance of all predictors for the incumbency subset. Note: Permutation variable importance of all predictors in the random forest model fitted on the subset of candidates who are already members of the *Bundestag*, i.e. incumbents, prior to the election.

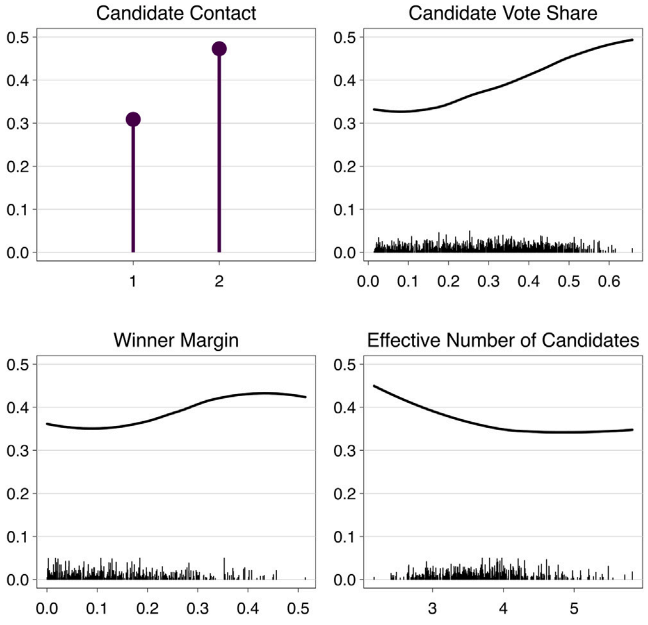

Fig. 4. Partial dependence plots for the four most important variables. x-axes represent the predictor variables, y-axes show predicted probabilities to recall a district candidate based on our random forest model.

Consequently, incumbency status cannot be as clearly conceptualized as in first-past-the-post systems. There are two different types of incumbents depending on their mode of election: district incumbents, who won the nominal district race in the previous election and list incumbents, who were elected through their party's list even if they might have lost their district race. We find that there is no difference between both incumbent types in terms of predicting voters' candidate awareness. Respondents in our data do not recall their district incumbents better than list incumbents. This might be surprising as candidates elected through a party list should have a priori no strong incentives to make themselves known to potential voters so that voters can recall their name and party affiliation correctly. As dual candidates, however, even list incumbents have such incentives. They compete in their local district as well, running more candidate-centered election campaigns (Gschwend and Zittel, 2015; Zittel and Gschwend, 2008) and serve there as 'shadow' district representatives to increase their chances of winning this district the next time or serve as the local representative of their party (Lundberg, 2006; Manow, 2015). This finding has also implications for the current electoral reform debate in Germany. The supposed importance of district incumbents for local representation seems to be more of a myth used by some to discredit particular reform proposals. At least our finding that voters are not more aware of district incumbents than list incumbents suggests that voters do not share this myth. District MPs are, for that matter, no better MPs than list MPs.

We perceive our study as a first step towards a more thorough understanding of voters' candidate awareness in mixed-member electoral systems. While voters' candidate awareness is an essential topic in democratic theory and a must-have, at least in some minimal form, for representation and accountability to work, current research is surprisingly innocent about its causes, especially in mixed-member electoral systems. Our methodological approach shows how scholars can use a data-driven research design to identify potential explanatory factors from a kitchen-sink list of conceivable predictors by treating the supposed data-generating process as unknown. In lieu of solid theoretical guidance, predictive modeling, and machine learning can offer a viable alternative to assuming the functional form of the relationship between various explanatory factors and voters' candidate awareness. Our criterion for identifying potentially explanatory factors is whether and how important they are in predicting voters' candidate recall as measured in surveys. Specifically, our quantity of interest is the *permutation variable importance* of each predictor in our random forest model. We rigorously evaluate the model's predictive ability out-of-sample, i.e., based on data not used during the estimation stage.

In the next step, research on this topic should turn to a more theoretically-driven approach. Our hope is that our results help future research to build causal theories on the determinants of the conditions under which voters can recall their candidates. We suggest that the characteristics identified as relevant predictors in our study should be put center stage in future research, both theoretically and empirically. Scholars should then develop implications of their theories and test them with causal research designs.

Our study is subject to a series of limitations. First, we note that our findings about candidate awareness depend on the assumption that recalling a candidate's name and party affiliation is diagnostic for thinking about candidates when making decisions. While we know of no research contradicting this assumption, it is conceivable that voters can remember a candidate's party affiliation but cannot recall their names (or vice versa). Fortunately, there are only a few respondents that do that. It is harder to imagine that respondents who neither recall their name nor their party do consider a candidate's identity seriously without wearing partisan lenses. Second, our list of potential predictors of candidate awareness is limited. Especially with respect to candidate characteristics, there are variables of potential interest for future research, including campaign spending or the distinction between high-rank and low-rank candidates. Third, the data-driven and exploratory nature of our approach needs to be emphasized. With our research design, we are not able to causally identify the effects of specific predictor variables on candidate awareness, and the results need to be interpreted accordingly.

Declaration of competing interest

The authors declare none.

Data availability

Replication Data for this article can be found in Harvard Dataverse at: https://doi.org/10.7910/DVN/T2XFXG.

Acknowledgments

We are grateful to Franziska Quoß, two anonymous reviewers, and the Editor for their constructive feedback on earlier versions of this manuscript.

Financial support

Oliver Rittmann acknowledges support by the University of Mannheim's Graduate School of Economic and Social Sciences (GESS).

Appendix. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.electstud.2023.102700.

References

Adler, Werner, Potapov, Sergej, Lausen, Berthold, 2011. Classification of repeated measurements data using tree-based ensemble methods. Comput. Statist. 26 (2), 355–369.

Athey, Susan, Imbens, Guido W., 2019. Machine learning methods that economists should know about. Annu. Rev. Econ. 11, 685–725.

Bergstra, James, Bengio, Yoshua, 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

Breiman, Leo, 2001a. Random forests. Mach. Learn. 45, 5–32.

Breiman, Leo, 2001b. Statistical modeling: The two cultures. Statist. Sci. 16 (3), 199–215.

Broockman, David E., Green, Donald P., 2014. Do online advertisements increase political candidates' name recognition or favorability? Evidence from randomized field experiments. Polit. Behav. 36 (2), 263–289.

van Buuren, Stef, Groothuis-Oudshoorn, Karin, 2011. Mice: Multivariate imputation by chained equations in R. J. Stat. Softw. 45 (3), 1–67.

Cain, Bruce E., Ferejohn, John A., Fiorina, Morris P., 1984. The constituency service basis of the personal vote for US representatives and british members of parliament. Am. Polit. Sci. Rev. 78 (1), 110–125.

Coleman, John J., Manna, Paul F., 2000. Congressional Campaign Spending and the Quality of Democracy. Technical Report 3.

Eisel, Stephan, Graf, Jutta, 2002. Bundestagswahl 2002 – die umstrittenen wahlkreise.

Elms, Laurel, Sniderman, Paul M., 2006. Informational rhythms of incumbent-dominated congressional elections. In: Brady, Henry E., Johnston, Richard (Eds.), Capturing Campaign Effects. The University of Michigan Press, Ann Arbor, MI, pp. 221–241.

Fiorina, Morris P., 1981. Retrospective Voting in American National Elections. Yale University Press, New Haven.

Frazer, Elizabeth, Macdonald, Kenneth, 2003. Sex differences in political knowledge in britain. Polit. Stud. 51 (1), 67–83.

Giebler, Heiko, Weßels, Bernhard, 2017. If you don't know me by now: Explaining local candidate recognition. German Polit. 26 (1), 146–169.

GLES, 2019a. Vorwahl-Querschnitt (GLES 2009). GESIS Datenarchiv, Köln, ZA5300 Datenfile Version 5.0.2, https://doi.org/10.4232/1.13228.

GLES, 2019b. Vorwahl-Querschnitt (GLES 2013). GESIS Datenarchiv, Köln, ZA5700 Datenfile Version 2.0.2, https://doi.org/10.4232/1.13231.

GLES, 2019c. Vorwahl-Querschnitt (GLES 2017). GESIS Datenarchiv, Köln, ZA6800 Datenfile Version 5.0.1, https://doi.org/10.4232/1.13234.

Grimmer, Justin, Roberts, Margaret E., Stewart, Brandon M., 2021. Machine learning for social science: An agnostic approach. Annu. Rev. Political Sci. 24, 395–419.

Grönlund, Kimmo, Milner, Henry, 2006. The determinants of political knowledge in comparative perspective. Scand. Polit. Stud. 29 (4), 386–406.

Gschwend, Thomas, 2007. Ticket-splitting and strategic voting under mixed electoral rules: Evidence from Germany. Eur. J. Polit. Res. 46 (1), 1–23.

Gschwend, Thomas, Rittmann, Oliver, Werner, Lisa-Marie, 2023. Zwischen wahlkreisreduzierung und bürgernähe: Zur aktuellen reformdiskussion des wahlrechts in baden-württemberg. Zeitschrift für Parlamentsfragen 54 (3), 611–624.

Gschwend, Thomas, Zittel, Thomas, 2015. Do constituency candidates matter in german federal elections? The personal vote as an interactive process. Elect. Stud. 39, 338–349.

Hajjem, Ahlem, Bellavance, François, Larocque, Denis, 2014. Mixed-effects random forest for clustered data. J. Stat. Comput. Simul. 84 (6), 1313–1328.

Holmberg, Soren, 2009. Candidate recognition in different electoral systems. In: Klingemann, Hans-Dieter (Ed.), The Comparative Study of Electoral Systems. Oxford University Press, Oxford, pp. 158–170.

Kam, Cindy D., Zechmeister, Elizabeth J., 2013. Name recognition and candidate support. Am. J. Polit. Sci. 57 (4), 971–986.

Karpievitch, Yuliya V., Hill, Elizabeth G., Leclerc, Anthony P., Dabney, Alan R., Almeida, Jonas S., 2009. An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. PLoS One 4 (9).

Key, V.O., 1964. Politics, Parties and Pressure Groups. Thomas Y. Crowell, New York, NY.

Kim, Seo-young Silvia, Zilinsky, Jan, 2022. Division does not imply predictability: Demographics continue to reveal little about voting and partisanship. Polit. Behav. 1–21.

Klingemann, Hans-Dieter, Wessels, Bernhard, 2001. The political consequences of Germany's mixed-member system: Personalization at the grass roots. In: Shugart, Matthew Soberg, Wattenberg, Martin P. (Eds.), Mixed-Member Electoral Systems: The Best of Both Worlds?. Oxford University Press, Oxford, pp. 279–296.

Kraft, Patrick W., Dolan, Kathleen, 2022. Asking the right questions: A framework for developing gender-balanced political knowledge batteries. Polit. Res. Q..

Kramer, Gerald H., 1971. Short-term fluctuations in us voting behavior, 1896-1964. Am. Polit. Sci. Rev. 65 (1), 131–143.

Kropko, Jonathan, Goodrich, Ben, Gelman, Andrew, Hill, Jennifer, 2014. Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. Polit. Anal. 22 (4), 497–519.

Laakso, Markku, Taagepera, Rein, 1979. "Effective" number of parties. Comp. Polit. Stud. 12 (1), 3–27.

Lang, Michel, Binder, Martin, Richter, Jakob, Schratz, Patrick, Pfisterer, Florian, Coors, Stefan, Au, Quay, Casalicchio, Giuseppe, Kotthoff, Lars, Bischl, Bernd, 2019. mlr3: A modern object-oriented machine learning framework in R. J. Open Source Softw. 4 (44), 1903.

Lundberg, Thomas Carl, 2006. Second-class representatives? Mixed-member proportional representation in britain. Parliam. Aff. 59 (1), 60–77.

Lupu, Noam, Warner, Zach, 2022. Why are the affluent better represented around the world? Eur. J. Polit. Res. 61 (1), 67–85.

Mann, Thomas E., Wolfinger, Raymond E., 1980. Candidates and parties in congressional elections. Am. Polit. Sci. Rev. 74 (3), 617–632.

Manow, Philip, 2015. Mixed Rules, Mixed Strategies: Parties and Candidates in Germany's Electoral System. ECPR Press.

Molina, Mario, Garip, Filiz, 2019. Machine learning for sociology. Annu. Rev. Sociol. 45, 27–45.

Montgomery, Jacob M., Olivella, Santiago, 2018. Tree-based models for political science data. Am. J. Polit. Sci. 62 (3), 729–744.

Neunhoeffer, Marcel, Sternberg, Sebastian, 2019. How cross-validation can go wrong and what to do about it. Polit. Anal. 27 (1), 101–106.

Parker, Glenn R., 1981. Interpreting candidate awareness in u. s. Congressional elections. Legislat. Stud. Q. 6 (2), 219–233.

Pattie, Charles J., Johnston, Ron J., 2004. Party knowledge and candidate knowledge: constituency campaigning and voting and the 1997 British general election. Elect. Stud. 23 (4), 795–819.

Pellagatti, Massimo, Masci, Chiara, Ieva, Francesca, Paganoni, Anna M., 2021. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. Stat. Anal. Data Min. 14 (3), 241–257.

Prinz, Timothy S., 1995. Media markets and candidate awareness in house elections, 1978–1990. Polit. Commun. 12 (3), 305–325.

Sandri, Marco, Zuccolotto, Paola, 2006. Variable selection using random forests. In: Data Analysis, Classification and the Forward Search: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6–8, 2005. pp. 263–270.

Schaub, Max, 2021. Acute financial hardship and voter turnout: Theory and evidence from the sequence of bank working days. Am. Polit. Sci. Rev. 115 (4), 1258–1274.

Shugart, Matthew Soberg, Valdini, Melody Ellis, Suominen, Kati, 2005. Looking for locals: Voter information demands and personal vote-earning attributes of legislators under proportional representation. Am. J. Polit. Sci. 49 (2), 437–449.

Sieberer, Ulrich, 2010. Behavioral consequences of mixed electoral systems: Deviating voting behavior of district and list MPs in the german bundestag. Elect. Stud. 29 (3), 484–496.

Sohnius, Marie-Lou, Gschwend, Thomas, Rittmann, Oliver, 2022. Welche auswirkungen haben größere wahlkreise auf das politische verhalten? Ein empirischer beitrag zur wahlrechtsreform. Politisc. Vierteljahresschr. 63 (4), 685–701.

Stratmann, Thomas, Baur, Martin, 2002. Plurality rule, proportional representation, and the german bundestag: How incentives to pork-barrel differ across electoral systems. Am. J. Polit. Sci. 46 (3), 506–514.

Zittel, Thomas, Gschwend, Thomas, 2008. Individualised constituency campaigns in mixed-member electoral systems: Candidates in the 2005 german elections. West Eur. Polit. 31 (5), 978–1003.