PA

When Correlation Is Not Enough: Validating Populism Scores from Supervised Machine-Learning Models

Michael Jankowski[1] **and Robert A. Huber**[2]

[1] *Institute for Social Sciences, Carl von Ossietzky University Oldenburg, Ammerländer Heerstraße 114-118, 26111 Oldenburg, Germany. Email: michael.jankowski@uol.de*
[2] *Department of Political Science, University of Salzburg, Rudolfskai 42, 5020 Salzburg, Austria. Email: robertalexander.huber@plus.ac.at*

Abstract

Despite the ongoing success of populist parties in many parts of the world, we lack comprehensive information about parties' level of populism over time. A recent contribution to Political Analysis by Di Cocco and Monechi (DCM) suggests that this research gap can be closed by predicting parties' populism scores from their election manifestos using supervised machine learning. In this paper, we provide a detailed discussion of the suggested approach. Building on recent debates about the validation of machine-learning models, we argue that the validity checks provided in DCM's paper are insufficient. We conduct a series of additional validity checks and empirically demonstrate that the approach is not suitable for deriving populism scores from texts. We conclude that measuring populism over time and between countries remains an immense challenge for empirical research. More generally, our paper illustrates the importance of more comprehensive validations of supervised machine-learning models.

Keywords: populism, manifestos, text-as-data, supervised machine learning, measurement validity

The rise of populist parties is one of the major transformations of party systems in the 21st century and has led to a plethora of research analyzing this development (e.g., Kriesi *et al.* 2012). However, it has only been in the last 10 years that scholars seem to have rallied behind a common definition of populism (Schäfer 2021, 2).[1] The now-dominant definition of populism goes back to Mudde (2004, 543) and describes populism as a "thin-centered ideology" that "considers society to be ultimately separated into two homogeneous and antagonistic groups, 'the pure people' and 'the corrupt elite'." It might be due to this struggle of how populism should be defined, that methodological research on how populism can be measured empirically developed only in recent years (Hawkins and Castanho Silva 2019).

In order to better understand the rise of populism, it is crucial to have valid measures of how parties' degree of populism developed over time and between countries. In a recent contribution to *Political Analysis*, Di Cocco and Monechi (2022a; in the folllowing: DCM) discuss a potential method to estimate such scores. They suggest applying supervised machine learning to election manifestos to measure parties' degree of populism. DCM, thus, contribute to recent research that utilized text-as-data techniques for measuring parties' level of populism (see, e.g., Hawkins and Castanho Silva 2019; Rooduijn and Pauwels 2011). DCM's approach suggests to provide *continuous* populism scores *over time* and *comparable across countries* (see page 313 in DCM). Moreover, the approach is easy-to-use and requires only little resources as it avoids labor-intensive annotations and "reduce[s] limitations inherent in human-coding techniques" (p. 312). As election manifestos are widely available for many countries, over time, and for different levels of government, the

1 For a summary of the development of political science research on populism and different research traditions, see Hunger and Paxton (2021).

Political Analysis (2023)
vol. 31: 591–605
DOI: 10.1017/pan.2022.32

Published
9 January 2023

Corresponding author
Michael Jankowski

Edited by
Jeff Gill

approach of DCM has the potential to solve the problem of missing data about the development of populism.

As we discuss in this article, a detailed inspection of DCM's approach unveils several concerns about the score's validity. While DCM apply sophisticated machine-learning models to their data and execute them exemplary, they provide insufficient validity checks to demonstrate that their approach captures meaningful variation in parties' degree of populism. DCM validate their approach by showing strong correlations between their measure of populism and external expert surveys. As we argue, these correlations are insufficient for demonstrating the validity of the approach because their models are trained on established expert surveys. We then proceed with a reanalysis of their models and conduct several additional validity checks. Our paper demonstrates that each of these checks calls the validity of their scores into question. We proceed as follows: First, we investigate the content validity of DCM's approach. As we demonstrate, DCM's models differentiate between populist and non-populist manifestos mainly by identifying party names, language differences, or policy positions. Second and more generally, we show that the approach by DCM is agnostic about the actual content of the manifestos. The approach assigns high populism scores to parties labeled as "populist" as long as there is some textual difference between the populists' and non-populists' manifestos. We demonstrate this point based on a coding error in one of DCM's text corpora as well as based on a systematic reshuffling analysis.

While our paper offers an in-depth discussion of the approach by DCM, it also contributes to the more general question of how supervised machine-learning models can be validated (e.g., de Vos and Verberne 2021). In their validation, DCM focus on assessing the predictive performance of their classifiers in the testing set (as suggested by, e.g., Grimmer and Stewart 2013, 295). Our study supplements recent debates which have argued that such validations can be insufficient (Baden *et al.* 2021). Supervised machine-learning models can have high predictive power without identifying the relevant concept (Hirst *et al.* 2014). Assessing the content validity of supervised machine-learning models is, thus, highly relevant when one is not only interested in high predictive power, but also in accurately measuring latent concepts of interest.

This paper starts by summarizing DCM's approach. We then describe why the provided validity checks of DCM are insufficient and proceed with a discussion of how supervised machine-learning models can be biased, even when they have high predictive power. We then empirically demonstrate that our concerns regarding the approach are warranted. The final section discusses more general implications for the validation of machine-learning models and for the study of populism.

1 Summary of the Approach by Di Cocco and Monechi

To predict parties' level of populism, DCM suggest the following approach. First, for each country, they take parties' election manifestos and use sentences as the unit of analysis. Second, they classify parties as populist or non-populist based on whether a party is listed as populist on the "PopuList" (Rooduijn *et al.* 2019). Each sentence from populist parties' manifestos receives a value of 1, whereas all other sentences get a value of 0.[2] Third, the data are split into a training (70%) and testing set (30%). Several machine-learning models are applied to the training set to predict whether a sentence comes from a populist party's manifesto. Fourth, the best-performing classifier (Random Forest) is selected for each country and is used for predicting whether a sentence from the *testing set* is populist or not. Finally, a manifesto's populism score is computed as the proportion of sentences predicted as "populist" in the testing set.[3] For example, an election

2 For the case of Italy, DCM also provide an additional analysis using manual coding. As DCM's paper advocates in favor of using the by far less resource-intensive coding of all sentences as either populist or non-populist, we do not discuss the analysis based on manual coding in this article.

3 DCM create two scores. The "global score" takes the fraction of sentences classified as populist for each party over all manifestos. The time-varying scores are created by taking the fraction of populist sentences for each manifesto.

manifesto with 100 sentences in the testing set of which 20 are classified to be "populist" receives a populism score of 0.2. DCM apply this procedure to six European countries since the early 2000s.

DCM explicitly state that their approach does not necessarily identify the *concept* of populism. Instead of measuring populism directly, the approach is supposed to be a "proxy" (see, e.g., page 313 in DCM). The approach, thus, circumvents the immense challenge of measuring populism directly from texts, which would require identifying anti-elitist and people-centrist ideas from the texts.[4] Instead, the approach predicts the probability that a sentence in the testing set comes from a populist party and *not* how populist a manifesto is. While DCM are explicit about this difference between their score's meaning and the definition of populism, the paper also creates the impression that their measure is an adequate substitute for measuring the concept of populism.[5]

2 Assessing the Validity of DCM's Populism Scores

When developing a new empirical measure for a theoretical concept, it is important to assess its validity to avoid systematic bias in the measurement. Several different types of validity have been suggested in the literature (e.g., Adcock and Collier 2001; Sartori 1970) among which *content validity* (does the measure capture the theoretical concept?), *construct validity* (does the data-generating process induce any systematic bias?), and *convergent validity* (does the measure correlate with other measures of the same concept?) are probably the most discussed aspects (see also McMann *et al.* 2021). DCM focus exclusively on assessing the convergent validity by correlating their measure with populism scores from established expert surveys. As they find strong correlations, DCM argue that their approach provides a valid measure (page 318 in DCM).

However, we argue that the strong correlation between DCM's populism scores and established expert surveys is insufficient for validating the approach. This is so because DCM train their models to identify populist manifestos by relying on the coding of the *PopuList*—an expert survey measuring whether a party is populist or not. Machine-learning models are trained to minimize the prediction error, that is, the models produce scores that are supposed to resemble the coding they were trained on (Grimmer, Roberts, and Stewart 2021). Thus, by design, DCM's scores resemble the coding of the PopuList. The only reason why DCM's scores do not perfectly resemble the binary coding of the PopuList is classification errors in which non-populist sentences are classified as populist or *vice versa*.[6] In other words, DCM's populism score for each party p in election t can be expressed as

$$\text{DCM's Populism Score}_{p,t} = \text{PopuList}_p + \varepsilon_{p,t}, \tag{1}$$

where PopuList_p is either 0 or 1, depending on whether the party is listed as non-populist or populist, and $\varepsilon_{p,t}$ is the proportion of classification errors for a specific manifesto. DCM's machine-learning models are trained to minimize $\varepsilon_{p,t}$, which in turn increases the correlation between their populism scores and the coding of the PopuList. The strong correlation between DCM's scores and external expert surveys on populism, such as the POPPA data by Meijers and Zaslove (2020), is,

4 There is a more general ongoing debate on how populism can be measured. Hawkins and Castanho Silva (2019) and Wuttke, Schimpf, and Schoen (2020) argue that the subdimensions of populism are non-compensatory. Hence, only observations that score high on all subdimensions should be coded as populist. Other studies show that many concepts, such as political trust or external political efficacy, are correlated with populism, and that it is hard to disentangle these concepts empirically despite being different theoretical concepts (Dolezal and Fölsch 2021; Geurkink *et al.* 2020). Finally, other recent studies demonstrate that even carefully designed survey scales measuring populist attitudes might lack validity and rather capture opposition to the government (Jungkunz, Fahey, and Hino 2021).

5 The title asks the question of "How populist are parties?" and also suggests that their approach measures "degrees of populism." The abstract states that their approach "derive[s] a score of parties' levels of populism," which "allows for obtaining a continuous score of populism." Only later it is clarified that they use the scores "as a proxy for parties' levels of populism."

6 This and all following arguments apply to the parties that were included in the training data. DCM exclude some parties with ambiguous populism levels from the training data.

thus, only logical. The PopuList and POPPA are both established expert surveys, and they strongly correlate. We can, thus, express the PopuList scores for each party as follows:

$$\text{PopuList}_p = \text{POPPA}_p + \zeta_p. \tag{2}$$

Hence, DCM's scores can be expressed as

$$\text{DCM's Populism Score}_{p,t} = \text{POPPA}_p + \zeta_p + \varepsilon_{p,t}. \tag{3}$$

This representation of DCM's validation approach clarifies two aspects. First, it shows that DCM's scores and external expert surveys—such as POPPA—will always strongly correlate as long as ζ_p and $\varepsilon_{p,t}$ are sufficiently low. Indeed, ζ_p is low as the PopuList and POPPA strongly correlate (see Section A1 of the Supplementary Material). $\varepsilon_{p,t}$ is by definition low because the models are trained to minimize the prediction error. Second, this representation also demonstrates that the temporal variation in DCM's populism scores is introduced by the classification errors of the machine-learning models. However, to validate their scores, DCM aggregate their scores to a single "global" populism score for each party which is the mean of all populism scores from a party. Thereby, this "global score" discards the temporal variation and, thus, DCM do not assess whether this variation is a valid reflection of parties' changes in populism over time.[7]

In other words, DCM essentially validate the PopuList since their scores are trained to resemble the coding of the PopuList.[8] What separates DCM's scores from the PopuList are classification errors, which introduce temporal variation for each party. However, in their validation, DCM do not assess whether this variation reflects meaningful changes in parties' degree of populism. This question can hardly be answered by assessing the model's overall predictive performance. Instead, it is necessary to analyze which concepts the machine-learning models have identified as being predictive for classifying sentences as populist.

3 Text Matters: Validating the Content Identified by Machine-Learning Models

Supervised machine-learning models classify texts by identifying features that are predictive for a certain class. In the case of DCM's approach, the models identify text features that help the algorithm to correctly classify sentences as coming from a populist or non-populist manifesto. Thus, it is important to look at the features that the machine-learning models have identified as being particularly predictive. In the following, we describe in more detail why such content validity checks are important and how even seemingly well-performing machine-learning models can induce systematic bias in the measurement.

3.1 How Supervised Machine-Learning Can Introduce Bias in DCM's Approach

The major advantage of supervised over unsupervised models is that they give researchers a certain degree of control over the concept that is supposed to be measured (Grimmer and Stewart 2013, 275). Moreover, it is often argued that the validation of supervised models is straightforward by evaluating the classifier's performance on the testing set (Grimmer and Stewart 2013, 279). This is in line with DCM's various model summaries, which generally imply a good model fit (see Table 1 in DCM). However, these advantages should not lead to the conclusion that supervised machine-learning models provide automatically valid measures of the desired concept once the

7 In Section A2 of the Supplementary Material, we correlate DCM's scores with the V-Party data, which is the only expert survey that provides time-varying populism scores (Lührmann *et al.* 2020). We find no correlation between DCM's and V-Party's populism scores within parties.

8 While DCM exclude some parties from the training set, the vast majority of parties from each country are included in the training set. Pooled across all countries, 75 of 91 (82%) parties are included in the training set (ignoring the regional parties in Spain for which no scores are computed by DCM). Moreover, the parties excluded from the analysis are often small and rather irrelevant parties for the respective party systems (such as the Pirate Party in Germany).

model performs well on the testing set (Hirst *et al.* 2014). As Baden *et al.* (2021, 4; emphasis added) concisely summarize:

> . . . operationalization [of a concept] is replaced by powerful algorithms trained to identify *any patterns and indicators that correlate with provided annotations*, effectively supplanting validity with predictive performance [. . .] In their effort to match human annotations or given ground truths, algorithmic classifiers show little interest in *separating valid variation in the material from accidental, meaningless regularities and confounding patterns*. Relying on salient, correlated patterns identified in the data, these tools may still frequently guess correctly, while *potentially introducing systematic biases* into the analysis.

This problem applies directly to the approach by DCM. Since DCM code *all sentences* from a populist party as 1, the models are not directly trained to identify populism. This coding introduces bias, because the models can rely on a wide range of other concepts than populism for making accurate predictions of whether a sentence comes from a populist party or not. Not every sentence of a populist party is populist. In fact, Rooduijn and Pauwels (2011) hand-code populist content in manifestos of European parties and find that the highest share of populist content among all analyzed manifestos is approximately 15%. Thus, many sentences contain other content than populist language, which might be picked-up by the machine-learning models as being predictive for manifestos of populist parties. The good performance of DCM's classifiers on the testing sets shows that their models often "guess correctly" and are, thus, able to distinguish between populist and non-populist manifestos. What is missing from their validation is an assessment of whether this variation is based on "valid variation" between the texts or only based on "accidental, meaningless regularities, and confounding patterns." The latter would be highly problematic as it indicates that variation in the scores (essentially the value of $\varepsilon_{p,t}$ in Equation (1)) is not based on a change in the degree of populism of a party. Instead, the score would vary due to some other either random or confounding factor that was identified by the algorithm. DCM's approach implicitly rests on the assumption that their models identify variation between populist and non-populist manifestos, which is meaningful for measuring variations in a party's degree of populism. However, they never test whether this is actually the case.

Referring to DCM's scores as a proxy measure of populism does not solve this problem. While a proxy measure does not need to measure the desired concept directly, Knox, Lucas, and Cho (2022) have recently highlighted that proxies can suffer from various sources of measurement error.[9] It is, thus, even more important to assess the validity of proxy measures to understand the "measurement gap" between the desired concept and its approximation. Specifically, Knox, Lucas, and Cho (2022, 421) argue that proxies only provide an imperfect measurement of the underlying concept for three potential reasons: "(a) [proxy] measures often fail to fully capture all aspects of the underlying concept, (b) they often contain some level of purely random noise, and (c) they are often systematically contaminated by other factors besides the concept of interest." We specifically focus on (c), that is, by which factors DCM's scores are contaminated and to what extent.[10] One might consider DCM's scores a reasonable proxy for populism, if they are contaminated by concepts that can serve as a reasonable approximation for populism. For example, some scholars have used political trust as a proxy for populism when no direct measure of populism was available (Dolezal and Fölsch 2021; Geurkink *et al.* 2020). In contrast, the validity of DCM's scores would be called into question, when the approach approximates populism based on less meaningful

9 Following the definition by Knox *et al.* (2022), supervised machine-learning models are *always* proxies, even when the model is trained on a coding that directly identifies the desired concept. In the case of DCM, the models are trained to replicate the coding of the PopuList, which is already a proxy for populism. Thus, DCM's scores represent, as one reviewer put it, "a proxy for a proxy."

10 Clearly, (a) and (b) are important as well. Regarding (a), we discuss this aspect in footnote 4 and in the conclusion. Concerning (b), we acknowledge that a certain degree of random noise cannot be avoided in a text-as-data approach.

concepts. Assume, for example, that the machine-learning model has simply identified party names as being predictive for classifying parties as either populist or non-populist. Party names are uninformative for measuring changes in the degree of populism. However, if party names are strong predictors for classifying parties as populist, then the machine-learning models will classify a party as more populist the more often it uses its party name in the manifesto. This type of variation is not meaningful for measuring populism or for approximating it. Thus, even when referring to the measure as a proxy, it is relevant to understand on which textual differences the classifications of DCM's model are based.

3.2 What Have DCM's Models Identified?

Following the discussion from above, we validate DCM's approach by assessing the feature importance of their models.[11] Feature importance describes how relevant a feature is for classifying a sentence as "non-populist" or "populist." There are different ways to estimate a feature's importance, and in Table 1 we report the feature importance based on the "mean decrease impurity" approach (Louppe *et al.* 2013). This measure does not tell us whether a feature has a positive or negative effect on the probability of predicting a sentence as "populist" or "non-populist"—it only shows which features are more relevant for the classification.[12]

Table 1 shows the five most important features for each of DCM's models. The features demonstrate that the models often identified concepts that are not or only vaguely connected to populism. In the case of Germany, the two most important features for classifying parties as populist are "link" (Left) and "afd" (AfD)—the party names of the two populist parties. As we discussed above, party names are uninformative for measuring different degrees of populism. A party does not become more (or less) populist when it mentions its party name more (or less) frequently in its manifesto. But this is what happens in DCM's approach. For example, 147 sentences of the Left Party in the testing set contain the feature "link" (which stands for "left"). Of these sentences, 143 (97.3%) are classified as "populist," demonstrating how strongly DCM's model rely on these terms. Among all of the sentences of the Left Party classified as populist, 26.6% (143/537) contain the feature "link". For the feature "afd", these patterns are even more pronounced. All of the sentences containing this term are classified as populist. In other words, the occurrence of the term "afd" leads almost directly to the classification of a sentence as populist. Among all sentences of the AfD classified as populist, 70.7% contain the term "afd". If one focuses on sentences of the AfD in which the term "afd" does not occur, then DCM's populism score for the AfD would decrease from 0.21 to 0.07, indicating that relatively few sentences of the AfD are classified as populist if the party does not mention its name. These findings clearly demonstrate how DCM's scores are driven by party names.

The other three features in Germany are "public", "social", and "employed". These features are indicative of a "thick" left-wing ideology. This finding is consistent with the observation that most of the "populist" sentences in Germany come from the Left Party. Because the Left Party contested in all elections and the AfD only in two of the analyzed elections, most sentences labeled as populist in the original corpus of DCM come from the Left Party. This creates a strong correlation between left-wing policy content and populism (see Section A4 of the Supplementary Material). One might argue that the identification of left-wing ideology is more informative compared to party names. While we agree, it remains a crude proxy for populism. Relying on thick ideologies as a proxy for populism would imply that all parties with a certain ideological leaning become more

11 All data and code to replicate our analyses are available from Political Analysis' Harvard Dataverse (Jankowski and Huber 2022).

12 Section A3 of the Supplementary Material to this paper provides a more in-depth discussion of the feature importance based on SHAP values (*SH*apley *A*dditive ex*P*lanations). SHAP values allow us to assess the direction of a feature, that is, whether its occurrence has a positive or negative impact on classifying a document as "populist" or "non-populist."

Table 1. Top-five most important features of the Random Forest model for each country.

Country	Feature		Feature importance
	Original	Translation†	
Germany	link	left•	0.0229
	afd	afd•	0.0068
	offent	public	0.0050
	sozial	social	0.0047
	beschaftigt	employed	0.0041
Spain	proposem‡	we propse	0.0058
	els‡	the	0.0057
	public	public	0.0037
	amb‡	with	0.0035
	dels‡	of the	0.0035
France	mesur	measure	0.0068
	constat	report	0.0057
	réalis	realize	0.0048
	suivant	following	0.0039
	écologiqu	ecological	0.0036
Italy	preved	forsee	0.0048
	stell	star•	0.0038
	attiv	active	0.0033
	access	access	0.0030
	abolizion	abolition	0.0029
Netherlands	d66	d66•	0.0094
	vvd	vvd•	0.0040
	christenunie	christian union•	0.0033
	nederland	netherlands	0.0031
	cda	cda•	0.0029

Note: † = Translations are based on nonstemmed versions, whereas the column "Original" reports the stemmed version of the feature. ‡ = Word is in Catalan. • = These features are party names. We do not report results from Austria due to a coding error in the original data used by DCM. We discuss this error in more detail below and in Section A5 of the Supplementary Material. In Section A10 of the Supplementary Material, we display the top-50 features for each country.

populist. It would also imply that a party becomes more (or less) populist when it changes its thick ideological position. This is in contrast to the conception of populism as a thin-centered ideology that is conceptually independent from thick ideologies. Moreover, studies connecting populism to left–right positions highlight that it is *radicalism* that is predictive for populism (Rooduijn and Akkerman 2017; Huber, Jankowski, and Juen 2022). The features in Germany—"public", "social", and "employed"—are not indicating radical language, but rather mainstream left-wing content. Given these arguments, it seems far-fetched to assume that ideology is a reasonable proxy for populism. This holds even more true for the case of Germany, given that the radical-right AfD is by far the most populist party in Germany (Meijers and Zaslove 2020), but DCM's models identified a left-wing ideology as predictive for populism.[13]

Similar patterns can be found in other countries. Four of the five most important features in the Netherlands are party names. In addition, the term "netherlands" is identified as important.

13 DCM acknowledge that they underestimate AfD's level of populism as it is underrepresented in the text corpus compared to the Left Party (see page 320 in DCM).

None of these terms is suitable for identifying or approximating populism. In France, Italy, and Spain, party names are not among the most important features. In Spain, however, the features point to a different aspect that was learned by the model. Apparently, four of the five most important features are in Catalan and not Spanish. This is caused by the fact that the Catalan party "In Common We Can" (En Comú Podem [ECP]) has been classified as populist party, and their manifestos are written in Catalan. As DCM exclude all other non-Spanish manifestos from the training data, ECP's manifestos are the only ones using Catalan language. Catalan is, thus, perfectly predictive of "populism."[14] Consequently, the ECP receives a high populism score (0.94). Finally, for France and Italy, it remains unclear which latent concept was identified by the models based on the relevant features.

In sum, the feature importance analysis suggests that the models have neither identified populism nor an approximate concept. In many cases, the predictions seem to be driven by the use of party names or the use of different languages. The high relevance of these features is not surprising considering how the models are trained since party names are highly predictive for identifying certain parties. But party names, language differences, and even thick ideologies cannot reasonably predict variations in populism.[15]

4 Irrelevance of Manifesto Content for Populism Scores in DCM's Approach

As we argued above, DCM's populism scores are primarily a transformation of the PopuList. DCM's scores will always correlate with external measures of populism as long as these external measures correlate with the PopuList and as long as populist and non-populist manifestos differ in some way from each other. In other words, DCM's approach is agnostic about the manifesto content. To demonstrate that this argument holds true, we first analyze the case of Austria in which a coding error in DCM's data caused the manifesto *content* to be misaligned with the manifesto *labels* used for training the data. Second, to demonstrate our point in a more systematic fashion and for a different case, we provide a reshuffling simulation for all cases. As we demonstrate, DCM's approach still predicts high populism values for populist parties although the underlying manifestos were *not* populist.

4.1 The Case of Austria

The data of DCM contain a coding error for the Austrian manifesto corpus.[16] The *labels* assigned to the manifestos do not identify the correct party manifesto *content*. For example, none of the manifestos labeled as being from the populist radical-right Austrian Freedom Party (FPÖ) are actually from this party. Instead, the *labels* of the FPÖ are assigned to manifesto *content* either from the Austrian Social Democratic Party (SPÖ) or from the Austrian People's Party (ÖVP)—both are non-populist parties (see the first five rows of Table 2). In total, only 4 of the 27 manifesto *contents* receive the correct manifesto labels. Section A5 of the Supplementary Material provides full information on which manifesto *label* was assigned to a certain manifesto *content*.

Because DCM use the *labels* for classifying parties as populist or non-populist when training their models, this coding error implies that the models were trained to identify populism (or approximate concepts) based on non-populist manifesto *content*. Yet, as we demonstrate in

14 DCM do not report populism scores for other non-Spanish manifestos since their scores "stand out as outliers" (page 319 in DCM).

15 In Sections A7 and A8 of the Supplementary Material, we present results from an analysis in which we excluded party names from the German and Dutch manifesto document-feature-matrix to analyze how the results are affected. In line with the findings described here, the AfD in Germany is no longer identified as populist by the model and the classification relies more strongly on left-wing ideology. Moreover, for Dutch manifestos, there is no evidence that the models identify populism when party names are excluded from the analysis.

16 The error is not in the original manifesto data from the Manifesto Project and was, thus, introduced by DCM. We contacted DCM to inform them about the coding error. They replied that the coding error happened accidentally due to a (non-reported) reshuffling analysis, and they published a Corrigendum (Di Cocco and Monechi 2022b).

Table 2. Example of coding errors for Austrian election manifestos.

DCM's manifesto *labels*		Manifesto *content*	
Party	**Year**	**Party**	**Year**
FPÖ	2002	SPÖ	2002
FPÖ	2006	ÖVP	2002
FPÖ	2008	SPÖ	2008
FPÖ	2013	ÖVP	2008
FPÖ	2017	ÖVP	2013
ÖVP	2002	SPÖ	2006
ÖVP	2006	ÖVP	2006
...	...	...	...

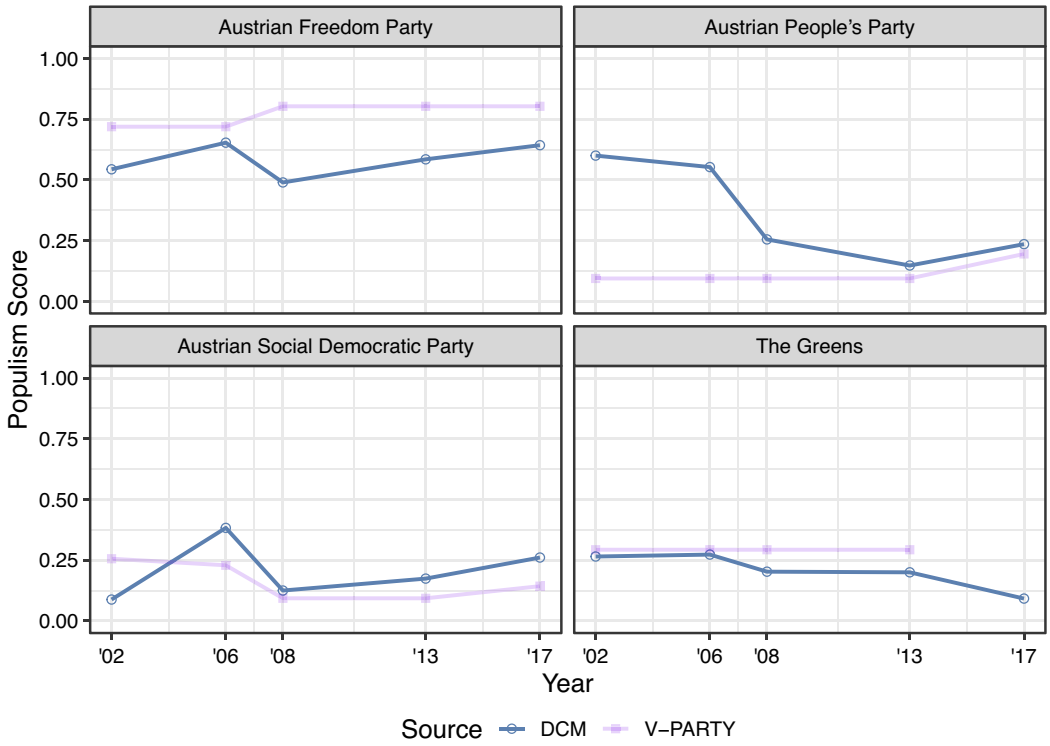

Figure 1. Populism scores of Di Cocco and Monechi (2022a) and V-Party for the four main parties in Austria. Note: The party names and years in the figure identify the *labels* used by DCM for training the machine-learning models in Austria. The *labels* are incorrectly assigned to the manifesto *content* (see Section A5 of the Supplementary Material).

Figure 1, the populist FPÖ still receives high "populism" scores based on the approach by DCM (we added the V-Party populism scores as a reference). In other words, despite not being trained on a single FPÖ manifesto, DCM's approach nonetheless predicts that the FPÖ is strongly populist. In contrast, the non-populist party *labels* still receive low populism scores, even when they are assigned to manifesto *content* of the populist FPÖ.[17] We also assessed the content of the manifestos which received the *labels* of the FPÖ. If these manifestos contain a high level of populist language, the classification of these manifestos as populist would be reasonable. However, based

17 The only exception to this pattern are the *labels* of the ÖVP in 2002 and 2006. These labels were assigned to manifesto *content* that is very similar to manifesto *content* on which the FPÖ labels were trained. Specifically, the *label* "ÖVP 2002" is assigned to the manifesto of the SPÖ in 2006 and the *label* "ÖVP 2006" was assigned to the correct manifesto. As can be seen from Table 2, the FPÖ labels were assigned to manifestos from these two parties in similar time periods.

on our reading of these manifestos, they contain only little populist language. Four of the five manifestos come from parties which were part of the government during that time and, thus, rather represented the established political elite against which populist parties try to mobilize. Only one manifesto, the SPÖ in 2002, contains some sentences which might be interpreted as populist. The SPÖ in 2002 was in the opposition and, thus, strongly criticized the government in their manifestos which shows some similarity to populist language. In sum, the coding error for Austria suggests that DCM's approach produces scores that correlate with external measures of populism, although the manifesto content was incorrectly assigned. In other words, the *content* of the manifestos seems to be largely irrelevant for DCM's populism scores.

4.2 Reshuffling Analysis

The case of Austria suggests that the manifestos' *content* is irrelevant for assigning populism scores in DCM's approach. In this section, we demonstrate that this finding is not limited to the Austrian case based on a reshuffling analysis for all text corpora used by DCM. To do so, we take the text corpus for each country and randomly reshuffle the party *labels* so that they identify different manifesto *content*. Then we apply DCM's approach to the reshuffled data using the code provided in DCM's replication material and store the resulting populism score for each party *label*.[18] We repeat this process 500 times for each country.[19] Thereby, we replicate the coding error from Austria for all countries, and we can systematically assess whether the populism scores are affected by the manifestos' *content*.

We illustrate the findings of this approach for the case of Germany in Figure 2. The results for the other countries are very similar and can be found in Section A9 of the Supplementary Material. We visualize the results in two ways. In Figure 2a, we show how the score of a manifesto changes when it gets assigned to a party *label* that is populist. In other words, each facet in Figure 2a displays the populism score of a specific manifesto conditional on whether the text randomly received a party *label* from a non-populist party (colored in blue) or populist party (colored in orange). Every manifesto receives a significantly higher "populism" score when it is *labeled* as populist although the *content* of the manifesto is always identical.[20] We also display the difference in means between both distributions including 95% confidence intervals. Overall, this analysis suggests that each manifesto can be "populist" when the manifestos are labeled as populist in the training stage. This is consistent with how the models are trained (to identify certain manifestos), but not with the idea of identifying populism. In Figure 2b, the results are displayed from the perspective of the party *labels*. Each box plot contains the scores of 500 randomly assigned manifesto *contents* to the respective *label*. The orange boxes highlight *labels* that are classified as populist. Just like in the case of Austria, we find that, even though the party labels are trained on random manifesto *content*, the populist *labels* receive a much higher populism score than the non-populist *labels*. In sum, the reshuffling analysis provides systematic evidence that DCM's approach is agnostic about the manifesto's *content* and assigns high populism scores to any text as long as it is labeled as populist when training the machine-learning models.

18 We use a Logistic Regression classifier as the Random Forest model is computationally demanding. The logistic regression classifier has a much shorter run time. As the information provided in DCM's appendix demonstrates, the logistic regression classifier performs quite similarly to the Random Forest model.

19 For the Netherlands and Spain, we only use 100 reshuffles as these text corpora are quite large and take a long time to compute.

20 In Figure 2, the PDS (Partei des Demokratischen Sozialismus) in 2005 and the Left Party in 2005 are both included although both were the same party and used a single manifesto. We explain this error in DCM's data in Section A6 of the Supplementary Material.

Figure 2. Scores for party manifestos (a) and manifesto labels (b) in Germany based on 500 reshuffling analyses.

4.3 Convergent Validity of DCM's and Random Models

Based on the reshuffling analysis, we can also provide evidence for our claim that DCM's scores will almost always correlate with external measure of populism. To do so, we correlate the scores generated in each iteration of the random reshuffling analysis with external expert data on populism. First, we use the V-Party data (Lührmann *et al.* 2020). These are the only data that provide populism scores over time. Thus, we can merge the scores from the iterations at the manifesto level. Second, we use the POPPA data that provide time-invariant populism scores. To this end, we aggregate the scores for each party to a single score from the reshuffling analysis and then merge these scores to the POPPA data at the party level.

For both cases, we correlate the scores produced by the random reshuffling approach with the data from the expert surveys and store the correlation coefficient. We display the distribution of the correlation coefficients for each country in Figure 3. It demonstrates that there is almost always a strong correlation between the reshuffling analysis' scores and external expert surveys. This is problematic, because when using random data, one would anticipate the scores not to be systematically correlated with external measures of populism. In contrast to this expectation, the average within-country correlation between POPPA's populism score and the "populism" scores from random manifesto content is 0.69 (0.56 for V-PARTY). The results show that, regardless of the

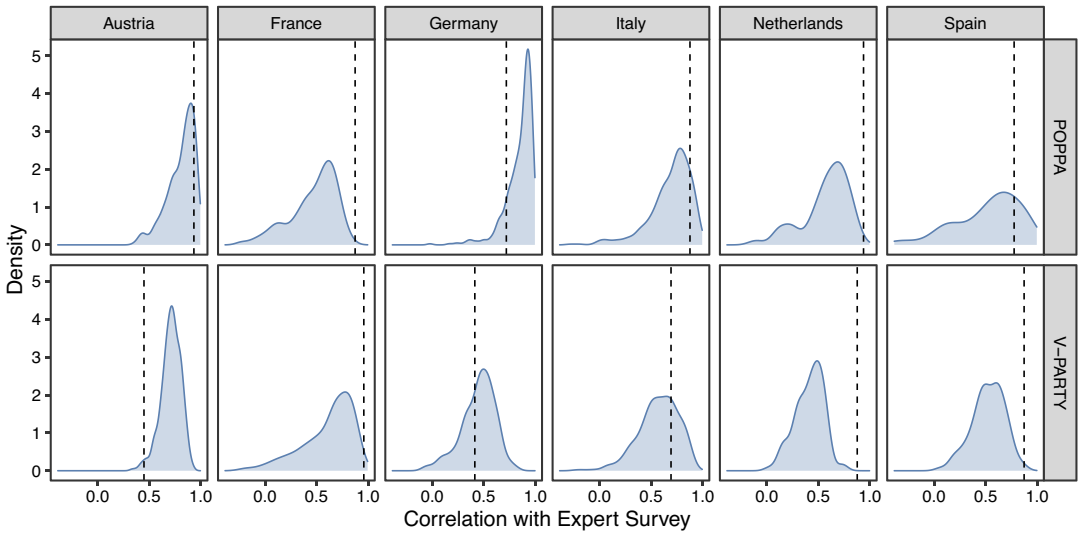

Figure 3. Correlation of "populism" scores from reshuffled text corpora with expert positions (V-Party and POPPA). Note: The dashed lines display correlation between DCM's populism scores with the respective expert surveys.

actual *content* of the manifestos, DCM's approach almost always results in scores which seem to correlate well with external measures of populism. In fact, in the majority of cases, the correlation between the random populism scores and expert data is nondistinguishable from DCM's populism scores as highlighted by the dashed line in Figure 3. For example, for the cases of Austria, Germany, Italy, and Spain, the correlation of DCM's scores with POPPA is not significantly better compared to the random models. Only for France and the Netherlands, the correlation of DCM's models is substantially higher compared to random models.[21] This, however, does not indicate that the models measured populism in these context. It only suggests that in these cases, the models are better in distinguishing between populist and non-populist parties when the correct manifestos are used. As we discussed in Section 3.2 (see Table 1), there is little evidence that the models actually identified populist language in these cases, but rather that they relied on party labels (e.g., the Netherlands) and different languages (e.g., Spain) for classifications.

The main conclusion from this analysis is that despite being trained on random manifesto content, the models of DCM produce values that strongly correlate with external measures of populism. Again, this is evidence that the models are agnostic about the manifestos' specific *content*. Our results show that DCM's central validity strategy, to assess the correlation of their scores with external expert surveys, is insufficient.

5 Conclusion

We currently lack systematic measures for populism that are comparable over time and between countries. DCM address this important research gap. However, as we demonstrated in this paper, DCM's approach has certain limitations. The models are trained to produce scores that resemble patterns found in the PopuList and, therefore, the scores also correlate with other expert surveys on populism. Thus, while the strong correlation with external expert surveys seems impressive, they rather validate the coding of the PopuList. We further demonstrate that DCM's approach will almost always predict high populism scores for the parties that are labeled as populist—even when the models are explicitly trained on manifestos from non-populist parties. DCM's models are largely agnostic about the actual content of the manifesto and rely on any textual

21 Likewise, the correlation between the scores provided by DCM's models and the expert scores is higher compared to the correlations between the random models and the expert surveys when the correlation is not conducted within countries.

differences between the populist and non-populist manifestos. The suggested method provides no mechanism to ensure that these textual differences between populist and non-populist parties are actually informative for measuring populism. In fact, as we demonstrated in this article, the models often rely on party names, language differences, or references to certain policy positions for classifying parties as populist or non-populist. Such "concepts" are not meaningful for analyzing systematic variation in a party's level of populism. Overall, our findings raise concerns about the approach's validity for measuring populism.

While primarily concerned with the approach by DCM, our paper has more general implications both for the measurement of populism and the validation of supervised machine-learning models. With regard to the measurement of populism, our paper highlights the tremendous challenges scholars face when measuring populism from texts. While extracting latent concepts from text is generally a nontrivial task, deriving valid measures of populism might be particularly challenging. Populist language is often highly context-specific because populism rests on the distinction between an "evil elite" (anti-elitism) and the "pure people" (people-centrism).[22] Which part of the society belongs to these groups differs between parties, countries, and over time.[23] Moreover, some authors argue that a valid measure of populism needs to measure both dimensions separately and then combine them to a single score in a noncompensatory manner (Hawkins and Castanho Silva 2019; Wuttke *et al.* 2020). Another challenge is that populism often occurs in combination with certain thick ideologies. Any measure of populism, thus, might be biased to a certain degree by a party's thick ideological position. Solving these methodological obstacles probably requires further advances in the development of text-as-data approaches. However, these limitations should not lead to the conclusion that measuring populism using supervised machine-learning is impossible. Many of the limitations identified in this paper can be traced back to DCM's decision not to code populism at the sentence or paragraph level. Recently, for example, Dai and Kustov (2022) have shown that supervised machine-learning models can identify populism in texts. They manually code paragraphs as populist or non-populist and then train machine-learning models using word embeddings. While such an approach is more resource-intensive, it seems to avoid several of the pitfalls identified in this paper. Of course, the resources required for such an approach are much higher, particularly when texts in different languages are coded.

Regarding the validation of supervised machine-learning models, our paper echoes Baden *et al.* (2021) that such models are often insufficiently validated by putting too much emphasis on the predictive performance in the testing set and paying too little attention on whether the models have actually identified the desired concept (see also Hirst *et al.* 2014). In fact, DCM followed many recommendations given in the literature on validating supervised machine-learning models, but following these recommendations does not always seem to be sufficient. Of course, a good performance on the training set is important for the validity of a supervised machine-learning model. But as we demonstrated in this article, it does not guarantee that the model is unbiased.[24] Since the number of applications of supervised machine-learning models will continue to increase in political science research (Grimmer *et al.* 2021), being aware of these potential pitfalls is important. Against this background, future research could develop a more systematic framework for validating machine-learning models in the social sciences.

22 This contextual nature of populism might also be issue-dependent within a party. For example, the British Conservatives may rely on populist language in EU-related issues, but not in other parts of their manifestos.

23 This is potentially one reason why qualitative techniques, such as holistic grading (Hawkins 2009), have been popular as they can incorporate the contextual factors more easily.

24 Another option is to develop more sophisticated methodological models that are able to detect and avoid such biases. For example, Rheault and Cochrane (2020, 114) developed a method that can address confounding in the estimation process.

Acknowledgments

We are grateful to Nicolai Berg, Thomas Bräuninger, Bruno Castanho Silva, Clint Claessen, Andreas Dür, Christina-Marie Juen, Rosa Kindt, Maurits Meijers, Stefan Müller, Marius Sältzer, Christian Stecker, Annika Werner, Andrej Zaslove, Lisa Zehnter, the participants of the "Text as Data Reading Group," five anonymous reviewers, and the Editor (Jeff Gill) for providing helpful feedback on previous versions of this paper. We also thank Jessica Di Cocco and Bernardo Monechi for the constructive communication. We rotate the order of authors across various publications.

Conflicts of Interest

The authors declare no conflicts of interest related to this work.

Data Availability Statement

Replication code for this article is available at https://doi.org/10.7910/DVN/DDXRXI (Jankowski and Huber 2022).

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2022.32.

References

Adcock, R., and D. Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3): 529–546. https://doi.org/10.1017/S0003055401003100. https://www.cambridge.org/core/product/identifier/S0003055401003100/type/journal_article.

Baden, C., C. Pipal, M. Schoonvelde, and M. A. C. G. van der Velden. 2021. "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda." *Communication Methods and Measures* 16 (1): 1–18. https://doi.org/10.1080/19312458.2021.2015574. https://www.tandfonline.com/doi/full/10.1080/19312458.2021.2015574.

Dai, Y., and A. Kustov. 2022. "When Do Politicians Use Populist Rhetoric? Populism as a Campaign Gamble." *Political Communication* 39 (3): 1–22. https://doi.org/10.1080/10584609.2022.2025505. https://www.tandfonline.com/doi/full/10.1080/10584609.2022.2025505.

De Vos, H.P., and S. Verberne. 2021. "Small Data Problems in Political Research: a critical replication study." In: I. Rehbein, G. Lapesa, G. Glavas, and S. Ponzetto (Eds.) 1st Workshop on computational linguistics for political text analysis (CPSS-2021).

Di Cocco, J., and B. Monechi. 2022a. "How Populist Are Parties? Measuring Degrees of Populism in Party Manifestos using Supervised Machine Learning." *Political Analysis* 30 (3): 311–327. https://doi.org/10.1017/pan.2021.29. https://www.cambridge.org/core/product/identifier/S1047198721000292/type/journal_article.

Di Cocco, J., and B. Monechi. 2022b. "Corrigendum and Addendum to: How Populist Are Parties? Measuring Degrees of Populism in Party Manifestos using Supervised Machine Learning." https://doi.org/10.48550/ARXIV.2201.07972.

Dolezal, M., and M. Fölsch. 2021. "Chapter 9: Researching Populism Quantitatively: Indicators, Proxy Measures and Data Sets." In *Political Populism*, edited by R. Heinisch, C. Holtz-Bacha, and O. Mazzoleni, 177–190. Nomos Verlagsgesellschaft mbH & Co. KG. https://doi.org/10.5771/9783748907510-177. https://www.nomos-elibrary.de/index.php?doi=10.5771/9783748907510-177.

Geurkink, B., A. Zaslove, R. Sluiter, and K. Jacobs. 2020. "Populist Attitudes, Political Trust, and External Political Efficacy: Old Wine in New Bottles?" *Political Studies* 68 (1): 247–267. https://doi.org/10.1177/0032321719842768. http://journals.sagepub.com/doi/10.1177/0032321719842768.

Grimmer, J., M. E. Roberts, and B. M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24 (1): 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921. https://www.annualreviews.org/doi/10.1146/annurev-polisci-053119-015921.

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297. https://doi.org/10.1093/pan/mps028. https://www.cambridge.org/core/product/identifier/S1047198700013401/type/journal_article.

Hawkins, K. A. 2009. "Is Chavez Populist? Measuring Populist Discourse in Comparative Perspective." *Comparative Political Studies* 42 (8): 1040–1067. https://doi.org/10.1177/0010414009331721. http://cps.sagepub.com/cgi/doi/10.1177/0010414009331721.

Hawkins, K. A., and B. Castanho Silva. 2019. "Textual Analysis: Big Data Approaches." In *The Ideational Approach to Populism: Concept, Theory, and Method*, edited by K. A. Hawkins, R. E. Carlin, L. Littvay, and C. Rovira Kaltwasser, 27–49. New York: Routledge.

Hirst, G., Y. Riabinin, J. Graham, M. Boizot-Roche, and C. Morris. 2014. *Text to Ideology or Text to Party Status?*, edited by B. Kaal, I. Maks, and A. van Elfrinkhof, Discourse Approaches to Politics, Society and Culture, Vol. 55, 93–116. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/dapsac.55.05hir. https://benjamins.com/catalog/dapsac.55.05hir.
Huber, R. A., J. Michael and J. Christina-Marie. 2022. "Populist parties and the two-dimensional policy space". *European Journal of Political Research*. https://doi.org/10.1111/1475-6765.12569. https://ejpr.onlinelibrary.wiley.com/doi/abs/10.1111/1475-6765.12569.
Hunger, S., and F. Paxton. 2021. "What's in a Buzzword? A Systematic Review of the State of Populism Research in Political Science." *Political Science Research and Methods* 10 (3): 617–633. https://doi.org/10.1017/psrm.2021.44. https://www.cambridge.org/core/product/identifier/S2049847021000443/type/journal_article.
Jankowski, M., and R. A. Huber. 2022. "Replication Data for: When Correlation Is Not Enough: Validating Populism Scores from Supervised Machine-Learning Models." Harvard Dataverse. https://doi.org/10.7910/DVN/DDXRXI
Jungkunz, S., R. A. Fahey, and A. Hino. 2021. "How Populist Attitudes Scales Fail to Capture Support for Populists in Power." PL*o*S One 16 (12): e0261658. https://doi.org/10.1371/journal.pone.0261658. https://dx.plos.org/10.1371/journal.pone.0261658.
Knox, D., C. Lucas, and W. K. T. Cho. 2022. "Testing Causal Theories with Learned Proxies." *Annual Review of Political Science* 25 (1): 419–441. https://doi.org/10.1146/annurev-polisci-051120-111443. https://www.annualreviews.org/doi/10.1146/annurev-polisci-051120-111443.
Kriesi, H et al. (eds). 2012. *Political Conflict in Western Europe*. Cambridge–New York: Cambridge University Press.
Louppe, G., L. Wehenkel, A. Sutera, and P. Geurts. 2013. "Understanding Variable Importances in Forests of Randomized Trees." *Advances in Neural Information Processing Systems* 26: 431–439.
Lührmann, A., et al. 2020. "Varieties of Party Identity and Organization (V-Party)." Dataset V1.
McMann, K., D. Pemstein, B. Seim, J. Teorell, and S. Lindberg. 2021. "Assessing Data Quality: An Approach and an Application." *Political Analysis* 30 (3): 426–449. https://doi.org/10.1017/pan.2021.27. https://www.cambridge.org/core/product/identifier/S1047198721000279/type/journal_article.
Meijers, M. J., and A. Zaslove. 2020. "Measuring Populism in Political Parties: Appraisal of a New Approach." *Comparative Political Studies* 54 (2): 372–407. https://doi.org/10.1177/0010414020938081. http://journals.sagepub.com/doi/10.1177/0010414020938081.
Mudde, C. 2004. "The Populist Zeitgeist." *Government and Opposition* 39 (4): 542–563. https://doi.org/10.1111/j.1477-7053.2004.00135.x.
Rheault, L., and C. Cochrane. 2020. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis* 28 (1): 112–133. https://doi.org/10.1017/pan.2019.26. https://www.cambridge.org/core/product/identifier/S1047198719000263/type/journal_article.
Rooduijn, M., and T. Akkerman. 2017. "Flank Attacks: Populism and Left-Right Radicalism in Western Europe." *Party Politics* 23 (3): 193–204. https://doi.org/10.1177/1354068815596514.
Rooduijn, M., and T. Pauwels. 2011. "Measuring Populism: Comparing Two Methods of Content Analysis." *West European Politics* 34 (6): 1272–1283. https://doi.org/10.1080/01402382.2011.616665. http://www.tandfonline.com/doi/abs/10.1080/01402382.2011.616665.
Rooduijn, M., et al. 2019. "The PopuList: An Overview of Populist, Far Right, Far Left and Eurosceptic Parties in Europe." http://www.popu-list.org.
Sartori, G. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64 (4): 1033–1053. https://doi.org/10.2307/1958356. https://www.cambridge.org/core/product/identifier/S0003055400133325/type/journal_article.
Schäfer, A. 2021. "Cultural Backlash? How (Not) to Explain the Rise of Authoritarian Populism." *British Journal of Political Science* 52 (4): 1977–1993. https://doi.org/10.1017/S0007123421000363. https://www.cambridge.org/core/product/identifier/S0007123421000363/type/journal_article.
Wuttke, A., C. Schimpf, and H. Schoen. 2020. "When the Whole Is Greater than the Sum of Its Parts: On the Conceptualization and Measurement of Populist Attitudes and Other Multidimensional Constructs." *American Political Science Review* 114 (2): 356–374. https://doi.org/10.1017/S0003055419000807. https://www.cambridge.org/core/product/identifier/S0003055419000807/type/journal_article.