Rallying around the flagging button?

An experimental investigation of intergroup discrimination in democratic norm enforcement

Roeland Endtz Amalia Álvarez-Benjumea

July 31st, 2026

Abstract

Citizen enforcement of norms against antidemocratic expression constitutes a line of defence against democratic backsliding. However, it remains unclear whether citizens uniformly enforce these norms between them. Two mechanisms – partisan motivated reasoning and intergroup norm differentiation – generate opposing predictions about the direction of politicised norm enforcement. This investigation tests which of these patterns dominates democratic norm enforcement in the Spanish context, decomposing politicised enforcement into its in-group and out-group dimensions. An online experiment examines whether cues of radical political allies or adversaries affect flagging antidemocratic posts as inappropriate. Results show that enforcement of democratic norms is indeed politicised, and generally more driven by denigration of political out-groups. Exploratory analyses suggest politicised enforcement varies in extent and nature between partisan groups. Radical-left partisans exempt their in-group from democratic norm enforcement; radical-right partisans enforce norms less overall, unless violated by an out-group. These findings emphasise that in-group protection and out-group denigration differentially shape democratic norm enforcement.

Recent scholarship on democratic backsliding has converged on an uncomfortable finding: democratic decline proceeds even where democratic attitudes remain stable (Castaldo and Memoli, 2024; Claassen, 2020; Valentim, 2024). This has shifted scholarly attention from studying what citizens think to studying what they do and, more recently, to the unwritten norms structuring what citizens are willing to express in public (Álvarez-Benjumea and Valentim, 2025; Helmke and Rath, 2025). According to this emerging view, the soft guardrails of democracy run not only between politicians but also between citizens, who sanction each other for political expressions deemed beyond the pale. Where these citizen-level sanctions hold, antidemocratic preferences remain private and democratic norms persist; where they erode, norm violations cascade into the political mainstream (Álvarez-Benjumea and Winter, 2018; Dinas et al., 2024).

This account rests on an assumption that has yet to be tested: that citizens enforce democratic norms impartially between them. Insights from social norm theory (Bicchieri, 2017) imply that this might not be the case. When deciding to follow a norm, individuals orient their attention mostly towards the actions and beliefs of their socially similar individuals, i.e., their reference group (Bicchieri and Dimant, 2022). Under high polarisation, partisan identity becomes a salient marker (Huddy and Bankert, 2017; Mason, 2018). With reference networks linked to perceived political allyship, democratic norm enforcement becomes politicised: whether authoritarian expressions are sanctioned depends on who says it. We test two possible mechanisms shaping the direction in which political in-group and out-group cues affect democratic norm enforcement, namely partisan motivated reasoning and intergroup norm differentiation. In doing so, we deconstruct the measurement of partisan discrimination into its in-group and out-group dimensions. We find that partisan democratic norm enforcement operates along the lines of partisan motivated reasoning, but that it varies systematically in extent and nature between partisan groups. Thus, our core theoretical claim is that politicised enforcement can operate through two distinct dimensions — in-group protection and out-group denigration — that are theoretically and empirically separable. Additionally, we find that these dimensions vary in strength across partisan groups in ways associated with their structural position in the political system.

We develop this claim across three theoretical steps. First, we argue that democratic norms governing public expression function as social norms in the sense of Bicchieri (2006, 2017) — informal behavioural rules sustained by individuals' beliefs about what others do and what others believe ought to be done. Once enforcement is at issue, the identity of the actor cannot be bracketed: the same behaviour is sanctioned differently depending on whether the actor falls inside or outside the observer's reference network. Second, the result is a politicised pattern of enforcement that we decompose into two analytically separable dimensions. The

in-group dimension refers to how perceived political in-groups are sanctioned relative to a neutral baseline; the out-group dimension concerns sanctioning of political out-groups relative to the same baseline. These dimensions are analytically independent. Politicised enforcement can occur through both, neither, or only one, and the presence of a neutral (no-cue) condition allows them to be distinguished. Third, we argue that politicised enforcement should vary in strength across partisan groups in ways predicted by their structural position in the party system.

This account speaks to the existing literature on partisan double standards concerning antidemocratic behaviour in three ways. Existing studies — most prominently Graham and Svolik (2020), Simonovits et al. (2022), Krishnarajan (2023), and Juratic et al. (2025) — treat partisan double standards as a single effect: in-group violators are evaluated more leniently than out-group violators. We treat it as the joint product of two distinct mechanisms whose relative weight is theoretically and empirically separable. Existing studies have also focused on attitudes (whether respondents condone antidemocratic behaviour) or on voting (whether respondents punish it at the ballot box); we focus on a third channel, citizen-to-citizen sanctioning of political expression, which Álvarez-Benjumea and Valentim (2024) and Dias et al. (2025) have recently begun to map but which has been little examined through the lens of intergroup discrimination. Finally, existing studies treat partisan double standards as a property of citizens in general; we treat them as a property of citizens that varies systematically with their position in the political system.

Democratic norms as citizen-level social norms

As authoritarianism is steeply on the rise around the globe (Nord et al., 2026), it is becoming increasingly apparent that this development cannot solely be explained through rising antidemocratic attitudes among the public. Firstly, levels of citizen support for democracy have remained stable in most countries (Claassen, 2020; Voeten, 2016; Wuttke et al., 2022). Additionally, state transitions away from democracy appear associated with increases in popular democratic support (Castaldo and Memoli, 2024; Claassen, 2020; Sadowski, 2024). Conversely, democratic support generally declines in countries experiencing democratisation (Olar and Neundorf, 2025). Furthermore, even in countries where a rise in authoritarian rule has been accompanied by a generational decline in pro-democratic attitudes (Claassen and Magalhães, 2023), increases in authoritarian political support have far outpaced any change in attitudes (Valentim, 2024). Therefore, recent scholarship has turned to non-attitudinal factors underlying the increasing popularity of antidemocratic actors among citizens. These include understandings of democracy (Kaftan and Gessler, 2025; Wunsch et al., 2025), partisan polarisation (Graham and Svolik, 2020) and, with increasing centrality, democratic

norms (Helmke and Rath, 2025).

Scholars have taken various approaches to the study of "the soft guardrails of democracy" (Levitsky and Ziblatt, 2018). Most commonly, democratic norms are treated as largely synonymous with democratic principles (Carey et al., 2022; Clayton et al., 2021; Kingzette et al., 2021). This line of research implicitly rests on the assumption that democratic norms are violated when political actors violate democratic principles. A second approach, prevalent within international political theory, views democratic norms as state conventions, internationally diffused by hegemonic actors (Cooley, 2015; Hyde, 2020; Jütersonke et al., 2021). A third approach, most relevant to the present investigation, draws on social norm theory. Social norms are informal behavioural rules, which are followed conditionally on social expectations regarding one's reference network (Bicchieri, 2006, 2017; Bicchieri and Dimant, 2022). Individuals comply with these norms so as to avoid social punishment by their peers (Fehr and Gächter, 2002); conversely, deviations from the norm face social sanctions (Fehr and Fischbacher, 2004). Recent scholarship has applied this framework to the realm of democracy (Bischof et al., 2024; Goldstein, 2024; Helmke and Rath, 2025). In this view, democratic norms are social norms regarding behaviours pertaining to democracy (Helmke and Rath, 2025) — thus being a subcategory of political norms (Álvarez-Benjumea and Valentim, 2024).

The benefits of the social norm framework for the study of citizen-level authoritarianism are twofold. First, this framework takes a constructivist approach to what constitutes a democratic norm and its violation. Antidemocratic behaviour is only assumed to violate norms if citizens find it inappropriate, and expect others to disapprove (Bicchieri, 2017). Thus, it avoids the pitfall of assuming that democratic norm violations are popularly recognised as such (Goldstein, 2024). Second, it extends the study of democratic norms beyond elite interactions to encompass interactions between citizens (Helmke and Rath, 2025). This is crucial, because these interactions constitute a key area of democratic norm contestation. Citizens enforce political norms among themselves by sanctioning counternormative political expressions (Ahn et al., 2024; Álvarez-Benjumea and Valentim, 2024). Consequently, the prospect of disapproval by other citizens deters individuals from sharing expressions perceived to violate social norms (Ammassari, 2023; Bursztyn et al., 2020; Valentim, 2024). When counternormative expressions are no longer constrained, this can lead to norm erosion through a cascade effect (Álvarez-Benjumea and Winter, 2018; Dinas et al., 2024). Thus, citizen norm enforcement constitutes a critical arena through which democratic norms are maintained — or eroded.

Citizen-level enforcement of democratic norms can take various forms, differing in their directness and costliness (Molho et al., 2020). Although indirect punishment is likely less

effective, individuals are more prone to engage in it (Álvarez-Benjumea and Valentim, 2025). Furthermore, a growing body of scholarship has demonstrated the effectiveness of norm enforcement in digital environments, curbing misinformation (Chuai et al., 2026), hate speech (Munger, 2017) and partisan incivility (Munger, 2021). A form of indirect online norm enforcement of specific relevance to this investigation is flagging inappropriate content on social networks. This tool is a popular vehicle for users to enforce norms in online discussions (Watson et al., 2019; Wilhelm and Schulz-Tomančok, 2024; Zhang et al., 2023). Users perceive flagging as an attractive way of sanctioning due to its cost-free and anonymous nature impeding counter-sanctioning (Rashidi et al., 2020). Importantly, flagging has been shown to affect users' judgment about a post (Gaozhao, 2021) and intentions to share it (Lanius et al., 2021; Liang et al., 2023). Thus, by deterring the spread of antidemocratic content, community flagging could constitute a cost-free yet effective way to enforce democratic norms.

Norm enforcement as a politicised act

If a social norm against authoritarian expression exists among citizens, enforcement of this norm between citizens could constitute a way to prevent rising authoritarianism. However, the notion that citizens would enforce these norms uniformly is complicated by insights on the interaction between social identity and social norms. One such interaction concerns reference networks, defined as the groups of people whose expectations and actions individuals are most attentive to (Bicchieri, 2017). These reference networks can either consist of close friends and family, or simply individuals with the same social identity (Bicchieri, 2006, 2017; Hogg and Reid, 2006). It has been well documented that individuals orient themselves towards their reference network when deciding whether to comply with a norm (Bicchieri and Dimant, 2022; Goldstein et al., 2008; Neighbors et al., 2008; Sinclair, 2012). Our central claim is that a parallel dynamic applies to norm enforcement, more specifically that of democratic norms. In politically polarised contexts, co-partisans constitute a natural reference network, making perceived political identity an axis along which democratic norm enforcement is structured. In-group and out-group dynamics can independently drive this politicised enforcement, in ways that diverge between partisan groups.

Existing literature paints a mixed image with regard to the direction of the relationship between perceived social group membership and norm enforcement. We argue that this is because of two possible mechanisms running in opposite directions. The first such mechanism is partisan motivated reasoning. Social group membership functions as a mechanism through which individuals strive to maintain a positive social identity (Tajfel and Turner, 2004). In polarised contexts, partisanship functions as a highly salient social group (Huddy and Bankert, 2017; Mason, 2018). This, in turn, can interact with the evaluation of norm-

violating expressions. Individuals engage in motivated reasoning so as to arrive at conclusions that support their pre-existing motivations (Kunda, 1990; Mercier and Sperber, 2011). This occurs especially in the case of ambiguous political information, as it provides more leeway for biased interpretation (Su, 2022). Given diverging popular conceptions of democracy, this might especially hold for authoritarian expressions (Wunsch et al., 2025). As such, individuals have an incentive to engage in norm manipulation, i.e., self-serving interpretations of the norm violation at hand (Bicchieri and Chavez, 2010). This will lead to the prioritisation of sanctioning political out-groups for a certain violation, as compared to in-groups violating the same norm. At the attitudinal level, this notion is supported by findings on democratic hypocrisy (Juratic et al., 2025; Simonovits et al., 2022), partisan intolerance (Tilley et al., 2026) and democratic rationalisation (Krishnarajan, 2023).

However, scholarship on social norms and their enforcement highlights a second mechanism suggesting an opposite dynamic. Bicchieri (2017) observes that individuals mostly pay attention to the actions of their reference group. As the subjective group dynamics model postulates, this dynamic is amplified in a context of status competition between groups: individuals maintain a positive group image by engaging in intergroup norm differentiation (Abrams et al., 2000; Marques et al., 1998). Simply put, individuals have an incentive to positively distinguish themselves from out-groups through group norms. Punishment is an effective way to maintain these group norms (Fehr and Gächter, 2002; Sääksvuori et al., 2011). This leads to a phenomenon dubbed the black sheep effect, where norm-deviant in-group members are derogated more harshly than out-group members deviating from the same norm (Marques et al., 1988; Marques and Paez, 1994; Tang et al., 2023). Importantly, this has behavioural consequences: shared group membership with a norm violator has been linked with increased sanctioning (Habyarimana et al., 2007; Shinada et al., 2004). Turning to politicised norm enforcement, it has been observed that partisans indeed perceive democratic norms to only be adhered to by their political in-groups (Braley et al., 2023; Pasek et al., 2022). Thus, antidemocratic expressions from an in-group might threaten one's group image, while the same expression from an out-group only confirms one's negative out-group perceptions. This could lead individuals to enforce norms against authoritarian expression more harshly in the case of perceived shared political identity. In short, while partisanship likely conditions social punishment for authoritarian expression, two opposing mechanisms push this effect in opposite directions.

Decomposing politicised norm enforcement

The investigation of politicised norm enforcement is further complicated by a mostly unaddressed gap in the literature. Positive attitudes towards political in-groups do not necessarily

go hand in hand with negative attitudes towards political out-groups (Bankert, 2021; Brewer, 1999). Extrapolating this to the behavioural domain, helping in-groups (in-group protection) and harming political out-groups (out-group denigration) are two empirically distinct phenomena (Amira et al., 2021; Martin-Gutierrez et al., 2024; Taflinger and Hudde, 2026). Which one of these phenomena predominates depends — among other things — on citizens' ideology (Yu et al., 2024), affective polarisation (Dimant, 2024) and the presence of symbolic threat to the in-group (Amira et al., 2021). Additionally, these phenomena are not contained to a dichotomous in-party/out-party distinction: in multiparty systems, citizens also favour partisans of in-bloc parties (Turnbull-Dugarte and López Ortega, 2025). Meanwhile, out-group denigration only occurs if the out-party is perceived as immoral (Weisel and Böhm, 2015). As such, intergroup discrimination likely varies between partisan groups, while also depending on the party at hand.

The distinction between in-group protection and out-group denigration is not considered in research on partisan discrimination in the implementation of democratic principles (Krishnarajan, 2023; Simonovits et al., 2022; Tilley et al., 2026). Similarly, scholarship on subjective group dynamics has typically established the black sheep effect through dichotomous comparisons between in-groups and out-groups (Abrams et al., 2000; Habyarimana et al., 2007; Reiman and Killoran, 2023). As such, not only does existing literature identify two mechanisms driving politicised norm enforcement in opposing directions; it is also not clear whether politicised enforcement is driven by differential treatment of in-groups or out-groups. To disentangle both the direction and dimensions of politicised norm enforcement, we thus formulate two sets of opposing hypotheses, comparing both in-groups and out-groups to a neutral baseline. The first set of hypotheses focuses on the in-group dimension of politicised norm enforcement:

> **H1a:** *In-group political cues are associated with decreased willingness to sanction antidemocratic expressions (Partisan motivated reasoning).*
>
> **H1b:** *In-group political cues are associated with increased willingness to sanction antidemocratic expressions (Intergroup norm differentiation).*

In contrast to this, the second set of hypotheses disentangle the out-group dimension of politicised norm enforcement:

> **H2a:** *Out-group political cues are associated with increased willingness to sanction antidemocratic expressions (Partisan motivated reasoning).*
>
> **H2b:** *Out-group political cues are associated with decreased willingness to sanction antidemocratic expressions (Intergroup norm differentiation).*

These hypotheses inform our knowledge on intergroup discrimination in democratic norm enforcement across multiple levels. First, we want to know whether it exists; second, in what direction it occurs; third, if it manifests through in-group or out-group discrimination. Additionally, we consider the possibility that in a multiparty context, these three levels of politicised enforcement might vary between partisan groups. We address this exploratory question after testing our main hypotheses.

Case study: freedom of expression in Spain

The specific democratic principle against which we test politicised norm enforcement is freedom of expression. This is a prominent element of liberal democracy in both scholarly conceptualisations (Claassen et al., 2025; Dahl, 1971; Teorell et al., 2016) and in popular understanding of democracy (Baviskar and Malone, 2004; Chu et al., 2024). Additionally, freedom of expression — at least in its abstract sense — is supported across the ideological spectrum (Stefanelli et al., 2025). In line with this, citizens widely recognise freedom of speech restrictions as antidemocratic (Sacher and Reinemann, 2025). At the same time, beliefs on the limits of freedom of expression are subject to social expectations (Bischof et al., 2024). As such, if publicly opposing democratic principles indeed constitutes a social norm violation, this will most likely apply to freedom of expression. Thus, it remains an open question whether our findings can be generalised to more abstract democratic principles not popularly recognised as such, i.e., governments abiding by court rulings (Ferrer et al., 2025).

Next, Spain is particularly suitable for the study of politicised enforcement regarding the norm of freedom of expression. This is firstly for theoretical reasons: existing literature on discrimination in the application of democratic principles tends to focus on the polarised two-party system of the United States (Graham and Svolik, 2020; Kingzette et al., 2021; Littvay et al., 2024). The Spanish case allows us to test whether these findings travel to a context similar in levels of affective polarisation (Garzia and Ferreira Da Silva, 2025), but different in its bloc-based party competition (Bantel, 2023; Simón, 2020). Second, Spain constitutes a highly functioning democracy where civil liberties are well protected (Nord et al., 2026). Because of this, it has been the basis of recent investigations focusing on norms against authoritarian expression (Álvarez-Benjumea and Valentim, 2024; Dinas et al., 2024; Jaziri-Arjona, 2025). Third, as shown in Figures A.1–A.3 in the Appendix, freedom of expression is firmly supported among the Spanish population. In a recent 40-country study, Spanish citizens were most likely to prioritise freedom of expression over curbing misinformation (Spampatti et al., 2026). This makes it more likely that calling for the curbing of freedom of expression is perceived as a norm violation among the Spanish populace. Finally, freedom of expression has been politicised in the parliamentary arena. Politicians of both the radical-left

Podemos party and the radical-right Vox party have called for restrictive measures regarding freedom of expression (Altimira, 2026; La Vanguardia, 2019), making these parties credible cues for our experiment. Thus, if citizens are not found to engage in politicised norm enforcement in Spain, the same will likely hold for citizens in less polarised contexts.

Analytical strategy

Preliminary study

The analysis of whether perceived political identity affects the likelihood to enforce the norm of freedom of expression rests on a fundamental assumption. This is the question of whether publicly opposing freedom of expression is perceived to be a norm violation in the first place. After all, no norm enforcement can be expected if a behaviour does not violate a perceived norm. To empirically test this precondition, we fielded a pre-registered repeated-measures incentivised survey among a sample of 500 Spanish participants in December 2025. Crossed quotas were applied for gender and age. After preregistered exclusion for attentiveness, speeding and AI use[1], the final sample consisted of 447 individuals. For completing the survey on Limesurvey, participants were rewarded with 12 Korus points, exchangeable for goods on the Nicequest platform.

The preliminary study presented respondents with a selection of screenshots of online posts expressing support for the curbing of freedom of expression, later to be used in the experiment. These posts were sourced from real-life posts on X, but modified to ensure anonymity, clarity and absence of partisan cues. Additionally, language expressing interpersonal incivility in these expressions was minimised, as this could constitute a norm violation in itself (Muddiman, 2017). A convenience sample answered questions about the perceived source of these posts to ascertain that they could plausibly be voiced by either the Spanish radical left or right. The antidemocratic posts were compared to a selection of baseline posts, similarly sourced, in which freedom of expression was defended. To prevent survey fatigue, participants were shown a random subset of three out of six posts for both the democratic and antidemocratic condition, presented in random order. All twelve statements and their translation are shown in Appendix B.1-12. To establish whether the antidemocratic posts were viewed as a norm violation, we applied an adaptation to the norm elicitation task used by Álvarez-Benjumea (2023). Specifically, for each post shown, participants were asked to

[1]Following findings by Westwood (2025) on potential AI use by survey participants, we included a text prompt injection item. This question, presented as an attention check, required participants to fill out their favourite colour, requesting that they fill in yellow as to show that they were paying attention. An invisible text elsewhere on the screen instructed participants to fill in "Mauve" instead. This has been found to be an effective way to detect AI usage (Zhang et al., 2026). Over the preliminary and main study, ten participants failed this check by filling in this invisible option.

fill in a battery consisting of the following items:

- **Attitudes:** participants were asked to rate their agreement with each expression presented to them on a 1–6 Likert scale, ranging from 1 ("Completely disagree") to 6 ("Completely agree")[2]. This item is traditionally not asked in social norm-related questionnaires and is therefore not included in our outcome variable. We nonetheless included it in the survey as to allow participants to report their (dis)agreement with a statement separately from their beliefs and expectations on its appropriateness. Additionally, this measure is a prerequisite for our incentivised measure of empirical expectations explained below.

- **Personal normative beliefs:** whether one personally believes a behaviour ought to be done (Bicchieri, 2017). Thus, for every statement, we asked participants to indicate whether, regardless of their personal agreement, they considered it appropriate to voice this statement to others, ranging from 1 ("Completely inappropriate") to 6 ("Completely appropriate").

- **Empirical expectations:** how common one believes a given behaviour to be (Bicchieri, 2017). As it would be counterintuitive to ask participants how common voicing this exact expression was, we requested them to predict how common indicated agreement with this expression was. In other words, participants were told other participants completed the same survey, and then asked to predict the modal answer to the 6-point attitude item. This corresponds with the measure of descriptive norm expectations by Pickup et al. (2024).

- **Normative expectations:** second-order beliefs about whether one perceives others to believe a given behaviour ought to be done (Bicchieri, 2017). Parallel to our measure of empirical expectations, we instructed participants to predict the modal score for the 6-point item on personal normative beliefs. This corresponds with the measure of injunctive norm perceptions by Bischof et al. (2024).

Consistent with recommendations by Szekely and Rodriguez (2026), we combine the items on personal normative beliefs, empirical expectations and normative expectations into

[2]Here, we somewhat deviate from Bicchieri (2017), who argues attitudes and personal normative beliefs to be conflated. This is likely to hold for evaluations of some behaviours, i.e. distribution decisions in an economic game. However, literature on hate speech shows that people can disagree with political expressions while still finding them appropriate to voice (Chong and Levy, 2018; Papcunová et al., 2023). In line with this, we contend that in the case of political expressions, there is a meaningful difference between attitudes and normative beliefs.

an equally weighted index of norm perceptions. To verify that these index components are internally consistent, Cronbach's α was calculated prior to the analysis. Secondly, we followed the recommendation by Bicchieri (2017) to incentivise items on empirical and normative expectations as to prevent participants from projecting their own beliefs and attitudes. Participants were informed that they would earn a point for every correct prediction of other participants' answers. They were also told that the 10% of participants who scored highest in predicting others' answers would receive a bonus of 15 Korus points, more than double their initial compensation. Participants were reminded of this fact before every set of incentivised questions. The 55 participants best predicting the answers of other participants received the bonus a few days after their participation.

Turning to the analytical approach, we opt for a crossed random effects model (individuals $\times$ statements), given that individuals evaluate a subset from a pool of statements. We include the antidemocratic nature of the statement as the independent variable, with random intercepts for participants and statements. Additionally, a random slope for the antidemocratic indicator is included at the respondent level, accounting for individual variation in differentiation between democratic and antidemocratic posts. The statistical notation of the model specification is provided below, with β_0 as the baseline norm perception, i.e. that of democratic posts; β_1 represents the fixed effect of antidemocratic publication A_j (1 = antidemocratic); the participant random intercept is given by u_{0i}; the random slope for the indicator at the participant level by $u_{1i}A_j$; and finally, the statement random intercept by v_{0j}.

$$Y_{ij} = \beta_0 + \beta_1 A_j + u_{0i} + u_{1i}A_j + v_{0j} + \epsilon_{ij} \tag{1}$$

As the random effect structure absorbs any between-individual confounds, no additional controls are included. However, regional and educational post-stratification weights were applied, with the maximum weight capped at 4. To ascertain that any effects were not driven by the weights, we also carry out a separate unweighted analysis. We maintain a conservative definition of a perceived norm violation as a substantial deviation from what is considered acceptable. As such, we contend that a significant effect of the democratic status of the expression is necessary but not sufficient. In line with this, we preregistered a smallest effect size of interest (SESOI) with a Cohen's d of 0.5, corresponding to a medium to large effect size (Lovakov and Agadullina, 2021). This means that we only accept that the antidemocratic statements are perceived as a norm violation if the outcome of the crossed random effects model is significant and the effect size exceeds this threshold. Note that the SESOI is only applied to establish a perceived norm violation; it does not speak to the severity of the perceived violation.

Main experiment

To test our main hypotheses on the relation between perceived political identity and norm enforcement, we recruited 1,210 Spanish citizens to a pre-registered online within-individual experiment. This sample size was the outcome of a power calculation assuming an effect size with a Cohen's h of 0.2 with a threshold of 80% at $p = .05$, using the `DeclareDesign` package on R.[3] In Spain and other democracies, a primary axis of competition is between left and right blocs (Bantel, 2023; Simón, 2020). Consequently, only participants identifying with the five largest Spanish parties adhering to this axis of political competition were selected. This proportion represents about two-thirds of the Spanish population at the time of the experiment (Centro de Investigaciones Sociológicas, 2025). Additionally, this study was only accessible to Netquest panel members that had not completed the survey pertaining to the preliminary study. Lastly, this experiment applied the same exclusion criteria as the preliminary study, leading to a final sample of 1,046 participants. Given that the political composition of our sample inherently differs from that of the general population, no post-stratification weights were used in the analysis. The distribution of the final sample in terms of region, education and partisan affiliation is provided in Appendix A.9-11.

After filling in a pre-survey with demographic items and questions on political identity, participants were asked to evaluate six posts by flagging the posts they found to be socially inappropriate or unacceptable. As such, for each post, individuals could either click a flagging button or a button allowing them to move on to the next post. After finishing the flagging task, participants were debriefed and rewarded for their participation. A visualisation of the instructions is given in Figure 1 below.

As part of the flagging task, all participants were presented with the same selection of six posts, presented in random order. These posts were selected by leveraging results from the preliminary study: the three antidemocratic posts scoring lowest in the norm perceptions index were selected alongside the three democratic posts scoring highest. The within-person manipulation consisted of the political cue randomly assigned to each post, namely Podemos, Vox or no cue. Specifically, as shown in Appendix B.13, the user profile header contained the hashtag #SiemprePodemos, #SoloQuedaVox or no hashtag whatsoever. These hashtags are popular slogans for their corresponding Spanish radical left and radical right parties, implying that the profiles belonged to partisans of these parties. The in-group and out-group conditions were then computed based on the indicated partisanship of the participant measured prior to the experiment. Following Turnbull-Dugarte and López Ortega (2025), we consider Podemos, PSOE and Sumar partisans to pertain to the left bloc and PP and Vox

[3]For the outcome of the power calculation for both the preliminary study and the main experiment, see Appendix D.

Figure 1

Task instructions online experiment, translated to English.

partisans to be the right bloc. Thus, the Podemos cue constitutes the in-group condition for the left bloc and the out-group condition for the right bloc; the opposite holds for the Vox cue. In this way, each participant got to see three antidemocratic posts, each one randomly assigned to one of the three conditions[4] . So as to facilitate the specificity test, all participants were additionally exposed to three democratic posts assigned to the same conditions. This design has the advantage of maximising statistical power by accounting for variation within individuals rather than between individuals.

Several reasons support the choice to include cues of the radical-left Podemos and radical-right Vox. First, Spanish politics are marked by a two-bloc logic in recent years (Madariaga and Riera, 2023; Simón, 2020). Within these blocs, partisans of radical parties are highly disliked by out-bloc partisans, while being relatively liked by their mainstream counterparts (see Table C.4). In line with this, Turnbull-Dugarte and López Ortega (2025) find that radical right citizens are treated as in-partisans by their mainstream counterparts. As such, partisans of radical Spanish parties are likely to attract the highest amount of dislike by their partisan opponents, while also enjoying relatively high liking by their mainstream partners. Second, both parties were opposition parties at the time of the experiment. This criterion,

[4]As reported in Appendix E, we ran a follow-up survey in July 2026 with 510 Spanish participants. One treatment arm served as a manipulation check to ensure participants registered the party cues. Additionally, we carried out a spillover check to verify that participants who saw cued posts first did not assume that non-cued posts shown afterwards came from the same parties. Both checks passed.

which excludes the radical-left Sumar, is important because democratic violations coming from (supporters of) an incumbent party are likely taken more seriously (Littvay et al., 2024; Van Der Brug et al., 2021). Third, these parties were chosen because they most clearly diverge on the ideological spectrum (see Appendix C.5). Correspondingly, radical left and right groups and parties are generally understood to be more restrictive regarding freedom of speech than their mainstream counterparts (Sacher and Reinemann, 2025). This also holds for Spain, where both Vox and Podemos have been accused of rhetoric limiting freedom of expression (El Debate, 2024; Petit, 2026). Because of this, cues of these parties are more credible options for the general public than mainstream-left or mainstream-right cues.

Our decision to operationalise in-group cues and out-group cues at the bloc level has certain theoretical consequences that should be noted. Voters of radical parties are more liked by their co-partisans than by partisans of their mainstream bloc partners. In other words, the in-group cues function as an in-party cue for some participants, while functioning as a weaker in-bloc cue for others. This means that the analysis of our hypotheses is conservative, also including groups for whom the in-group cue is less strong. Next to this, higher levels of affective polarisation among some partisan groups make intergroup discrimination more likely among those groups (Crespo-Martínez et al., 2024). For a clearer picture of how the extent and nature of politicised enforcement might differ, we disentangle the main effects by partisan group in an exploratory subgroup analysis.

With perceived in/out-bloc cues as dichotomous independent variables (baseline = no cue) and flagging as the dichotomous outcome variable, we carry out a mixed-effects logistic regression to test H1 and H2. This is in line with recommendations by Judd et al. (2012), who warn against the use of fixed effects for the analysis of experimental stimuli. Nonetheless, given the low number of clusters, we additionally carry out a robustness check with item-level fixed effects. Next, we run a separate specificity test using democratic posts with the same treatments in order to test to what extent differential norm enforcement uniquely applies to antidemocratic expressions. The model specification for the main analysis is given below, with P_{ij} as the probability that individual i sanctions antidemocratic post j. β_0 represents the baseline condition (antidemocratic content with no political cue); β_1 represents the fixed effect for the in-bloc condition; the fixed effect for the out-bloc condition is given by β_2, and the random intercepts for participants and statements by u_{0i} and v_{0j} respectively. No random slopes have been included because the single exposure to each condition would result in convergence issues. This study relies on two-tailed hypothesis testing ($\alpha = .05$) throughout, with graphics reporting confidence intervals at 95%.

$$P_{ij} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \,\mathrm{InBloc}_{ij} + \beta_2 \,\mathrm{OutBloc}_{ij} + u_{0i} + v_{0j})}} \tag{2}$$

Findings

Prior to the main analysis, we first discuss the central underlying assumption that publicly opposing freedom of expression is perceived as a social norm violation. The norm perceptions index we constructed to do so is sufficiently reliable with a Cronbach's α of 0.836; Table C.1 in the Appendix shows that correlations between items range from moderate to strong. Second, Table C.2 in the Appendix shows the outcome of the crossed random effects model comparing the index scores between democratic and antidemocratic posts. For robustness, the unweighted model is displayed alongside the model with regional and educational weights. In the main weighted model, expressions containing antidemocratic content are associated with a decrease of 0.59 on the 6-point norm perceptions index, equivalent to 12% of the scale range. This effect is highly significant ($p < .001$), and statistically substantial with a Cohen's d of .54, well above our preregistered SESOI. Additionally, Table C.3 in the Appendix reveals that it is substantively similar across index items.

We now turn to the question of whether and how perceived political identity affects the sanctioning of these perceived norm violations. The odds ratios resulting from the mixed-effects logistic regression model are outlined in Table 1 below. The insignificant intercept demonstrates that in the baseline condition (no cue), just over half of antidemocratic posts were marked as inappropriate. As the random effects show, variance was far higher between participants than between items, suggesting that flagging rates were similar between posts. With respect to the main effects, we find that the presence of an in-bloc cue is negatively associated with the likelihood to flag an antidemocratic post ($b = -.19$, SE $= .10$, OR $= .82$, $p = .045$). Perceiving that an antidemocratic post comes from an in-bloc account slightly decreases the odds of flagging by 18% as compared to when no political information is provided. Nevertheless, the effect of the out-bloc cue is considerably larger and more robust ($b = .32$, SE $= .10$, OR $= 1.38$, $p < .001$). Concretely, this means that the perception that an antidemocratic post was written by an out-bloc member is associated with a 38% increase in the odds of flagging said post. As shown in Table C.11 in the Appendix, both effects are robust to an alternative model specification with item-level fixed effects. However, Table C.12 in the Appendix shows that the out-group cue is more robust across stimuli than the in-group effect, where effects are more dependent on the stimulus. Additionally, as we discuss in the exploratory section, in-group protection is highly concentrated among Podemos partisans. Nonetheless, these observations are in line with the partisan motivated reasoning hypotheses (**H1a**, **H2a**), while contradicting the intergroup norm differentiation hypotheses (**H1b**, **H2b**). Nevertheless, they come with a caveat: out-group denigration mainly drives politicised enforcement, while the effect of in-group protection is more fragmented.

Table 1

Mixed-effect logistic regression: effect of in-bloc and out-bloc cues on flagging probability (baseline: no cue).

Predictors	*Odds Ratios*	*CI*
Intercept	1.127	0.852 – 1.491
In-Bloc	0.825*	0.683 – 0.995
Out-Bloc	1.382***	1.143 – 1.671
Random Effects		
τ_{00} part_id	0.80	
τ_{00} item_id	0.04	

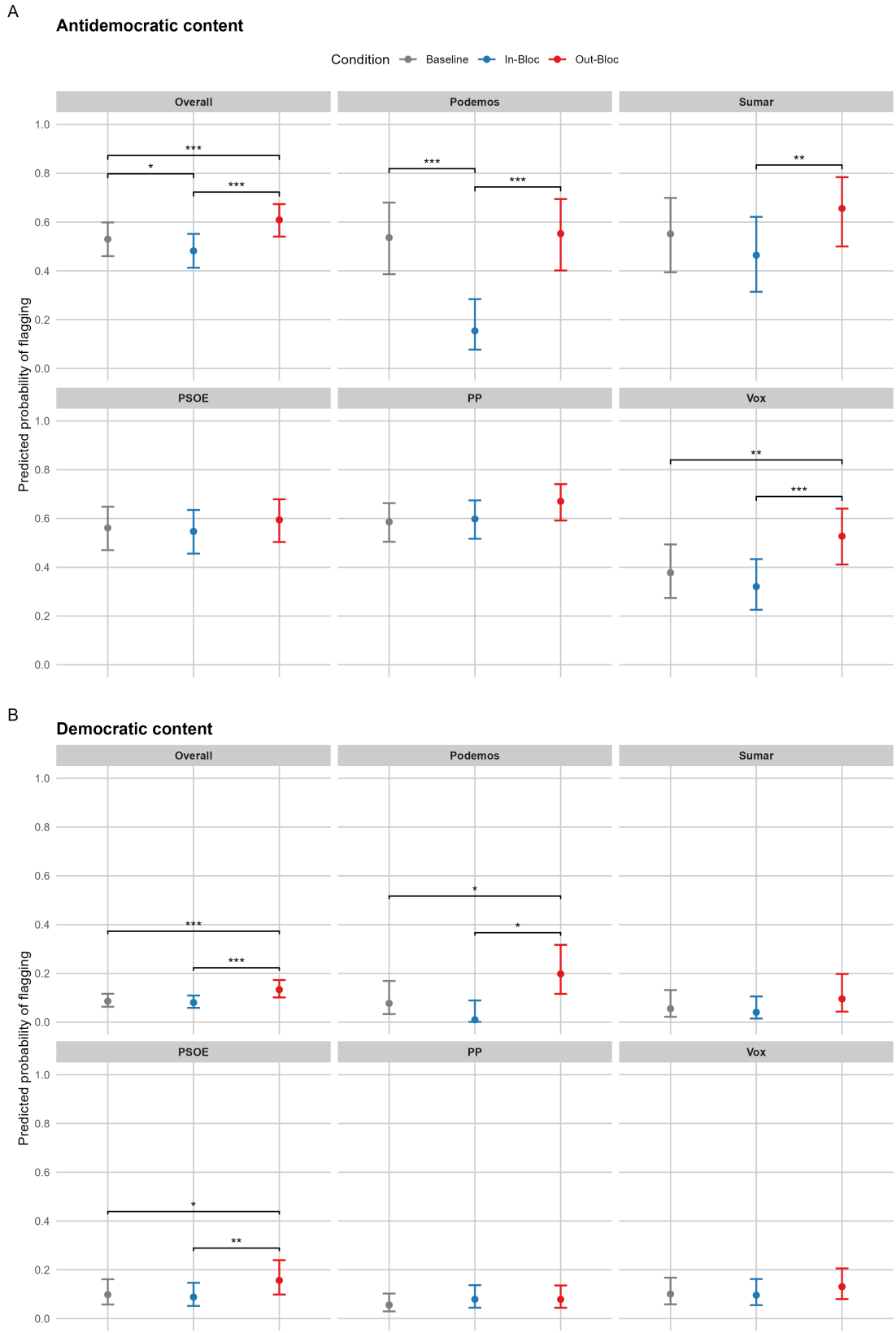
Note. 3138 observations, 1046 participants.
*$p < 0.05$ **$p < 0.01$ ***$p < 0.001$.

Exploratory analysis

Having identified politicised enforcement of democratic norms, we now present exploratory findings on three relevant follow-up questions. The first question is whether politicised enforcement manifests uniformly across partisan groups. The answer to this question is visualised below in Figure 2A, containing the predicted probabilities of flagging antidemocratic posts overall and by partisan group. As the repeated-measures design of the analysis results in 95% confidence intervals not capturing significance thresholds, significance brackets have been added for clarity. From top to bottom, these brackets compare the out-bloc and baseline condition, the in-bloc and baseline condition, and the out-bloc and in-bloc condition. Focusing on the latter comparison, the analysis suggests that politicised enforcement is concentrated among the Spanish radical left and radical right. Among partisans of the mainstream left and right, the differences between the in-group and out-group dimensions are small and not significant. On the other hand, partisans of Podemos and Vox differentiate strongly in their flagging, with a difference of 39.8 and 20.1 percentage points between the in-bloc condition and out-bloc condition ($p < .001$). Importantly, politicised enforcement is not confined to the parties cued in this experiment: the difference between the in-bloc and out-bloc condition is also highly significant among Sumar voters ($p < .01$).

Figure 2

Subgroup analysis: predicted probabilities of flagging by partisan group and condition.



Next, the brackets comparing the baseline condition to the cued conditions reveal that politicised enforcement not only varies in strength between partisan groups, but also in its nature. Among Vox partisans, baseline enforcement is significantly lower, by about 11 percentage points: flagging is only "activated" by the presence of an out-group cue (see Appendix C.13). In contrast to this, Podemos partisans do not differ from the general population in their norm enforcement when antidemocratic content is presented with no political cue. However, the in-group cue is uniquely strong among this group, cutting the likelihood of flagging by about 40 percentage points. While this effect is imprecisely estimated given the modest representation of Podemos partisans in the sample ($N = 68$), it exceeds the in-bloc effect for other partisan groups even in its most conservative estimate. Additionally, an interaction model with a dummy for Podemos partisanship (see Appendix C.14) shows that the in-bloc cue is significantly stronger ($p < .001$) for Podemos partisans than for the rest of the sample. As Tables C.15 and C.16 show, the differential effects found among Vox and Podemos partisans are robust to controlling for demographic differences between these groups and the general population.

The second question we investigate is how partisan cues affect flagging of norm-concordant expressions, i.e., expressions defending freedom of expression. Figure 2B presents the results of logistic mixed models using the three democratic items presented with the same cues, overall and by party. In this condition the in-group cue does not significantly affect flagging across any group, but a strong association between the presence of out-group cues and flagging is found ($b = .49$, $p < .001$). Table C.17 in the Appendix shows that while antidemocratic content drives flagging much more than out-group cues, the effect of the latter is similar between conditions. Nevertheless, two additional checks show that this finding should not be interpreted as blind partisan flagging of out-groups. First, the within-individual correlation between flagging out-group antidemocratic and democratic posts is nearly zero ($r = 0.02$; $p = .57$). Second, as Table C.18 in the Appendix shows, excluding participants who flag out-group democratic posts (i.e., potential blind flaggers) from the main analysis does not weaken the effect of the out-group cue, if anything strengthening it. Consistent with this, the subgroup analysis reveals a pattern opposite to what blind out-partisan flagging would predict. Here, the effect of the out-group cue is concentrated among Podemos and PSOE partisans, while no such effect is detected for antidemocratic expressions. Conversely, out-group cues coupled with norm-concordant content have a near-zero effect on Vox partisan flagging rates, while strongly driving flagging in the antidemocratic condition. Thus, left-wing partisans only sanction out-group cues accompanied with democratic behaviour; Vox partisans only do so when presented with antidemocratic behaviour.

Finally, we examine whether the observed effects could be explained by baseline differ-

ences in how respondents interpreted the partisan nature of our stimuli. In other words, we investigate whether our effects might be driven by individuals attributing (anti)democratic expressions to a specific partisan group even in the absence of any political cues. To answer this question, we fielded an additional survey on Prolific in July 2026 among 510 Spanish respondents, presenting them with the expressions used in the experiment without cues. We asked respondents to indicate the likelihood that a partisan of each of the five main parties wrote each expression[5]. As Tables C.19 and C.20 in the Appendix show, the dispersion of partisan attribution is high for the stimuli used in our experiment, both on an individual and an aggregate level. This means that when presented without cues, our stimuli were not interpreted as clearly partisan by individual respondents, nor did respondents agree in their partisan attributions. Still, as Figure A.12 shows, partisan attributions leaned towards Vox in two out of the three antidemocratic stimuli used; they leaned towards Podemos in the other stimulus. Nonetheless, Table C.12 demonstrates that the overall out-bloc effect is robust to excluding individual stimuli; Figure A.14 suggests that the same holds for the in-bloc effect among Podemos partisans. This means that our main findings hold even if we correct for the partisan lean of our stimuli in the no-cue condition.

To sum up, our findings show that democratic norms are not enforced uniformly: the perceived partisanship of the violator matters. Politicised enforcement occurs mainly because individuals sanction antidemocratic expressions more if they perceive a political out-group to be behind it. It can also be driven by partisans excusing political in-groups, although we find that this effect is highly concentrated among Podemos partisans. The exploratory analysis additionally suggests that politicised enforcement varies in extent and nature between partisan groups: in-group protection drives enforcement among Podemos partisans, whereas Vox partisans enforce less overall, only responding to out-group cues. Next, we find that even in the absence of antidemocratic expression, out-group cues affect sanctioning behaviour. The data do not support the interpretation of this being an artefact of blind out-group sanctioning. Rather, left-wing partisans only sanction out-group cues more in the case of democratic content, while out-group cues do not move flagging rates of democratic content among Vox partisans. Our follow-up study indicates that this cannot be explained by respondents' perceptions of what partisans might be behind the expressions used in our experiment.

[5]The preregistration of this study, as well as an elaboration on its results, is provided in Appendix E.

Discussion and conclusion

Our findings speak to existing literature in several ways. First, they support the budding literature conceptualising democratic norms as social norms (Bischof et al., 2024; Goldstein, 2024; Helmke and Rath, 2025). Citizens perceive antidemocratic expressions as violations of a social norm and are much more likely to flag these expressions than their democratic counterparts. This optimistic finding somewhat alleviates concerns raised by scholars on popular recognition of democratic violations (Wunsch and Gessler, 2023; Eroglu et al., 2025). Additionally, it bolsters confidence in the citizen channel of democratic norm maintenance (Álvarez-Benjumea and Winter, 2018; Helmke and Rath, 2025). However, our identification of sanctioning of antidemocratic expression comes with an important caveat. Given its anonymity and the impossibility of countersanctions, our measure of norm enforcement is likely inflated compared to levels of social sanctioning in interpersonal settings. As such, more scholarship is needed on whether — and how — citizens sanction antidemocratic expressions "in the wild".

Second, and more pessimistically, our research identifies that enforcement of democratic norms is politicised: antidemocratic expressions are sanctioned more often by out-bloc rivals than by in-bloc allies. This has both practical and theoretical consequences. On a practical level, this finding points to the importance of democratic deliberation between, and not just within, partisan groups. Where political deliberation remains contained in politically homogeneous spaces, this increases the likelihood of a cascade effect where antidemocratic expressions become normalised (Álvarez-Benjumea and Winter, 2018; Dinas et al., 2024). The increasingly "post-social" nature of digital communication, with online discussion taking place in semi-private discussion groups, constitutes a concerning backdrop for this observation (Törnberg and Rogers, 2026). Theoretically, we show that partisan double standards are not contained to political attitudes and voting behaviour, but also extend to the enforcement of democratic norms. This further underlines the relevance of recent findings on democratic hypocrisy, democratic rationalisation and partisan intolerance in multiparty systems (Juratic et al., 2025; Krishnarajan, 2023; Tilley et al., 2026). Meanwhile, no support for intergroup norm differentiation is found in the domain of antidemocratic expression. We suggest that this could possibly be due to the implied targets of antidemocratic rhetoric. Although the stimuli were carefully designed so no explicit targets were named, the political cues could lead participants to infer whose freedom of expression was targeted. Denigration of political in-groups is weakened when the out-group is targeted (Reiman and Killoran, 2023), while out-group denigration is heightened when the in-group is victimised (Amira et al., 2021; Goette et al., 2006; Rabellino et al., 2016). Rather than being an artefact of

this experiment, we argue that this dynamic is representative of antidemocratic expression in general: politicised enforcement could precisely be explained by authoritarian expression generally being directed towards out-groups (Stenner, 2005).

Third, our study expands on our understanding of partisan double standards by comparing in-group favouritism and out-group denigration to a neutral baseline. This yields two findings that connect to previous literature. On the one hand, previous scholarship on partisan double standards was not able to identify whether these standards were mostly driven by in-group favouritism or out-group denigration (Juratic et al., 2025; Krishnarajan, 2023; Tilley et al., 2026). Our findings show that politicised enforcement occurs both through out-group denigration and in-group favouritism, although the former is more dominant and the latter more case-specific. On the other hand, our results suggest that politicised enforcement does not manifest uniformly across the population. Mainstream partisans show little differentiation in their enforcement; Vox partisans enforce democratic norms less unless exposed to an out-group cue, whereas Podemos partisans exempt in-groups from otherwise regular norm enforcement. Theoretically, these findings highlight the importance of disentangling in-group favouritism and out-group denigration when studying partisan double standards. Practically, they provide a tentative explanation for why the effectiveness of intervention strategies reducing out-group hostility varies (Harteveld et al., 2026). For some partisans, tackling in-group favouritism could be a more effective strategy (Huang, 2026). An important limitation of our design however is that we are not fully able to disentangle in-bloc cues from in-party cues. We note that the strong in-party protection displayed by Podemos partisans does not extend to Vox partisans. Still, mainstream partisan groups could show similar in-group protection when presented with in-party cues instead of in-bloc cues. We recommend that more targeted designs test this possibility, although this likely comes with a credibility cost when studying antidemocratic expression.

A fourth take-away is that partisans sanction differently depending on the democratic nature of the expression. Out-group cues are effective in eliciting sanctioning behaviour among radical-right partisans when coupled with norm-violating content, but not for norm-concordant expression. The opposite holds for left-wing partisans, who only respond to radical-right cues in the case of regular democratic expression. We interpret this asymmetry as partially consistent with previous findings: Spanish radical-right identity markers are perceived as a norm violation, while the same does not hold for the left (Álvarez-Benjumea and Valentim, 2024). As such, the former category is sanctioned even when coupled with otherwise regular behaviour, while the latter is not (Álvarez-Benjumea and Valentim, 2025). However, the fact that this pattern reverses for norm-violating behaviour is an unexpected extension to the literature. If participants uniformly interpreted antidemocratic content as

radical right, this could provide a tentative explanation for this reversal. After all, this would mean that the antidemocratic condition functioned as a radical-right cue in its baseline condition. However, our follow-up survey shows that attribution is dispersed, and the observed patterns are similar when accounting for the perceived partisan lean of our stimuli. Thus, we note that more scholarship is needed to investigate whether this pattern extends to other contexts, and what drives it.

Fifth, our findings speak to the literature on social norm theory. Apart from strengthening the finding that political expression is a relevant arena of social norms, our results show that the perceived social identity of a norm violator matters for norm enforcement. Existing research has already shown that compliance with social norms hinges upon one's reference network, i.e., socially similar others whose behaviour and expectations one cares about (Bicchieri and Dimant, 2022; Goldstein et al., 2008). We find that in the context of online norm enforcement, individuals are more likely to sanction outside of their reference network. Taken together, these observations create a somewhat paradoxical dynamic. Those who violate democratic norms are disproportionately sanctioned precisely by the individuals toward whom they are most hostile. This might impede the effectiveness of norm enforcement, thereby facilitating norm erosion. In line with this, we recommend that future investigations assess whether out-group sanctioning is effective in bringing about norm compliance, or whether it leads to a backlash effect instead (Balcells et al., 2025; Munger, 2017).

Finally, our results have consequences for our understanding of the citizen-level predictors of democratic backsliding. Levitsky and Ziblatt (2018) argue that democratic norms function as soft guardrails, broadening the scope of acceptable political behaviour as they erode. Other scholars emphasise the role of affective polarisation in predicting democratic breakdown (Gidron et al., 2025; Haggard and Kaufman, 2021; Orhan, 2022). Nevertheless, experimental findings challenge the link between affective polarisation and democratic support (Broockman et al., 2023). Our research suggests an alternative pathway connecting affective polarisation to democratic deconsolidation. In the highly polarised Spanish context (Torcal and Comellas, 2022), the erosion of democratic norms could be facilitated by the fact that their enforcement is politicised along partisan lines. This dynamic is likely exacerbated by the relatively high level of politically homogeneous communication networks in Spain, further decreasing the likelihood of intergroup norm enforcement (Morales, 2010). Although this phenomenon is less likely to occur in democracies with low partisan animosity, it remains relevant in light of the general upwards trend in global levels of affective polarisation (Garzia et al., 2023). As of now, we can only make suggestive claims regarding this pathway. Thus, we contend that future investigations on the link between partisan animosity

and citizen-level democratic norm enforcement are needed.

References

Abrams, D., Marques, J. M., Bown, N. and Henson, M. (2000), 'Pro-norm and anti-norm deviance within and between groups', *Journal of Personality and Social Psychology* **28**(5), 906–912.

Ahn, C., Lelkes, Y. and Levendusky, M. (2024), 'Sanctioning political speech on social media is driven by partisan norms and identity signaling', *PNAS Nexus* **3**(12), pgae534.

Altimira, O. S. (2026), 'Absuelto el humorista Jair Domínguez, acusado por Vox: Llamar a dar 'un puñetazo' al fascismo no es delito de odio', elDiario.es. https://www.eldiario.es/catalunya/absuelto-humorista-jair-dominguez-acusado-vox-llamar-dar-punetazo-fascismo-no-delito-odio_1_13147082.html.

Álvarez-Benjumea, A. (2023), 'Uncovering hidden opinions: Social norms and the expression of xenophobic attitudes', *European Sociological Review* **39**(3), 449–463.

Álvarez-Benjumea, A. and Valentim, V. (2024), 'The enforcement of political norms', *British Journal of Political Science* pp. 1–24.

Álvarez-Benjumea, A. and Valentim, V. (2025), The social costs of showing far-right political preferences. Unpublished manuscript.

Álvarez-Benjumea, A. and Winter, F. (2018), 'Normative change and culture of hate: An experiment in online environments', *European Sociological Review* **34**(3), 223–237.

Amira, K., Wright, J. C. and Goya-Tocchetto, D. (2021), 'In-group love versus out-group hate: Which is more important to partisans and when?', *Political Behavior* **43**(2), 473–494.

Ammassari, S. (2023), 'It depends on personal networks: Feelings of stigmatisation among populist radical right party members', *European Journal of Political Research* **62**(3), 723–741.

Balcells, L., Martínez, S., Valentim, V. and vanderWilden, E. (2025), 'Re-Stigmatizing the Radical Right: A One-Way Street?', *Journal of Experimental Political Science* pp. 1–12.

Bankert, A. (2021), 'Negative and positive partisanship in the 2016 U.S. presidential elections', *Political Behavior* **43**(4), 1467–1485.

Bantel, I. (2023), 'Camps, not just parties. The dynamic foundations of affective polarization in multi-party systems', *Electoral Studies* **83**, 102614.

Baviskar, S. and Malone, M. F. T. (2004), 'What democracy means to citizens – and why it matters', *European Review of Latin American and Caribbean Studies* **76**, 3–23.

Bicchieri, C. (2006), *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge University Press.

Bicchieri, C. (2017), *Norms in the Wild: How to Diagnose, Measure and Change Social Norms*, Oxford University Press.

Bicchieri, C. and Chavez, A. (2010), 'Behaving as expected: Public information and fairness norms', *Journal of Behavioral Decision Making* **23**(2), 161–178.

Bicchieri, C. and Dimant, E. (2022), 'Nudging with care: The risks and benefits of social information', *Public Choice* **191**(3-4), 443–464.

Bischof, D., Allinger, T. L., Le Corre Juratic, M. and Frederiksen, K. V. S. (2024), How citizens perceive others: The role of social norms for democracies. Open Science Framework.

Braley, A., Lenz, G. S., Adjodah, D., Rahnama, H. and Pentland, A. (2023), 'Why voters who value democracy participate in democratic backsliding', *Nature Human Behaviour* **7**(8), 1282–1293.

Brewer, M. B. (1999), 'The psychology of prejudice: Ingroup love and outgroup hate?', *Journal of Social Issues* **55**(3), 429–444.

Broockman, D. E., Kalla, J. L. and Westwood, S. J. (2023), 'Does affective polarization undermine democratic norms or accountability? Maybe not', *American Journal of Political Science* **67**(3), 808–828.

Bursztyn, L., Egorov, G. and Fiorin, S. (2020), 'From extreme to mainstream: The erosion of social norms', *American Economic Review* **110**(11), 3522–3548.

Carey, J., Clayton, K., Helmke, G., Nyhan, B., Sanders, M. and Stokes, S. (2022), 'Who will defend democracy? Evaluating tradeoffs in candidate support among partisan donors and voters', *Journal of Elections, Public Opinion and Parties* **32**(1), 230–245.

Castaldo, A. and Memoli, V. (2024), 'Political support and democratic backsliding trends', *Journal of Contemporary European Studies* **32**(1), 138–156.

Centro de Investigaciones Sociológicas (2025), Barómetro de diciembre 2025, Technical report, Centro de Investigaciones Sociológicas. https://www.cis.es/documents/20117/13695327/es3536mar.pdf/d9378a63-c29d-e2a0-7ab3-aed20d0cd656.

Chong, D. and Levy, M. (2018), 'Competing norms of free expression and political tolerance', *Social Research: An International Quarterly* **85**(1), 197–227.

Chu, J. A., Williamson, S. and Yeung, E. S. F. (2024), 'People consistently view elections and civil liberties as key components of democracy', *Science* **386**(6719), 291–296.

Chuai, Y., Pilarski, M., Renault, T., Restrepo-Amariles, D., Troussel-Clément, A., Lenzini, G. and Pr linenllochs, N. (2026), 'Community-based fact-checking reduces the spread of misleading posts on X (formerly Twitter)', *Nature Communications* **17**(1), 4070.

Claassen, C. (2020), 'In the mood for democracy? democratic support as thermostatic opinion', *American Political Science Review* **114**(1), 36–53.

Claassen, C., Ackermann, K., Bertsou, E., Borba, L., Carlin, R. E., Cavari, A., Dahlum, S., Gherghina, S., Hawkins, D., Lelkes, Y., Magalhães, P. C., Mattes, R., Meijers, M. J., Neundorf, A., Öztürk, A., Sarsfield, R., Self, D., Stanley, B., Tsai, T.-h., Zaslove, A., and Zechmeister, E. J. (2025), 'Conceptualizing and measuring support for democracy: A new approach', *Comparative Political Studies* **58**(6), 1171–1198.

Claassen, C. and Magalhães, P. C. (2023), 'Public support for democracy in the united states has declined generationally', *Public Opinion Quarterly* **87**(3), 719–732.

Clayton, K., Davis, N. T., Nyhan, B., Porter, E., Ryan, T. J. and Wood, T. J. (2021), 'Elite rhetoric can undermine democratic norms', *Proceedings of the National Academy of Sciences* **118**(23), e2024125118.

Cooley, A. (2015), 'Countering democratic norms', *Journal of Democracy* **26**(3), 49–63.

Crespo-Martínez, I., Mora-Rodrıguez, A. and Rojo-Martínez, J. M. (2024), 'Political identity bias in interpersonal attitudes: Explaining affective polarisation in the 2023 Spanish general election', *Revista Latinoamericana de Psicología* **56**, 233–241.

Dahl, R. A. (1971), *Polyarchy: Participation and opposition*, Yale University Press.

Dias, N. C., Druckman, J. N. and Levendusky, M. S. (2025), 'Unraveling a "cancel culture" dynamic: When, why, and which americans sanction offensive speech', *The Journal of Politics* **87**(2), 588–600.

Dimant, E. (2024), 'Hate Trumps Love: The impact of political polarization on social preferences', *Management Science* **70**(1), 1–31.

Dinas, E., Martínez, S. and Valentim, V. (2024), 'Social norm change, political symbols, and expression of stigmatized preferences', *The Journal of Politics* **86**(2), 488–506.

El Debate (2024), 'La amenaza a la libertad de prensa llega al Congreso: Podemos registrará una ley para controlar a los medios', elDebate.com. https://www.eldebate.com/espana/20240517/comienzan-amenazas-libertad-prensa-podemos-registrara-ley-controlar-medios_197748.html.

El País & SER (2024), 'El 'desorden democrático' en España'. https://ep00.epimg.net/infografias/encuestas40db/2024/09-barometro/2024_09_barometro_democracia.pdf.

Eroglu, M. H., Finkel, S., Neundorf, A., Öztürk, A. and Ramírez, E. R. (2025), 'Choosing democracy over party? How civic education can mitigate the anti-democratic effects of partisan polarization', *British Journal of Political Science* **55**, e65.

Fehr, E. and Fischbacher, U. (2004), 'Third-party punishment and social norms', *Evolution and Human Behavior* **25**(2), 63–87.

Fehr, E. and Gächter, S. (2002), 'Altruistic punishment in humans', *Nature* **415**(6868), 137–140.

Ferrer, S., Hernández, E., Prada, E. and Tomic, D. (2025), 'The value of liberal democracy: Assessing citizens' commitment to democratic principles', *European Journal of Political Research* pp. 1475–6765.70021.

Gaozhao, D. (2021), 'Flagging fake news on social media: An experimental study of media consumers' identification of fake news', *Government Information Quarterly* **38**(3), 101591.

Garzia, D. and Ferreira Da Silva, F. (2025), 'In-Party Love, Out-Party Hate, and affective polarization in twelve established democracies', *Public Opinion Quarterly* **89**(2), 459–467.

Garzia, D., Ferreira Da Silva, F. and Maye, S. (2023), 'Affective polarization in comparative and longitudinal perspective', *Public Opinion Quarterly* **87**(1), 219–231.

Gidron, N., Margalit, Y., Sheffer, L. and Yakir, I. (2025), 'Why masses support democratic backsliding', *American Journal of Political Science* p. ajps.12958.

Goette, L., Huffman, D. and Meier, S. (2006), 'The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups', *The American Economic Review* **96**(2), 212–216.

Goldstein, D. A. N. (2024), The social foundations of democratic norms. SSRN Electronic Journal.

Goldstein, N. J., Cialdini, R. B. and Griskevicius, V. (2008), 'A room with a viewpoint: Using social norms to motivate environmental conservation in hotels', *Journal of Consumer Research* **35**(3), 472–482.

Graham, M. H. and Svolik, M. W. (2020), 'Democracy in america? partisanship, polarization, and the robustness of support for democracy in the united states', *American Political Science Review* **114**(2), 392–409.

Habyarimana, J., Humphreys, M., Posner, D. N. and Weinstein, J. M. (2007), 'Why does ethnic diversity undermine public goods provision?', *American Political Science Review* **101**(4), 709–725.

Haggard, S. and Kaufman, R. (2021), 'The anatomy of democratic backsliding', *Journal of Democracy* **32**(4), 27–41.

Harteveld, E., Berntzen, L. E., Kokkonen, A., Kelsall, H., Linde, J. and Dahlberg, S. (2026), 'The (Alleged) consequences of affective polarization: A survey experiment in nine democracies', *European Journal of Political Research* .

Helmke, G. and Rath, J. (2025), 'Defining and measuring democratic norms', *Annual Review of Political Science* **28**, 233–251.

Hogg, M. A. and Reid, S. A. (2006), 'Social identity, self-categorization, and the communication of group norms', *Communication Theory* **16**(1), 7–30.

Huang, Y.-S. (2026), 'Love blinds? Winners, in-party favoritism, and support for violations of democratic norms', *British Journal of Political Science* **56**, e35.

Huddy, L. and Bankert, A. (2017), Political partisanship as a social identity, *in* 'Oxford Research Encyclopedia of Politics', Oxford University Press.
URL: *https://oxfordre.com/politics/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-250*

Hyde, S. D. (2020), 'Democracy's backsliding in the international environment', *Science* **369**(6508), 1192–1196.

Jaziri-Arjona, T. (2025), Who breaks the stigma? social stratification and the normalization of the radical right. OSF Preprint.

Judd, C. M., Westfall, J. and Kenny, D. A. (2012), 'Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem', *Journal of Personality and Social Psychology* **103**(1), 54–69.

Juratic, M. L. C., Wagner, M. and Bischof, D. (2025), Democratic hypocrisy: Unequal tolerance for protest in germany. Unpublished manuscript.

Jütersonke, O., Kobayashi, K., Krause, K. and Yuan, X. (2021), 'Norm contestation and normative transformation in global peacebuilding order(s): The cases of China, Japan, and Russia', *International Studies Quarterly* **65**(4), 944–959.

Kaftan, L. and Gessler, T. (2025), 'The democracy I like: Perceptions of democracy and opposition to democratic backsliding', *Government and Opposition* **60**(2), 358–381.

Kingzette, J., Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M. and Ryan, J. B. (2021), 'How affective polarization undermines support for democratic norms', *Public Opinion Quarterly* **85**(2), 663–677.

Krishnarajan, S. (2023), 'Rationalizing democracy: The perceptual bias and (Un)Democratic behavior', *American Political Science Review* **117**(2), 474–496.

Kunda, Z. (1990), 'The case for motivated reasoning', *Psychological Bulletin* **108**(3), 480–498.

La Vanguardia (2019), 'Podemos y bildu piden que C.Tangana no actúe en las fiestas de bilbao por sus letras "machistas"'. https://www.lavanguardia.com/local/paisvasco/20190808/463944810211/podemos-bildu-c-tangana-fiestas-bilbao-letras-machistas.html.

Lanius, C., Weber, R. and MacKenzie, W. I. (2021), 'Use of bot and content flags to limit the spread of misinformation among social networks: A behavior and attitude survey', *Social Network Analysis and Mining* **11**(1), 32.

Levitsky, S. and Ziblatt, D. (2018), *How Democracies Die*, Crown.

Liang, F., Zhu, Q. and Li, G. M. (2023), 'The effects of flagging propaganda sources on news sharing: Quasi-Experimental evidence from Twitter', *The International Journal of Press/Politics* **28**(4), 909–928.

Littvay, L., McCoy, J. L. and Simonovits, G. (2024), 'It's not just Trump: Americans of both parties support liberal democratic norm violations more under their own president', *Public Opinion Quarterly* **88**(3), 1044–1058.

Lovakov, A. and Agadullina, E. R. (2021), 'Empirically derived guidelines for effect size interpretation in social psychology', *European Journal of Social Psychology* **51**(3), 485–504.

Madariaga, A. G. and Riera, P. (2023), 'The nationalisation of subnational elections in polarised Spain: The May 2023 regional and local elections', *South European Society and Politics* **28**(2), 231–257.

Marques, J. M., Abrams, D., Paez, D. and Martinez-Taboada, C. (1998), 'The role of categorization and in-Group norms in judgments of groups and their members', *Journal of Personality and Social Psychology* **75**, 976–988.

Marques, J. M. and Paez, D. (1994), 'The 'black sheep effect': Social categorization, rejection of ingroup deviates, and perception of group variability', *European Review of Social Psychology* **5**(1), 37–68.

Marques, J. M., Yzerbyt, V. and Leyens, J.-P. (1988), 'The "black sheep effect": Extremity of judgments towards ingroup members as a function of group identification', *European Journal of Social Psychology* **18**, 1–16.

Martin-Gutierrez, S., Robles Morales, J. M., Torcal, M., Losada, J. C. and Benito, R. M. (2024), 'In-party love spreads more efficiently than out-party hate in online communities', *Scientific Reports* **14**(1), 15700.

Mason, L. (2018), *Uncivil Agreement: How Politics Became Our Identity*, University of Chicago Press.

Mercier, H. and Sperber, D. (2011), 'Why do humans reason? Arguments for an argumentative theory', *Behavioral and Brain Sciences* **34**(2), 57–74.

Molho, C., Tybur, J. M., Van Lange, P. A. M. and Balliet, D. (2020), 'Direct and indirect punishment of norm violations in daily life', *Nature Communications* **11**(1), 3432.

Morales, L. (2010), Getting a single message? The impact of homogeneous political communication contexts in Spain in a comparative perspective, *in* L. Morales and M. R. Wolf, eds, 'Political Discussion in Modern Democracies: A Comparative Perspective', Routledge, pp. 201–223.

Muddiman, A. (2017), 'Personal and public levels of political incivility', *International Journal of Communication* **11**, 3182–3202.

Munger, K. (2017), 'Tweetment effects on the tweeted: Experimentally reducing racist harassment', *Political Behavior* **39**(3), 629–649.

Munger, K. (2021), 'Don't @ me: Experimentally reducing partisan incivility on Twitter', *Journal of Experimental Political Science* **8**(2), 102–116.

Neighbors, C., O'Connor, R. M., Lewis, M. A., Chawla, N., Lee, C. M. and Fossos, N. (2008), 'The relative impact of injunctive norms on college student drinking: The role of reference group', *Psychology of Addictive Behaviors* **22**(4), 576–581.

Nord, M., Altman, D., Fernandes, T., Good God, A. and Lindberg, S. I. (2026), Democracy report 2026: Unraveling the democratic era?, Technical report, V-Dem Institute.

Olar, R.-G. and Neundorf, A. (2025), 'Winds of change: Democratic transitions and long-term democratic support', *Democratization* **32**(1), 123–145.

Orhan, Y. E. (2022), 'The relationship between affective polarization and democratic backsliding: Comparative evidence', *Democratization* **29**(4), 714–735.

Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M. and Adamkovič, M. (2023), 'Perception of hate speech by the public and experts: Insights into predictors of the perceived hate speech towards migrants', *Cyberpsychology, Behavior, and Social Networking* **26**(7), 489–498.

Pasek, M. H., Ankori-Karlinsky, L.-O., Levy-Vene, A. and Moore-Berg, S. L. (2022), 'Misperceptions about out-partisans' democratic values may erode democracy', *Scientific Reports* **12**(1), 16284.

Petit, Q. (2026), 'Vox amenaza con echar "si hace falta a patadas" al presidente de RTVE', El País. https://elpais.com/comunicacion/2026-02-24/vox-amenaza-con-echar-si-hace-falta-a-patadas-al-presidente-de-rtve.html.

Pickup, M., Groenendyk, E. W., Kimbrough, E. O., Stephenson, L., French Bourgeois, L. and Harell, A. (2024), Is this how democracy dies? democratic norms and belief systems in mass publics. SSRN.

Rabellino, D., Morese, R., Ciaramidaro, A., Bara, B. G. and Bosco, F. M. (2016), 'Third-party punishment: Altruistic and anti-social behaviours in in-group and out-group settings', *Journal of Cognitive Psychology* **28**(4), 486–495.

Rashidi, Y., Kapadia, A., Nippert-Eng, C. and Su, N. M. a. (2020), 'It's easier than causing confrontation': Sanctioning strategies to maintain social norms and privacy on social media.', *Proceedings of the ACM on Human-Computer Interaction* **Proceedings of the ACM on Human-Computer Interaction**(4), 1–25.

Reiman, A.-K. and Killoran, T. C. (2023), 'When group members dissent: A direct comparison of the black sheep and intergroup sensitivity effects', *Journal of Experimental Social Psychology* **104**, 104408.

Rovny, J., Polk, J., Bakker, R., Hooghe, L., Jolly, S., Marks, G., Steenbergen, M. and Vachudova, M. A. (2025), 'The 2024 Chapel Hill expert survey on political party positioning in Europe: Twenty-five years of party positional data', *Electoral Studies* **97**, 102981.

Sääksvuori, L., Mappes, T. and Puurtinen, M. (2011), 'Costly punishment prevails in intergroup conflict', *Proceedings of the Royal Society B: Biological Sciences* **278**(1723), 3428–3436.

Sacher, A.-L. and Reinemann, C. (2025), 'What is freedom of speech? How citizens define and perceive the "bulwark of liberty"', *The International Journal of Press/Politics* pp. 1–22.

Sadowski, I. (2024), 'Democratic backsliding against a rising wave of support for democracy?', *International Sociology* **39**(6), 698–720.

Shinada, M., Yamagishi, T. and Ohmura, Y. (2004), 'False friends are worse than bitter enemies', *Evolution and Human Behavior* **25**(6), 379–393.

Simón, P. (2020), 'Two-bloc logic, polarisation and coalition government: The November 2019 general election in Spain', *South European Society and Politics* **25**(3–4), 533–563.

Simonovits, G., McCoy, J. and Littvay, L. (2022), 'Democratic hypocrisy and out-group threat: Explaining citizen support for democratic erosion', *The Journal of Politics* **84**(3), 1806–1811.

Sinclair, B. (2012), *The Social Citizen: Peer Networks and Political Behavior*, University of Chicago Press.

Spampatti, T., Globig, L. K., Rathje, S., Van Bavel, J. J. and Sjåstad, H. (2026), Preferences for speech governance are polarized across 40 countries. OSF preprint. https://osf.io/preprints/psyarxiv/9jtdn_v1.

Stefanelli, A., Abts, K. and Meuleman, B. (2025), 'Freedom for all? Populism and the instrumental support of freedom of speech', *West European Politics* pp. 1–24.

Stenner, K. (2005), *The Authoritarian Dynamic*, Cambridge University Press.

Su, S. (2022), 'Updating politicized beliefs: How motivated reasoning contributes to polarization', *Journal of Behavioral and Experimental Economics* **96**, 101799.

Szekely, A. and Rodriguez, I. (2026), 'Identifying social norms using validated survey measures'. Workshop presentation, Workshop Social norms in the Political Domain.

Taflinger, S. and Hudde, A. (2026), 'Why do young US Americans avoid cross-partisan dating? A closer look at mediators and variation by gender and party', *European Sociological Review* pp. 1–19.

Tajfel, H. and Turner, J. C. (2004), The social identity theory of intergroup behavior, *in* 'Political Psychology', Psychology Press, pp. 276–293.
URL: *https://www.taylorfrancis.com/books/9781135151355/chapters/10.4324/9780203505984-16*

Tang, S., Shepherd, S. and Kay, A. C. (2023), 'Morality's role in the black sheep effect: When and why ingroup members are judged more harshly than outgroup members for the same transgression', *European Journal of Social Psychology* **53**(7), 1605–1622.

Teorell, J., Coppedge, M., Skaaning, S.-E. and Lindberg, S. I. (2016), Measuring electoral democracy with V-Dem data: Introducing a new polyarchy index. SSRN Electronic Journal.

Tilley, J., Bejan, T. and Hobolt, S. B. (2026), 'Partisan (in)tolerance and affective polarization', *British Journal of Political Science* .

Torcal, M., Carty, E., Comellas, J. M., Bosch, O. J., Thomzon, Z. and Serani, D. (2023), 'The dynamics of political and affective polarisation: Datasets for spain, portugal, italy, argentina, and chile (2019-2022)', *Data in Brief* **48**, 109219.

Torcal, M. and Comellas, J. M. (2022), 'Affective polarisation in times of political instability and conflict. spain from a comparative perspective', *South European Society and Politics* **27**(1), 1–26.

Törnberg, P. and Rogers, R. (2026), Towards a post-social media studies. OSF preprint. https://osf.io/preprints/socarxiv/6nue7_v1.

Turnbull-Dugarte, S. J. and López Ortega, A. (2025), 'Far right normalization & centrifugal affect. Evidence from the dating market', *The Journal of Politics* .

Valentim, V. (2024), *The Normalization of the Radical Right: A Norms Theory of Political Supply and Demand*, Oxford University Press.

Van Der Brug, W., Popa, S., Hobolt, S. B. and Schmitt, H. (2021), 'Democratic support, populism, and the incumbency effect', *Journal of Democracy* **32**(4), 131–145.

Voeten, E. (2016), 'Are people really turning away from democracy?', *Journal of Democracy, Web Exchange* .

Watson, B. R., Peng, Z. and Lewis, S. C. (2019), 'Who will intervene to save news comments? Deviance and social control in communities of news commenters', *New Media & Society* **21**(8), 1840–1858.

Weisel, O. and Böhm, R. (2015), '"Ingroup love" and "outgroup hate" in intergroup conflict between natural groups', *Journal of Experimental Social Psychology* **60**, 110–120.

Westwood, S. J. (2025), 'The potential existential threat of large language models to online survey research', *Proceedings of the National Academy of Sciences: Political Sciences* **122**(0), 1–10.

Wilhelm, C. and Schulz-Tomančok, A. (2024), 'Predicting user engagement with anti-gender, homophobic and sexist social media posts – a choice-based conjoint study in Hungary and Germany', *Information, Communication & Society* **27**(11), 2094–2113.

Wunsch, N. and Gessler, T. (2023), 'Who tolerates democratic backsliding? a mosaic approach to voters' responses to authoritarian leadership in hungary', *Democratization* **30**(5).

Wunsch, N., Jacob, M. S. and Derksen, L. (2025), 'The demand side of democratic backsliding: How divergent understandings of democracy shape political choice', *British Journal of Political Science* **55**, e39.

Wuttke, A., Gavras, K. and Schoen, H. (2022), 'Have europeans grown tired of democracy? new evidence from eighteen consolidated democracies, 1981–2018', *British Journal of Political Science* **52**(1), 416–428.

Yu, X., Wojcieszak, M. and Casas, A. (2024), 'Partisanship on social media: In-Party Love among American politicians, greater engagement with Out-Party Hate among ordinary users', *Political Behavior* **46**(2), 799–824.

Zhang, A. Q., Montague, K. and Jhaver, S. (2023), Cleaning up the streets: Understanding motivations, mental models, and concerns of users flagging social media content. arXiv preprint. https://arxiv.org/abs/2309.06688.

Zhang, G., Walatka, R., Chen, S. Y., Urminsky, O., Fernandez, K., Low, A., Bogard, J. E. and Fox, C. R. (2026), Estimating the threat of AI-agent responding across online survey platforms. PsyArXiv preprint. https://osf.io/preprints/psyarxiv/xcg26_v1.

Appendix A

Figure A.1. Support among Spanish citizens for censoring media in times of crisis. Calculation using data from Torcal et al. (2023).

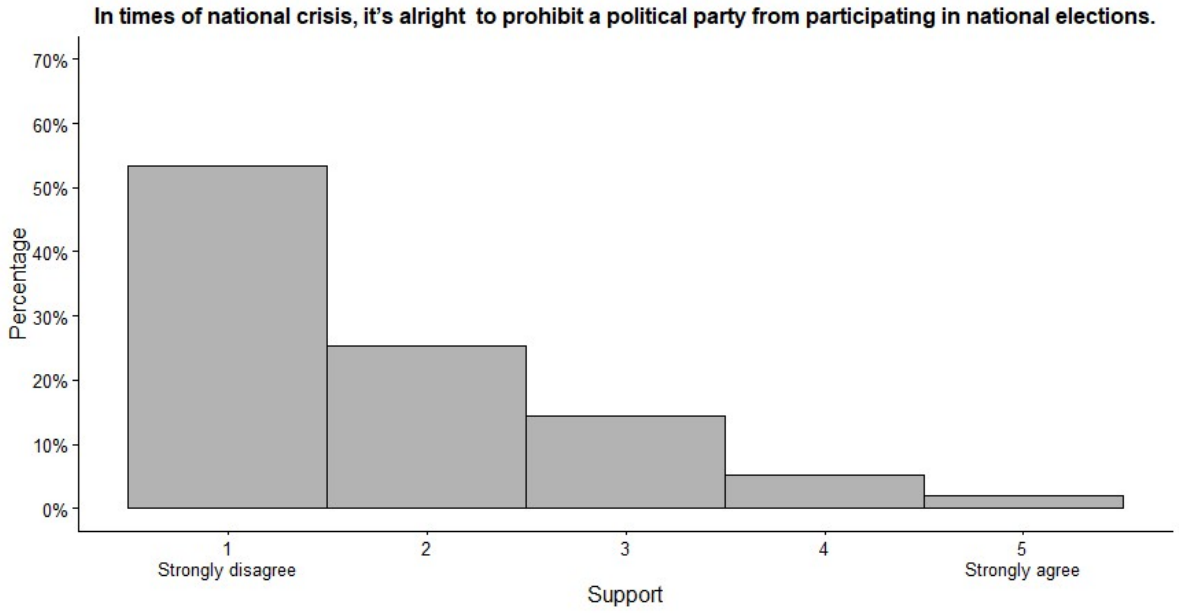


Figure A.2. Perceived importance of freedom to criticise the government for democracy among Spanish citizens. Calculation using data from Torcal et al. (2023).

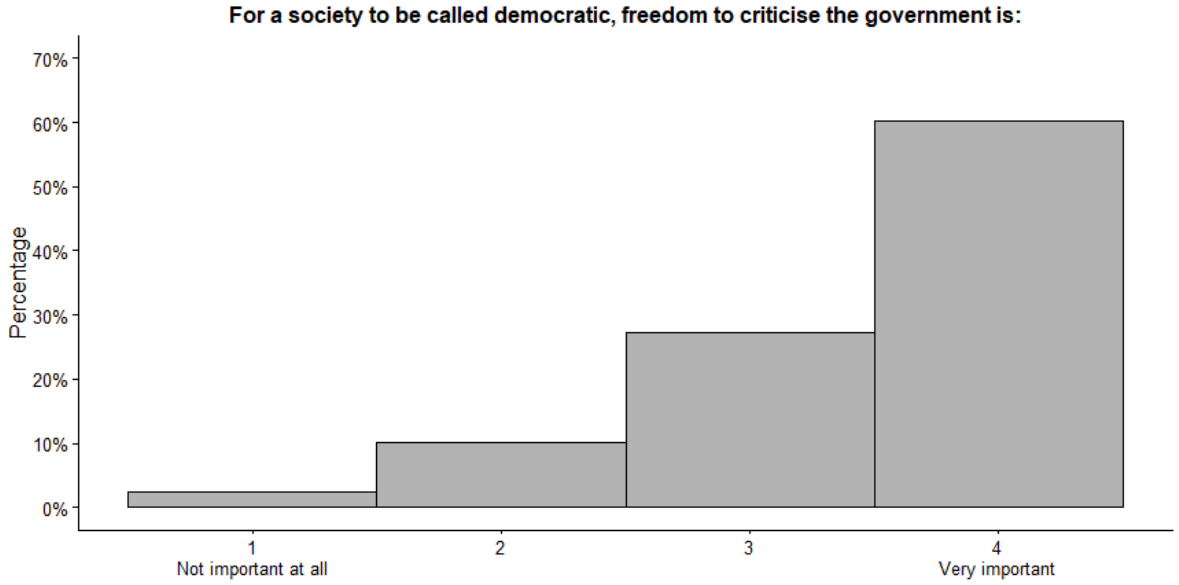


Figure A.3. Perceived importance of free and uncensored media for democracy among Spanish citizens. Calculation using data from Torcal et al. (2023).

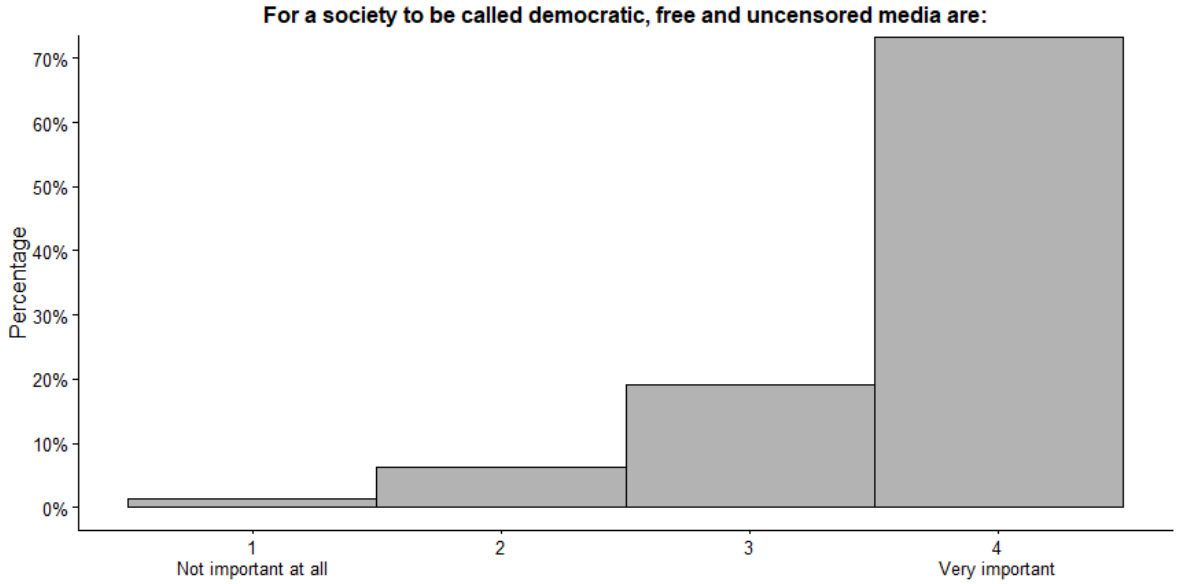
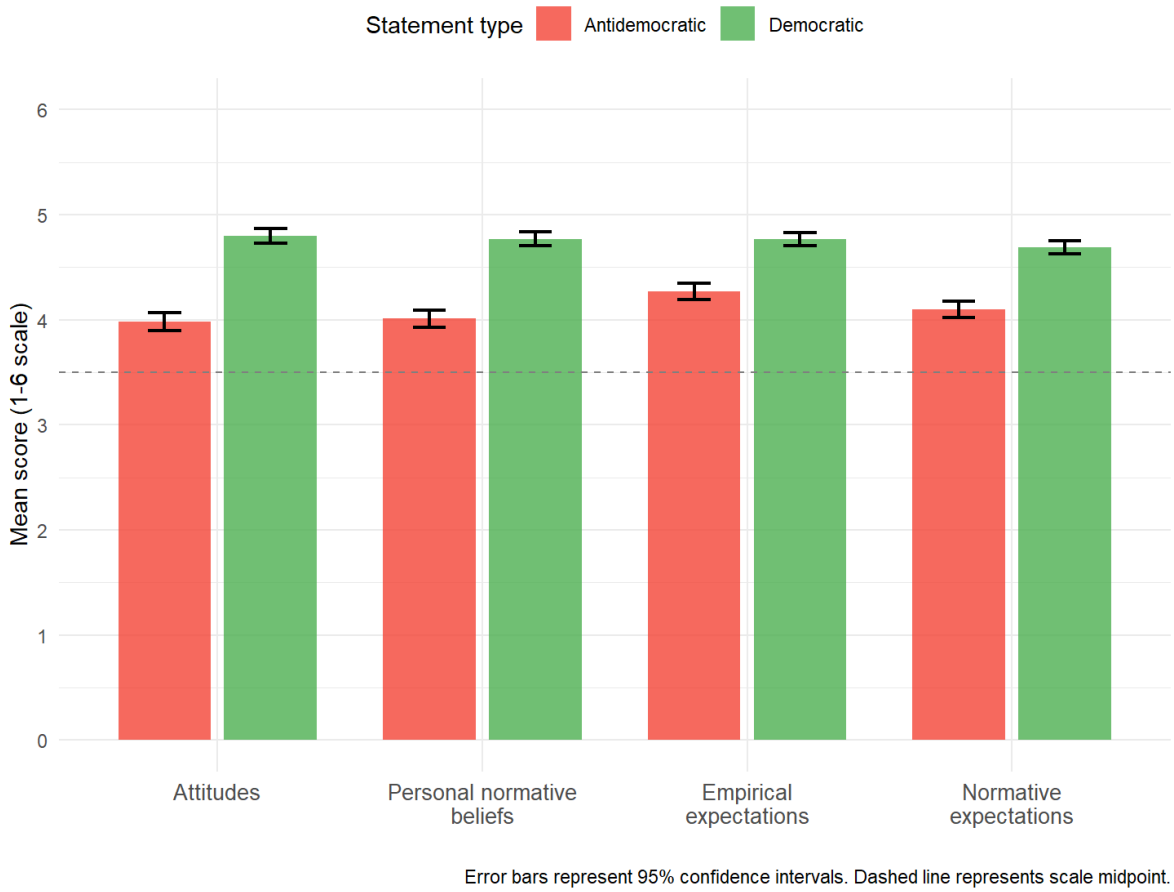


Figure A.4. Component breakdown by statement type.



Note: Attitudes not included in index.

Figure A.5. Index score by statement.

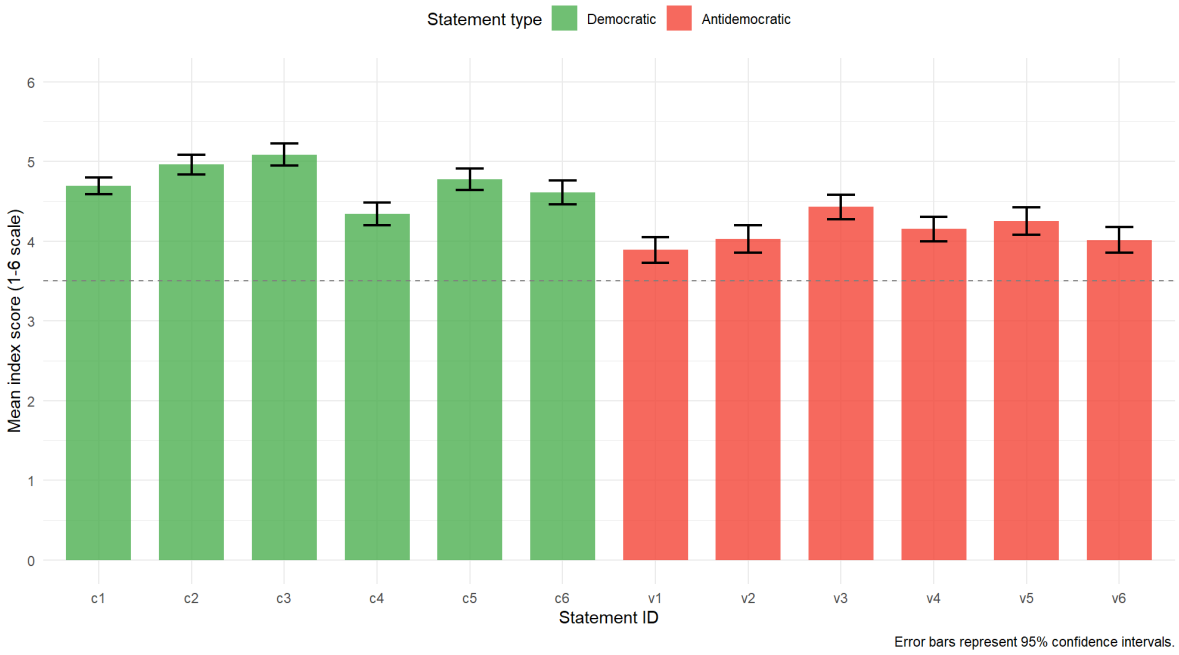


Figure A.6. Density plot, antidemocratic vs. democratic statements.

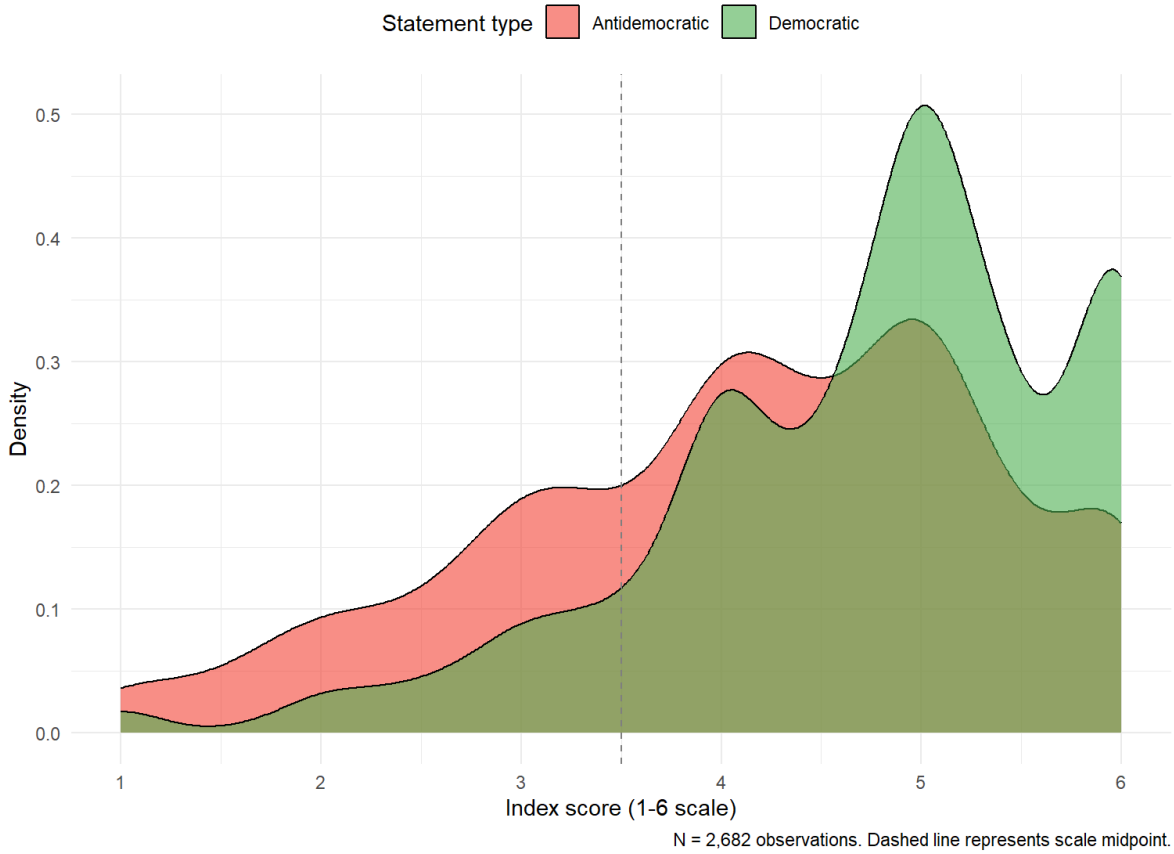
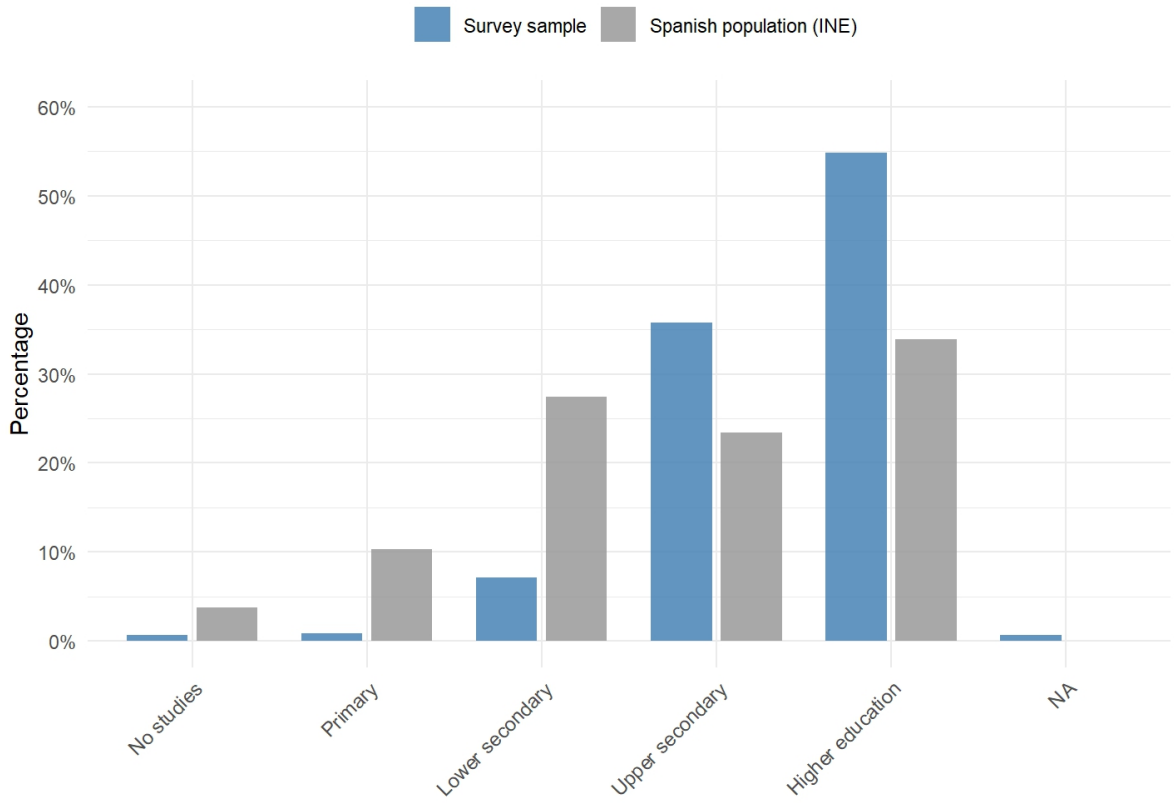
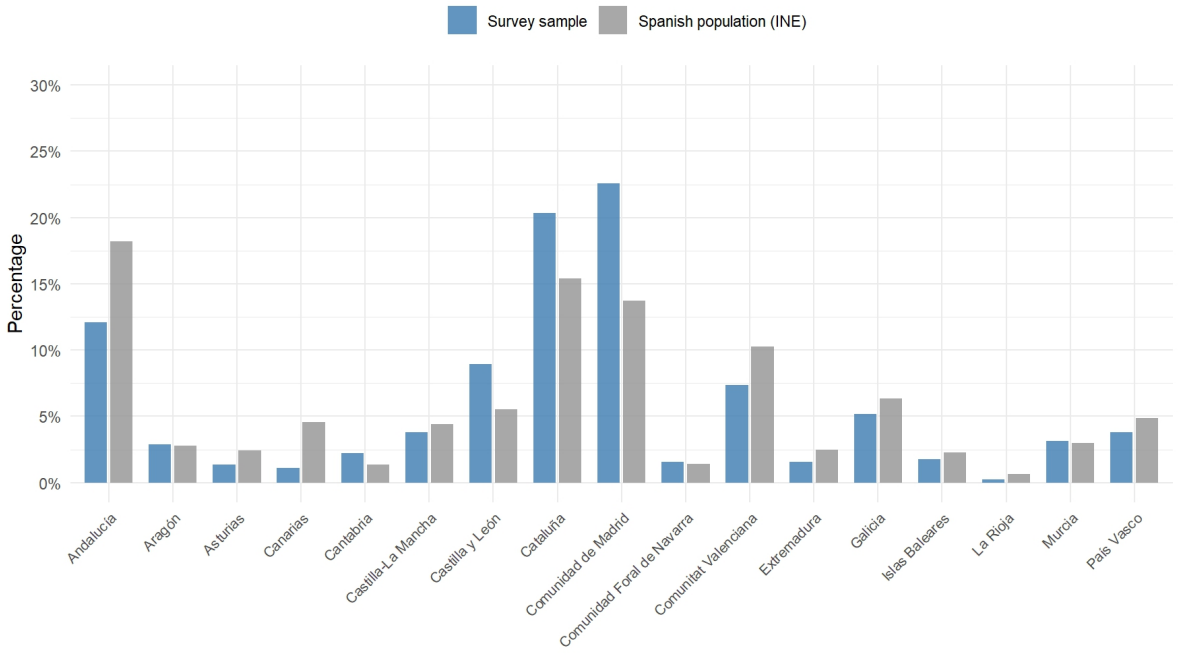


Figure A.7. Educational distribution preliminary study, sample versus population.



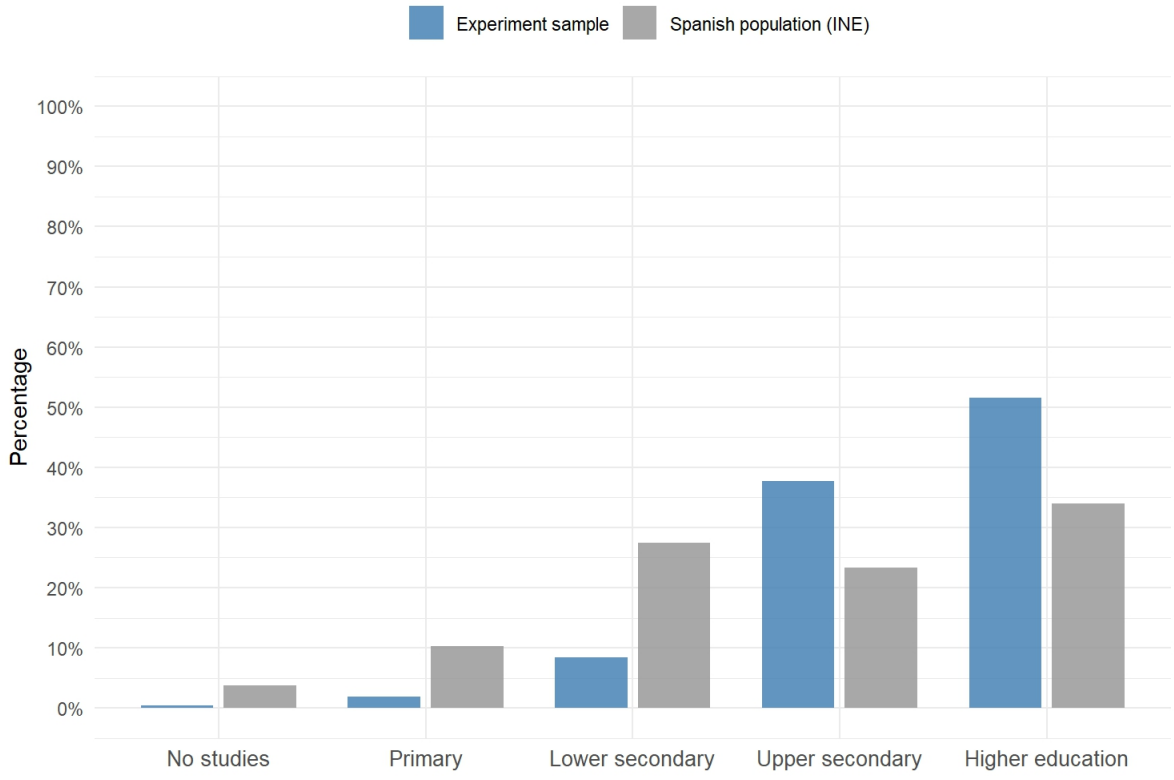
Source: Instituto Nacional de Estadística (2025).

Figure A.8. Regional distribution preliminary study, sample versus population.



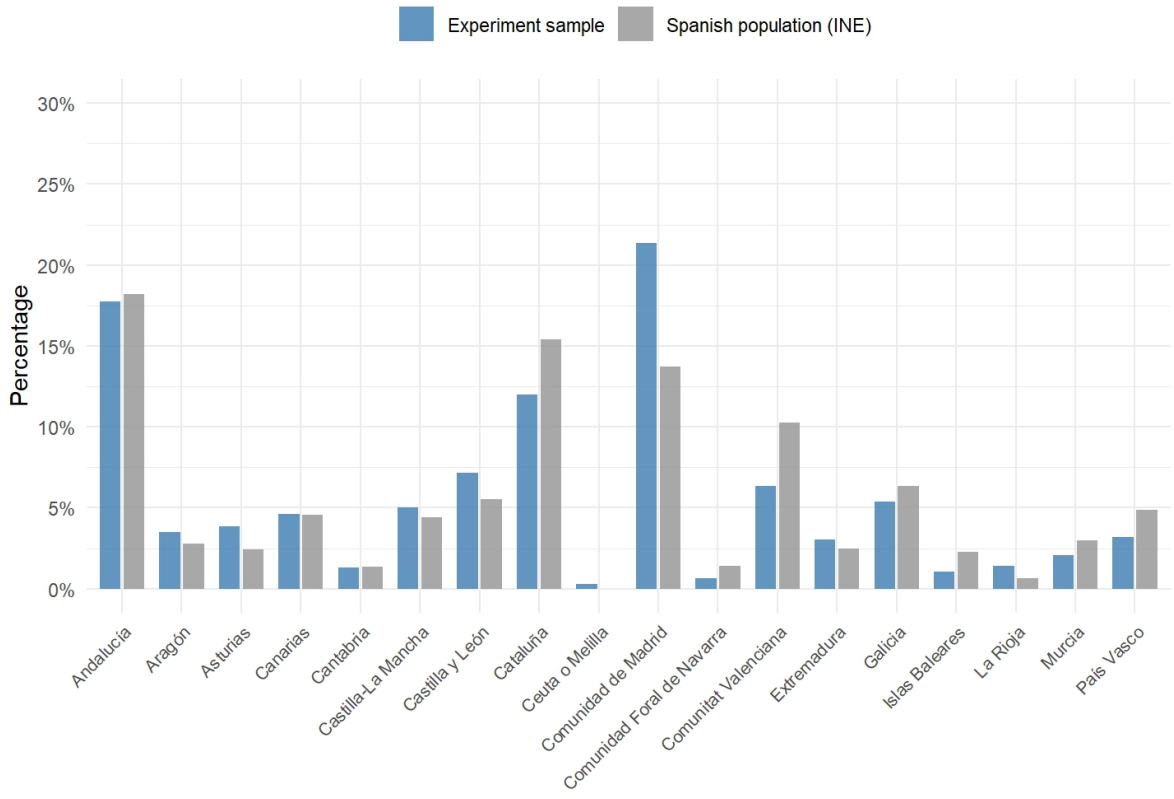
Source: Instituto Nacional de Estadística (2025).

Figure A.9. Educational distribution experiment, sample versus population.



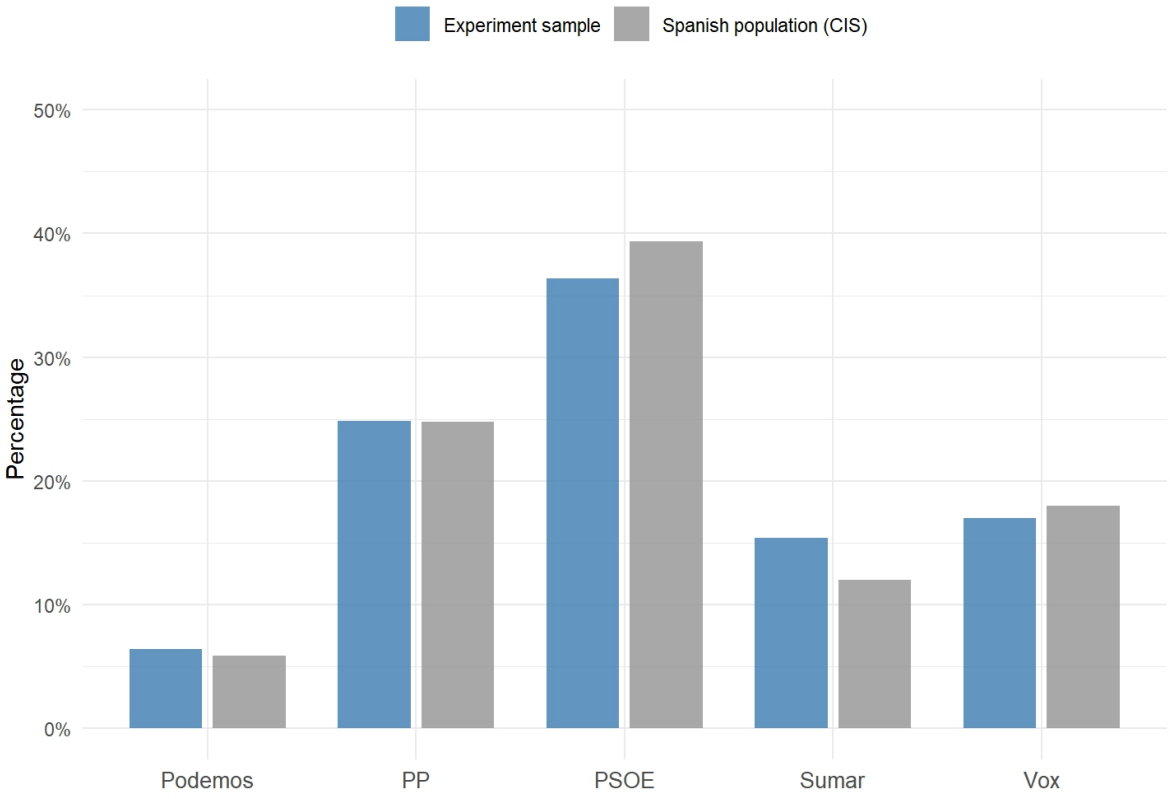
Source: INE (2023). Población de 16 y más años por nivel de formación alcanzado, sexo y grupo de edad.

Figure A.10. Regional distribution experiment, sample versus population.



Source: Instituto Nacional de Estadística (2025).

Figure A.11. Partisan distribution experiment, sample versus population.



Source: Centro de Investigaciones Sociológicas (2025). Population % rescaled to relative distribution among parties included in experiment.

Figure A.12. Partisan attribution of expressions in democratic versus antidemocratic condition.

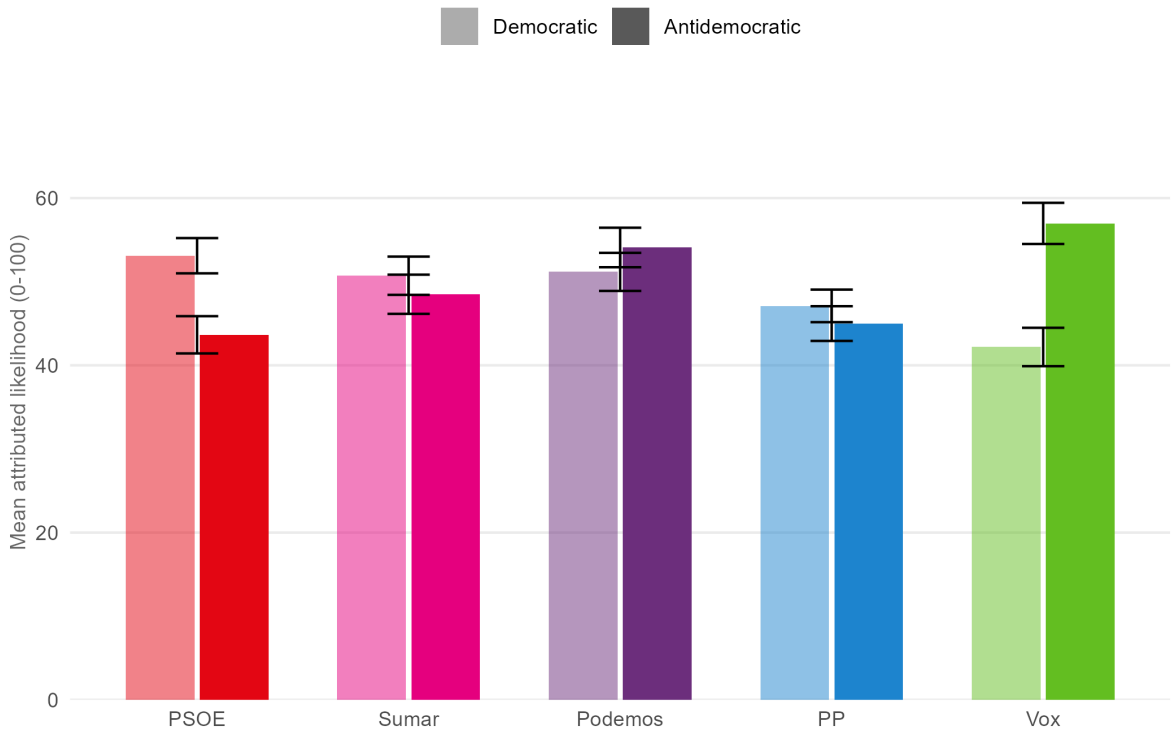


Figure A.13. Flagging rates by antidemocratic expression by condition, Podemos partisans.

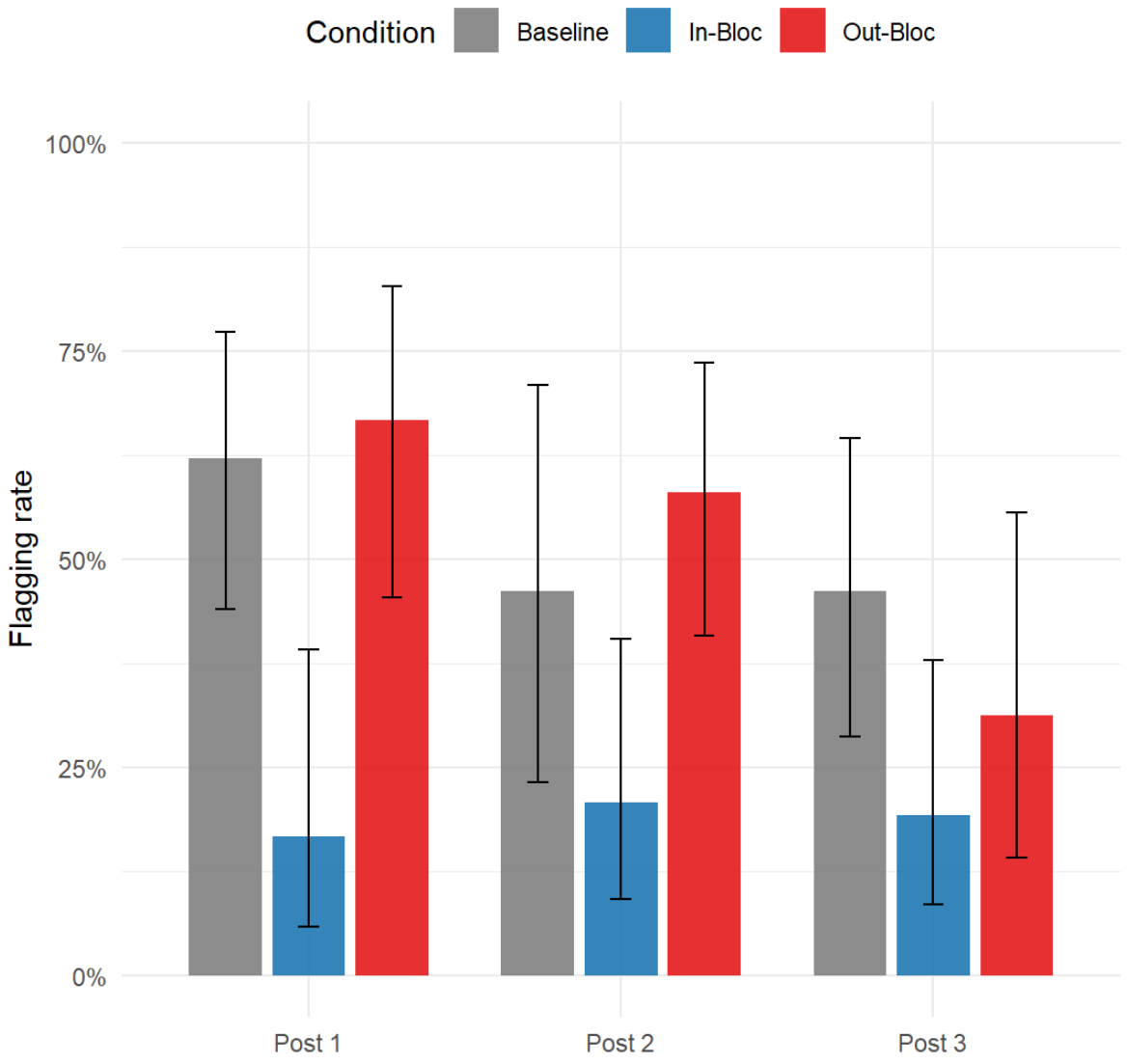


Figure A.14. Flagging rates by antidemocratic expression by condition, Sumar partisans.

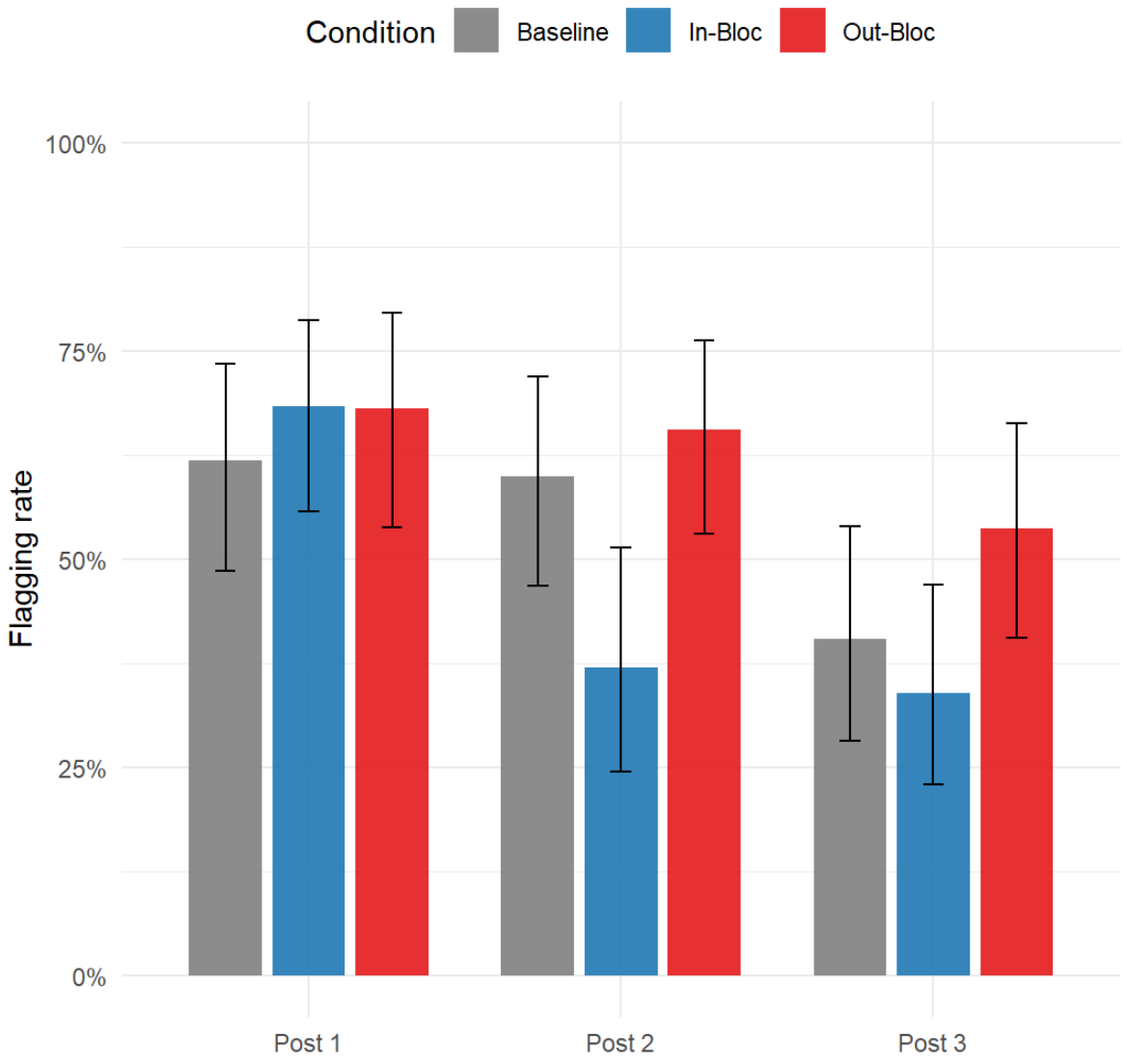


Figure A.15. Flagging rates by antidemocratic expression by condition, PSOE partisans.

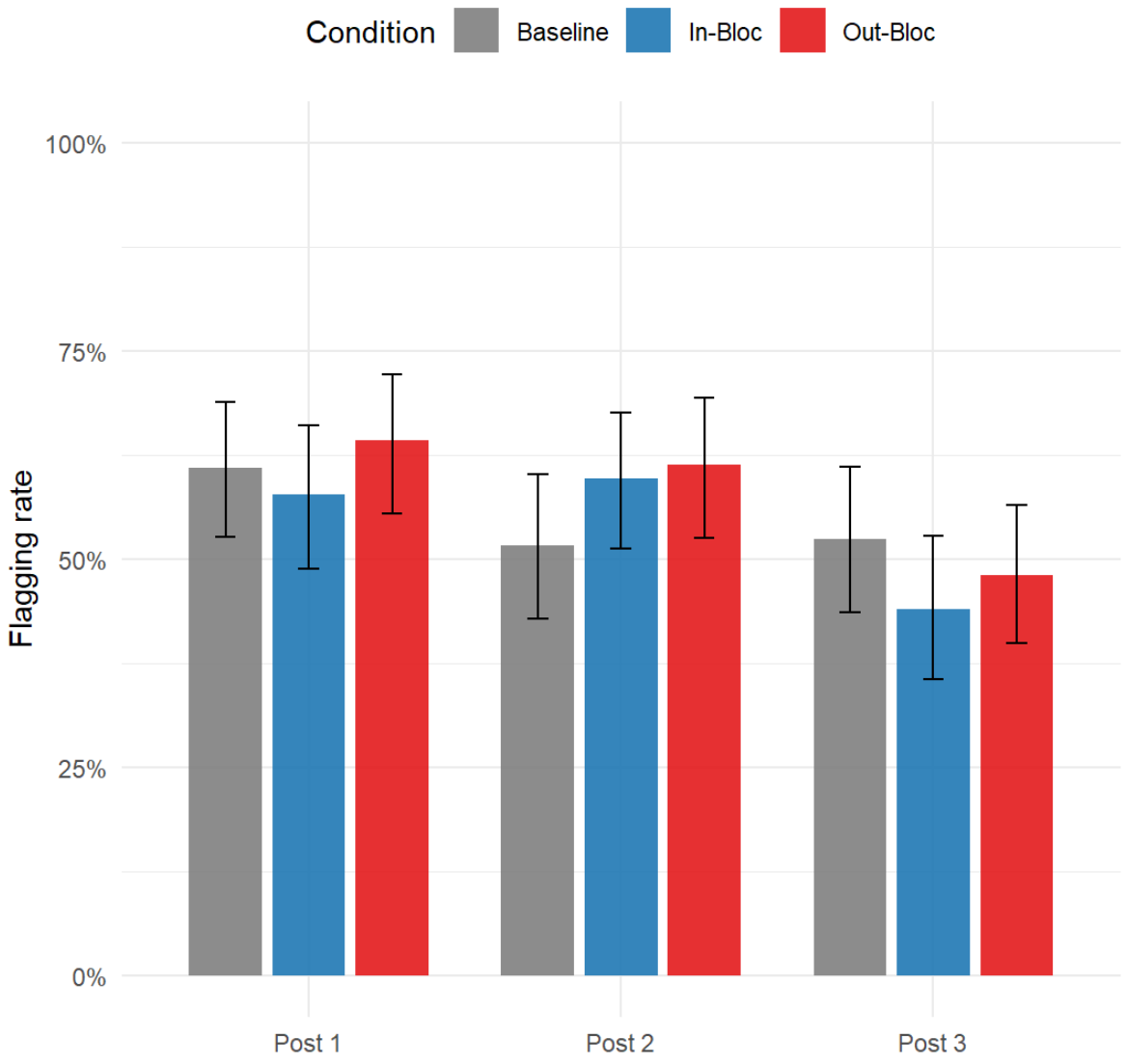


Figure A.16. Flagging rates by antidemocratic expression by condition, PP partisans.

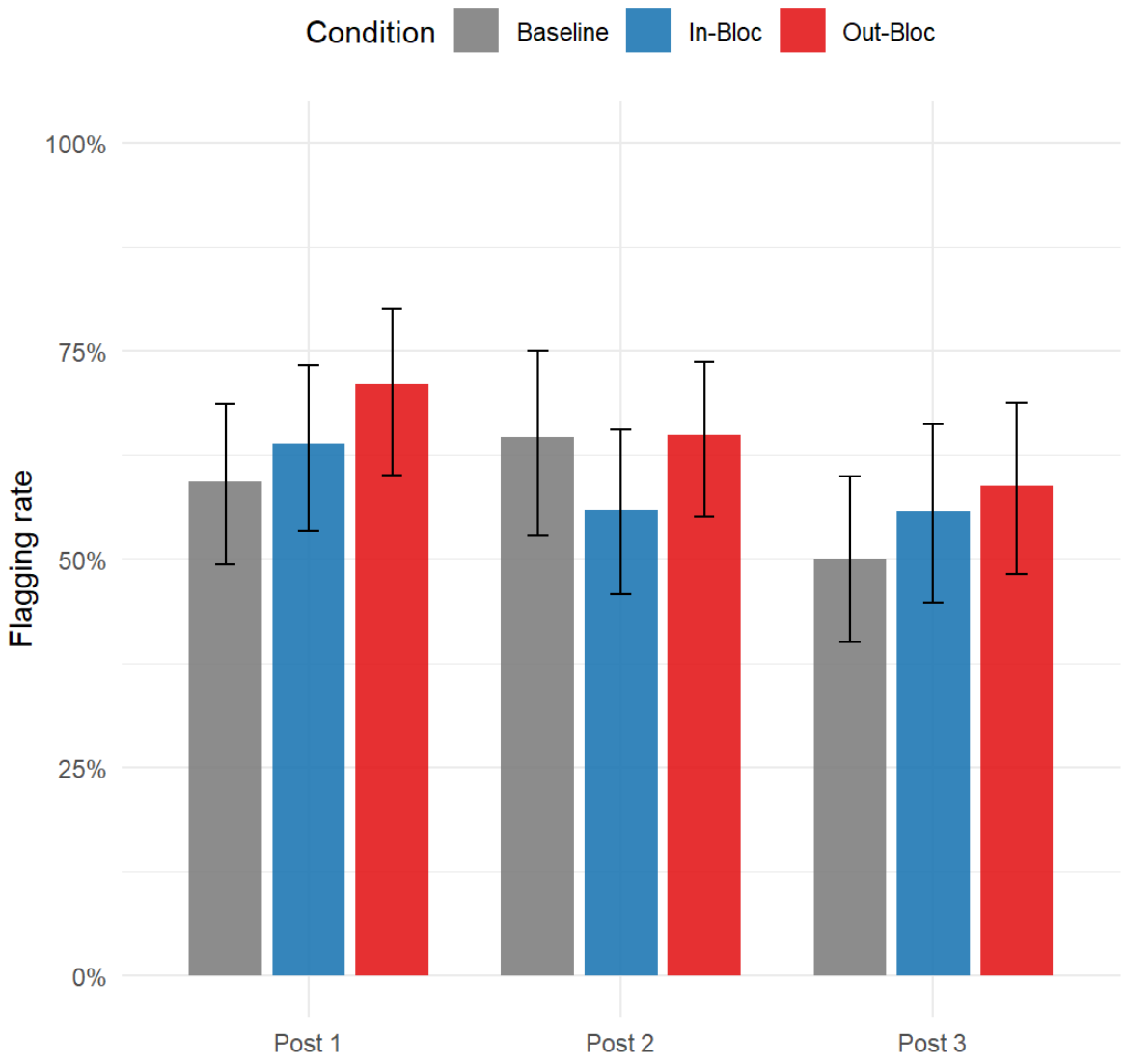
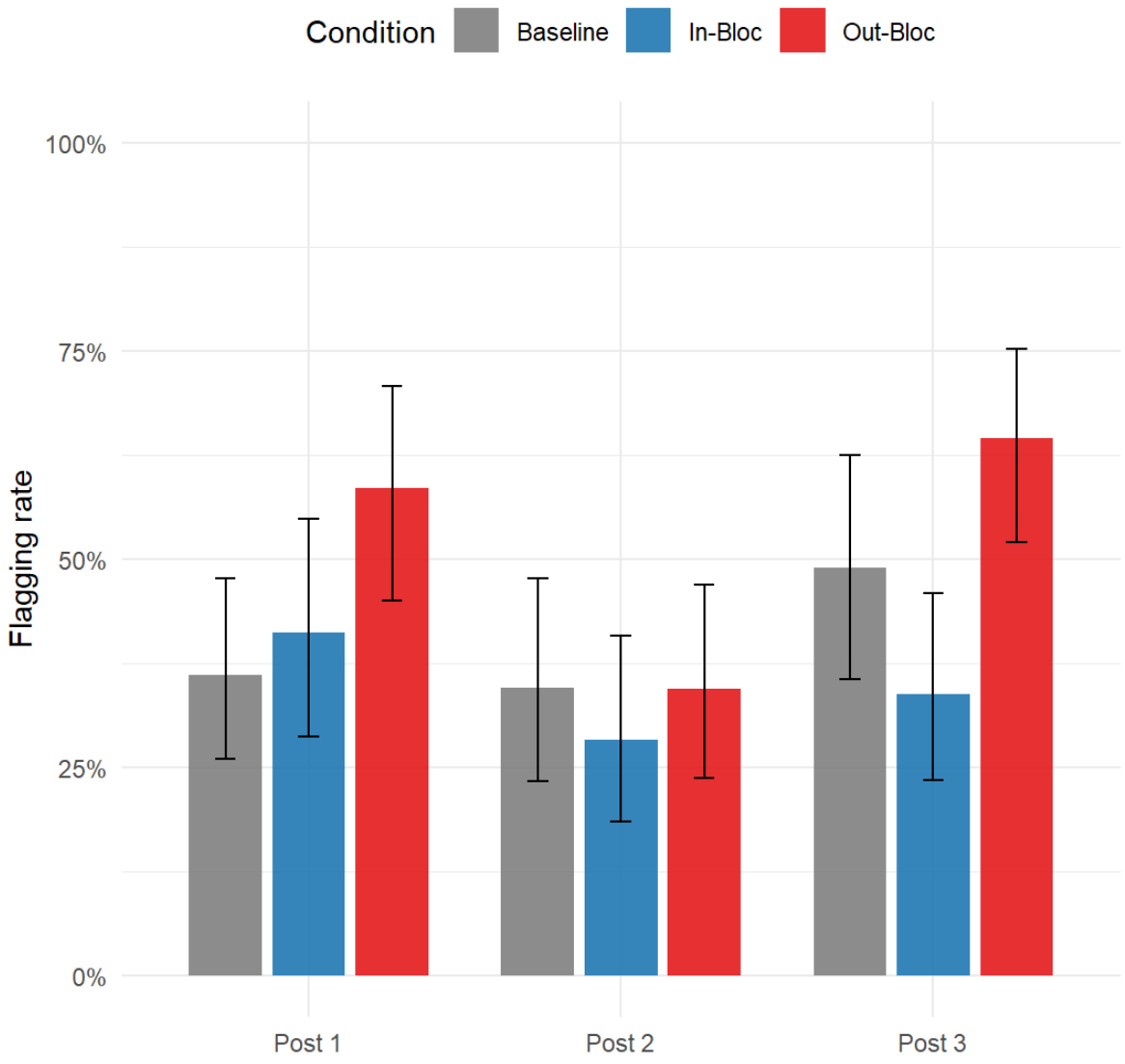


Figure A.17. Flagging rates by antidemocratic expression by condition, Vox partisans.



Appendix B

B.1: Democratic expression 1 (c1). Original alongside translation.

Usuario3113

La libertad de expresión es necesaria para ir corrigiendo lo que va saliendo mal en el proceso de desarrollo de cualquier sociedad. Cada vez que penalizan la crítica, crece la corrupción.

User3113

Freedom of expression is necessary to go about addressing things that could go wrong in any society's progress. Whenever criticism is punished, corruption will increase.

B.2: Democratic expression 2 (c2). Original alongside translation.

Usuario8126

El periodismo no es el enemigo. Por eso, el libre ejercicio de esta profesión es tan importante para el desarrollo y el buen funcionamiento de la democracia.

User8126

Journalism isn't the enemy. That's why the free exercise of this profession is important for the development and functioning of democracy.

B.3: Democratic expression 3 (c3). Original alongside translation.

Usuario2539

Amenazar y acosar a periodistas por hacer su trabajo es un ataque directo a la democracia. No deberíamos tolerar que los políticos usen el poder para intimidar a quienes informan.

User2539

Threatening and harassing journalists for doing their jobs is a direct attack on democracy. We shouldn't tolerate politicians using their power to intimidate those who keep us informed.

B.4: Democratic expression 4 (c4). Original alongside translation.

Usuario5542

Los límites a la libertad de expresión son el Código penal y la Ley de protección civil del derecho al honor... Todo lo demás es censura.

User5542

The limits of freedom of expression are the Criminal Code and the Civil Protection of the Right to Honour Act... Anything else is censorship.

B.5: Democratic expression 5 (c5). Original alongside translation.

Usuario8352	User8352
Puedes publicar lo que quieras en las redes, pero tienes que aceptar que habrá respuestas y opiniones que no siempre te van a gustar. Así funciona la democracia.	You can post anything you'd like on social media, but you have to accept that there will be replies and opinions that you won't always like. That's how democracy works.

B.6: Democratic expression 6 (c6). Original alongside translation.

Usuario4261	User4261
La libertad de expresión se refiere a poder expresarse sin la persecución del Estado, no a que alguien no te pueda llamar tonto si dices tonterías.	Freedom of expression means that you should be able to express yourself without being prosecuted by the state, not that someone can't call you stupid if you talk nonsense.

B.7: Antidemocratic expression 1 (v1). Original alongside translation.

Usuario1210	User1210
Los medios de este país no son medios de comunicación, son un aparato de agitación y de propaganda. Es hora de frenarlos.	The media in this country are not news outlets, they are a tool for agitation and propaganda. It's time to rein them in.

B.8: Antidemocratic expression 2 (v2). Original alongside translation.

Usuario7429	User7429
La prensa española está plagada de mercenarios del poder. No es periodismo, es corrupción y, como tal, debe ser perseguido.	The Spanish press is plagued by mercenaries of the establishment. That's not journalism, it's corruption and it should be persecuted as such.

B.9: Antidemocratic expression 3 (v3). Original alongside translation.

Usuario4630	User4630
Nos conviene a todos ponerle límites a la libertad de publicar. Igual que imprimir dinero en tu garaje es un delito y merece cárcel, como sociedad necesitamos tomar medidas para que la prensa no mienta.	It would benefit all of us to impose limits on the freedom of publication. Just like printing money in your garage is a crime deserving of prison, we should take measures as a society to make sure that the press doesn't lie.

B.10: Antidemocratic expression 4 (v4). Original alongside translation.

Usuario3940

Ojalá metan en la cárcel a los agitadores y a los que propagan odio. Me da mucho asco este tipo de gente.

User3940

I hope agitators and those who propagate hate will be thrown into prison. Those people disgust me.

B.11: Antidemocratic expression 5 (v5). Original alongside translation.

Usuario2751

España necesita urgentemente un cambio en los medios de comunicación. Entre esos cambios haría falta la ilegalización de la prensa manipuladora.

User2751

Spanish media urgently need change. One of the changes needed would be outlawing the manipulative press.

B.12: Antidemocratic expression 6 (v6). Original alongside translation.

Usuario6463

Quien reparte odio provoca asesinatos. Si instas a odiar a millones de ciudadanos, que te persigan como cómplice de la violencia que provocas.

User6463

Those who sow hatred provoke murders. If you encourage people to hate millions of citizens, you should be prosecuted as an accomplice in the violence you provoke.

B.13: Antidemocratic expression v1, shown in the Podemos, Vox and no-cue condition.

Appendix C

Table C.1. Means, standard deviations, and correlations of index components with confidence intervals.

Variable	*M*	*SD*	1	2	3
1. Attitudes	4.39	1.53			
2. Personal normative beliefs	4.39	1.42	.69** [.67, .71]		
3. Empirical expectations	4.52	1.32	.55** [.53, .58]	.52** [.50, .55]	
4. Normative expectations	4.39	1.35	.47** [.44, .50]	.61** [.58, .63]	.77** [.75, .78]

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation.

Table C.2. Crossed random effects model: antidemocratic vs. democratic expressions.

	Unweighted		**Weighted**	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
Intercept	4.75***	4.57 – 4.93	4.73***	4.58 – 4.87
Antidemocratic content	−0.62***	−0.87 – −0.36	−0.59***	−0.80 – −0.39
Random Effects				
σ^2	0.79		0.77	
τ_{00} participant	0.29		0.32	
τ_{00} expression	0.04		0.02	
ICC	0.39		0.39	
Observations	2,682		2,682	
Marginal R^2 / Conditional R^2	0.068 / 0.435		0.065 / 0.425	

Note. $N_{\text{participants}} = 447$; $N_{\text{expressions}} = 12$. $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.3. Disaggregated analysis of antidemocratic content on index components.

	Personal norm. beliefs		**Empirical expectations**		**Normative expectations**	
Predictors	*Est.*	*CI*	*Est.*	*CI*	*Est.*	*CI*
Intercept	4.74^{***}	$4.56 - 4.93$	4.76^{***}	$4.58 - 4.94$	4.68^{***}	$4.56 - 4.79$
Antidem. content	-0.70^{***}	$-0.96 - -0.44$	-0.50^{***}	$-0.76 - -0.25$	-0.58^{***}	$-0.74 - -0.41$
N participants	447		447		447	
N expressions	12		12		12	
Observations	2682		2682		2682	
Marg. R^2 / Cond. R^2	0.064 / 0.401		0.037 / 0.361		0.048 / 0.395	

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

Table C.4. Percentage of voters answering "No" to the question "would you sit down for a drink with someone who votes. . . "

	Podemos	*Sumar*	*PSOE*	*PP*	*Vox*
Podemos voters	0.0	7.8	26.8	57.6	66.5
Sumar voters	11.9	1.1	16.5	54.6	63.3
PSOE voters	12.1	8.4	1.4	38.3	65.8
PP voters	54.0	30.0	40.8	0.6	29.5
Vox voters	85.4	65.1	64.5	26.4	2.4

Source: El País & SER (2024). Representative survey data ($N = 2000$) gathered by 40db.

Table C.5. Party positions on key ideological dimensions (CHES 2024), 0–10 scale.

Party	*General left-right*	*Economic*	*GAL/TAN*	*Nationalism*
Podemos	1.56	1.62	1.44	2.38
PSOE	3.81	3.94	2.94	4.12
Sumar	2.31	2.27	1.20	2.12
Vox	9.38	9.00	9.50	9.62
PP	7.19	7.25	7.00	6.38

Source: Chapel Hill Expert Survey (Rovny et al., 2025).

Table C.6. Mixed-effect logistic regression, Podemos subset: effect of in-bloc and out-bloc cues on flagging probability (baseline: no cue).

Predictors	*Odds Ratios*	*CI*
Intercept	1.16	0.63 – 2.12
In-Bloc	0.16***	0.06 – 0.39
Out-Bloc	1.07	0.50 – 2.27
Random Effects		
τ_{00} part_id	0.93	
τ_{00} item_id	0.03	

Note. 204 observations, 68 participants.

$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.7. Mixed-effect logistic regression, Sumar subset: effect of in-bloc and out-bloc cues on flagging probability (baseline: no cue).

Predictors	*Odds Ratios*	*CI*
Intercept	1.23	0.65 – 2.32
In-Bloc	0.71	0.43 – 1.15
Out-Bloc	1.55^{+}	0.95 – 2.54
Random Effects		
τ_{00} part_id	0.89	
τ_{00} item_id	0.21	

Note. 486 observations, 162 participants.

$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.8. Mixed-effect logistic regression, PSOE subset: effect of in-bloc and out-bloc cues on flagging probability (baseline: no cue).

Predictors	*Odds Ratios*	*CI*
Intercept	1.28	0.89 – 1.84
In-Bloc	0.94	0.69 – 1.29
Out-Bloc	1.14	0.83 – 1.57
Random Effects		
τ_{00} part_id	0.91	
τ_{00} item_id	0.06	

Note. 1146 observations, 382 participants.

$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.9. Mixed-effect logistic regression, PP subset: effect of in-bloc and out-bloc cues on flagging probability (baseline: no cue).

Predictors	*Odds Ratios*	*CI*
Intercept	1.42^{*}	$1.02 - 1.97$
In-Bloc	1.05	$0.72 - 1.53$
Out-Bloc	1.44^{+}	$0.98 - 2.11$
Random Effects		
τ_{00} part_id	0.70	
τ_{00} item_id	0.02	

Note. 774 observations, 258 participants.

$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.10. Mixed-effect logistic regression, Vox subset: effect of in-bloc and out-bloc cues on flagging probability (baseline: no cue).

Predictors	*Odds Ratios*	*CI*
Intercept	0.61^{*}	$0.38 - 0.97$
In-Bloc	0.78	$0.49 - 1.24$
Out-Bloc	1.84^{**}	$1.16 - 2.91$
Random Effects		
τ_{00} part_id	0.52	
τ_{00} item_id	0.08	

Note. 528 observations, 176 participants.

$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.11. Robustness check: effect of in-bloc and out-bloc cues on flagging probabilities. Main model displayed next to model with post-level fixed effects.

	Random effects		**Fixed effects**	
Predictors	*Log-Odds*	*CI*	*Log-Odds*	*CI*
Intercept	0.12	−0.16 − 0.40	0.41***	0.23 − 0.59
In-Bloc	−0.19*	−0.38 − −0.00	−0.19*	−0.38 − −0.00
Out-Bloc	0.32***	0.13 − 0.51	0.33***	0.14 − 0.52
Post 2 (viol2)			−0.31**	−0.50 − −0.12
Post 3 (viol3)			−0.55***	−0.74 − −0.36
Random Effects				
σ^2	3.29		3.29	
τ_{00} part_id	0.80		0.81	
τ_{00} item_id	0.04			
Observations	3138		3138	

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

Table C.13. Mixed-effect logistic regression: effect of in-bloc and out-bloc cue on flagging probability, with Vox dummy (baseline: no cue).

Predictors	*Log-Odds*	*CI*
Intercept	0.25+	−0.03 − 0.54
In-Bloc	−0.19+	−0.40 − 0.01
Out-Bloc	0.25*	0.04 − 0.46
Vox Dummy	−0.81***	−1.19 − −0.42
In-Bloc × Vox	−0.02	−0.53 − 0.50
Out-Bloc × Vox	0.43+	−0.08 − 0.93
Random Effects		
τ_{00} part_id	0.75	
τ_{00} item_id	0.05	

Note. 3138 observations, 1046 participants.

$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.12. Item-level sensitivity analysis: effect of in-bloc and out-bloc cues on flagging probabilities.

	Full sample		**Excl. Post 1**		**Excl. Post 2**		**Excl. Post 3**	
Predictors	*OR*	*CI*	*OR*	*CI*	*OR*	*CI*	*OR*	*CI*
Intercept	1.13	0.85 – 1.49	1.04	0.85 – 1.28	1.11	0.78 – 1.57	1.25	0.91 – 1.71
In-Bloc	0.82*	0.68 – 1.00	0.76*	0.60 – 0.95	0.86	0.70 – 1.07	0.93	0.71 – 1.22
Out-Bloc	1.38***	1.14 – 1.67	1.25+	1.00 – 1.56	1.35**	1.09 – 1.67	1.70***	1.29 – 2.24
Random Effects								
σ^2	3.29		3.29		3.29		3.29	
τ_{00} part_id	0.80		0.28		0.08		2.54	
τ_{00} item_id	0.04		0.01		0.05		0.03	
Observations	3138		2092		2092		2092	

$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.14. Mixed-effect logistic regression: effect of in-bloc and out-bloc cue on flagging probability, with Podemos dummy (baseline: no cue).

Predictors	*Log-Odds*	*CI*
Intercept	0.12	$-0.16 - 0.40$
In-Bloc	-0.10	$-0.29 - 0.10$
Out-Bloc	0.34^{***}	$0.14 - 0.54$
Podemos Dummy	0.01	$-0.58 - 0.59$
In-Bloc × Podemos	-1.68^{***}	$-2.53 - -0.83$
Out-Bloc × Podemos	-0.27	$-1.03 - 0.50$
Random Effects		
τ_{00} part_id	0.81	
τ_{00} item_id	0.04	

Note. 3138 observations, 1046 participants.

$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.15. Mixed-effect logistic regression: effect of in-bloc and out-bloc cue on flagging probability, with Vox dummy and demographic controls (baseline: no cue).

Predictors	*Log-Odds*	*CI*
Intercept	-1.08^{**}	$-1.81 - -0.34$
In-Bloc	-0.19^{+}	$-0.40 - 0.01$
Out-Bloc	0.25^{*}	$0.04 - 0.46$
Vox Dummy	-0.70^{***}	$-1.08 - -0.31$
Gender	-0.42^{***}	$-0.61 - -0.24$
Age	0.01^{***}	$0.01 - 0.02$
Education	0.20^{**}	$0.07 - 0.33$
In-Bloc × Vox	-0.01	$-0.53 - 0.50$
Out-Bloc × Vox	0.43^{+}	$-0.07 - 0.94$
Random Effects		
τ_{00} part_id	0.66	
τ_{00} item_id	0.05	

Note. 3123 observations, 1041 participants.

$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.16. Mixed-effect logistic regression: effect of in-bloc and out-bloc cue on flagging probability, with Podemos dummy and demographic controls (baseline: no cue).

Predictors	*Log-Odds*	*CI*
Intercept	-1.35^{***}	$-2.08 - -0.62$
In-Bloc	-0.10	$-0.29 - 0.10$
Out-Bloc	0.34^{***}	$0.14 - 0.54$
Podemos Dummy	0.02	$-0.57 - 0.60$
Gender	-0.46^{***}	$-0.64 - -0.27$
Age	0.01^{***}	$0.01 - 0.02$
Education	0.23^{***}	$0.10 - 0.35$
In-Bloc × Podemos	-1.68^{***}	$-2.53 - -0.82$
Out-Bloc × Podemos	-0.26	$-1.03 - 0.51$
Random Effects		
τ_{00} part_id	0.70	
τ_{00} item_id	0.05	

Note. 3123 observations, 1041 participants.

$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.17. Mixed-effects logistic regression: effect of antidemocratic content on flagging probability, in-bloc and out-bloc cue moderators (baseline: democratic posts).

Predictors	*Odds Ratios*	*CI*
Intercept	0.11^{***}	$0.08 - 0.16$
Antidemocratic	9.73^{***}	$6.25 - 15.17$
In-Bloc	0.92	$0.71 - 1.20$
Out-Bloc	1.57^{***}	$1.23 - 2.00$
Antidemocratic × In-Bloc	0.90	$0.66 - 1.25$
Antidemocratic × Out-Bloc	0.87	$0.64 - 1.18$
Random Effects		
σ^2	3.29	
τ_{00} part_id	0.49	
τ_{00} item_id	0.06	
N part_id	1046	
N item_id	6	
Observations	6276	

Note. 6276 observations, 1046 participants.
$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.18. Robustness check: effect of in-bloc and out-bloc cue excluding participants who flagged out-bloc democratic posts.

Predictors	*Odds Ratios*	*CI*
Intercept	1.03	$0.75 - 1.42$
In-Bloc	0.90	$0.73 - 1.11$
Out-Bloc	1.50^{***}	$1.21 - 1.85$
Random Effects		
τ_{00} part_id	0.97	
τ_{00} item_id	0.06	

Note. 2583 observations, 861 participants.
$^{+}p < 0.10$ $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$.

Table C.19. Linear mixed model: effect of antidemocratic content on individual-level entropy (baseline: democratic posts).

Predictor	*Estimate*	*se*	*CI*
(Intercept)	0.837^{***}	0.01	$0.82 - 0.85$
Antidemocratic	-0.017^{**}	0.01	$-0.03 - -0.00$
Random Effects			
σ^2	0.02		
τ_{00} respondent	0.02		
ICC	0.47		
N respondent	255		
Observations	1511		
Marginal R^2 / Conditional R^2		0.002 / 0.467	

$^{*}p < 0.05 \quad ^{**}p < 0.01 \quad ^{***}p < 0.001$.

Table C.20. Aggregate post-level entropy reported by condition.

Category	*Entropy*	*Min – max*	*Posts*
Antidemocratic	0.986	$0.981 - 0.992$	3
Democratic	0.996	$0.991 - 1.000$	3
All posts	0.991	$0.981 - 1.000$	6

Appendix D

Below, we report an anonymised version of the preregistration of the preliminary and main study, uploaded to OSF in November 2025. Here, we note deviations from the preregistration, in descending order of importance:

- A post-survey fielded on Prolific in July 2026 has been added to the section in which we report exploratory findings (see Appendix E); it is labelled as exploratory.
- H1, the preregistered hypothesis predicting that antidemocratic expressions are perceived as norm violations, has been demoted to a pre-analysis check. Note that this hypothesis is supported by findings from the preliminary study, clearing the preregistered SESOI.
- Due to a copying error, the preregistered hypotheses grouped incorrectly in the hypotheses section of the preregistration. This has been fixed in the manuscript, along with minor framing adjustments. Hypotheses 2a and 2b thus correspond to H1a and H1b (in-group hypotheses) in the paper; hypotheses H3a and H3b to hypotheses H2a and H2b (out-group hypotheses). The competing predictions themselves remain unchanged.
- Due to a higher proportion of respondents failing the attention checks, the final sample of the experiment consists of 1046 participants, slightly below the preregistered N. Results remain similar if we use the full sample.
- Given that Limesurvey does not allow for the identification of the order in which questions were answered when items are presented in random order, we have dropped the straightlining exclusion. Given that no participant failed the preregistered speeding check, no participants have been excluded for this.

Preregistration

Description

A growing body of scholarship has investigated the increasing normalisation of antidemocratic expressions and attitudes (Álvarez-Benjumea & Valentim, 2024; Dinas et al., 2024; Valentim, 2024). It is often found that this normalisation does not occur through a rise in antidemocratic attitudes, but rather due to erosion of democratic norms (Álvarez-Benjumea & Valentim, 2024; Álvarez-Benjumea & Winter, 2020; Bursztyn et al., 2020). One of the ways in which democratic norm erosion can be prevented is through norm enforcement between citizens (Álvarez-Benjumea & Valentim, 2024; Harteveld et al., 2019). Specifically, social sanctioning could serve as an important deterrent of the expression of contra-normative attitudes (Bicchieri, 2017; Munger, 2017). Nevertheless, research on the determinants of citizen political norm enforcement remains scarce (Helmke & Rath, 2025).

In a twofold design, this investigation explores why these norms are often not enforced among citizens (Álvarez-Benjumea & Valentim, 2025). It tests two possible explanations: firstly, that citizens might not recognise expressions violating democratic norms as such if made by citizens (Berk & Kollberg, upcoming). Secondly, this investigation explores the possibility that citizens selectively enforce democratic norms based on whether they share political identity with the norm violator. Using freedom of expression in Spain as a case study, this investigation will test the first explanation through an incentivised survey and the second explanation through an online experiment. Both studies will be conducted through Netquest, with a Spanish sample representative in terms of gender and age.

The first study, the incentivised survey, serves to answer RQ1 ("Do citizens who express themselves antidemocratically violate social norms?"). In this study, respondents will be exposed to three fictional social media posts containing expressions opposing freedom of expression and three posts on the same topic, but in line with this principle. Using the framework outlined by Bicchieri (2017), it will then be assessed whether the antidemocratic content of a post predicts higher scores on an index of items designed to measure social norm violations, consisting of personal normative beliefs, empirical expectations and normative expectations (Álvarez-Benjumea & Valentim, 2024; Turnbull-Dugarte, 2025). Extrapolating previous findings on citizen sanctioning of radical-right expressions, it is hypothesised that citizens do recognise antidemocratic expressions as social norm violations.

The second study, the online experiment, investigates RQ2 ("How does the political identity of citizens violating democratic norms affect willingness to socially sanction them?"). This study will use the antidemocratic posts deemed most inappropriate in Study 1 in a within-person experiment. Specifically, after filling in a pre-experimental survey containing

a question on political identity, respondents are presented with the three antidemocratic posts rated as most inappropriate in Study 1. Subsequently, they are instructed to aid the researchers in flagging posts containing inappropriate content. The manipulation consists of party cues accompanying the posts: either the Spanish radical right (Vox), the Spanish radical left (Podemos), or no party cue at all. Assignment of party cues is random; three democratic posts with random allocation of party cues will also be provided to exclude the possibility that citizens solely sanction these cues. It is expected that the presence of these party cues affects respondents' probability of flagging posts as inappropriate, although split hypotheses are formulated on the direction of this relationship. Firstly, literature on democratic hypocrisy (Gidron et al., 2025; Juratic et al., 2025; Simonovits et al., 2022) and democratic rationalisation (Krishnarajan, 2023) leads to the expectation that in-group political cues are associated with decreased willingness to socially sanction norm violators (H2a) while out-group cues are associated with increased willingness to do so (H3a). Conversely, combining findings on the black sheep effect (Marques & Paez, 1994; Pinto et al., 2010) and literature on social proximity and norm enforcement (Bicchieri et al., 2022; Moisuc & Brauer, 2019), it is hypothesised that in-group (out-group) political cues are associated with increased (decreased) willingness to socially sanction norm violators (H2b; H3b).

Hypotheses

H1: Publications by citizens containing antidemocratic expressions are perceived as social norm violations.

Democratic hypocrisy hypotheses:

H2a: In-group political cues are associated with decreased willingness to socially sanction norm violators.

H2b: In-group political cues are associated with increased willingness to socially sanction norm violators.

Black sheep effect hypotheses:

H3a: Out-group political cues are associated with increased willingness to socially sanction norm violators.

H3b: Out-group political cues are associated with decreased willingness to socially sanction norm violators.

Study Design

Study 1: Within-subjects design with 1 factor (political content) and 2 levels (democratic, antidemocratic). The design employs a crossed stimulus structure where participants evalu-

ate a random subset of 6 posts drawn from a larger pool of 12.

Study 2: Within-subjects experimental design with 1 factor (political cue) and 3 levels (in-bloc, out-bloc, no cue).

In both studies, the order of the posts is fully randomised to account for order and fatigue effects. In Study 2, political cues are randomly assigned to each post; norm-concordant posts with cues are included as a discriminant validity check.

Randomisation

Study 1: The survey software employs stratified random selection to draw the stimuli. For each participant, the software randomly selects 3 posts from the pool of 6 democratic posts and 3 posts from the pool of 6 antidemocratic posts. The presentation order of these 6 selected posts is fully randomised (simple randomisation).

Study 2: The survey software employs simple randomisation to assign the treatment conditions. For each participant, the three treatment conditions (in-bloc cue, out-bloc cue, no cue) are randomly assigned to the fixed stimuli. The six publications are presented in randomised order (simple randomisation).

Data Collection Procedures

Both studies will be conducted on the online platform Limesurvey, with a sample representative in terms of gender and age recruited from the Netquest panel. Study 1 will be conducted prior to Study 2 in December 2025, with a space of no more than two weeks between the implementation of the two studies. Respondents will receive compensation for participating in these experiments and will be fully debriefed afterwards; their anonymity is maintained throughout the research. Netquest compensates the participants through points which can be exchanged for goods on the Korus platform. Participants in both studies will receive 10 points for participating. Additionally, in Study 1, the 10% of participants who best predict the modal responses of other participants receive an extra payment of 15 points. This is following recommendations made by Bicchieri (2017) on the measurement of social norms.

Participants who are not Spanish citizens and/or are below 18 years of age will be excluded from the study. Additionally, for Study 2, only voters of the five biggest Spanish parties will be selected (i.e. PP, PSOE, Vox, Sumar and Podemos), to ensure that the respondents vote for parties pertaining to the left-right dimension of Spanish politics. Lastly, to account for possible AI usage among survey respondents, as raised by Westwood (2025), a hidden reverse shibboleth will be included in an attention check. In an open question about the respondent's favourite colour presented as an attention check, white text against a white

background will instruct the participant to answer with "Mauve". Participants who fail to answer this question, or those who provide the hidden shibboleth answer, will be excluded from the analysis. No upper exclusion threshold has been chosen, given the absence of any cues that the participants would need to remember. For Study 2, no exclusion criteria for straightlining will be applied, given the small number of questions answered by the participants; in terms of response time and AI usage, the same exclusion criteria as in Study 1 will be applied.

Sample size

The survey will be fielded in Spain among a sample of 500 respondents with quotas applied for gender and age. This assumes a 17.6% oversampling rate for participants who fail to complete the survey or meet the post-hoc inclusion criteria. The online experiment will be fielded among a sample of 1200 Spanish citizens with quotas applied for gender and age. This assumes a 6.67% oversampling rate for participants who fail to complete the survey or meet the post-hoc inclusion criteria. This lower oversampling rate can be justified given the shorter length of the experiment.

Sample size rationale

Figure 1 below plots the required N to reach a statistical power of 80%, assuming a baseline of 1.05 on a 1–6 Likert scale and a small to moderate effect size of 30% of the SD, using the DeclareDesign package in R. A relatively high standard deviation of the outcome variable (1.775) is assumed based on calculations using data from Álvarez-Benjumea and Valentim (2024).

Figure 2 below reports the power calculation of the experiment, assuming a baseline sanctioning incidence of 20% and a within-individual correlation of 0.365, based on calculations using data from Álvarez-Benjumea and Valentim (2024). It assumes a small effect size with a Cohen's h of 0.25 between conditions, and ± 0.125 between the control condition and each manipulation condition.

Figure 15

Power curve Study 1.

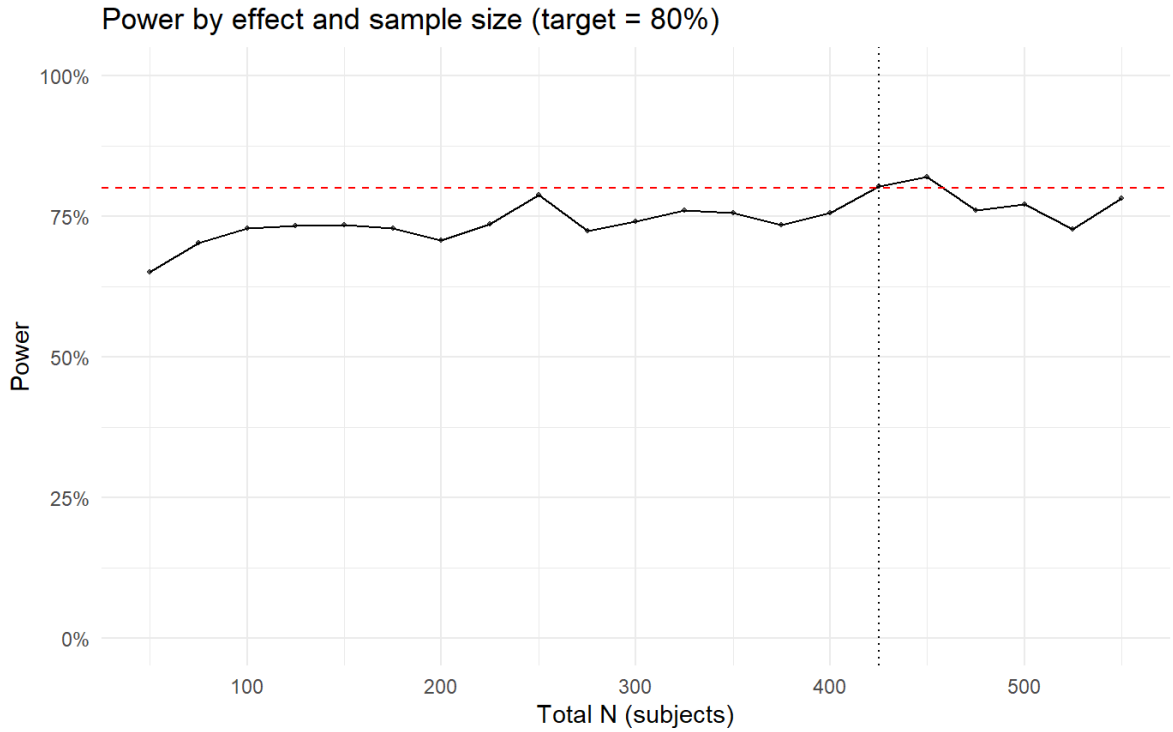
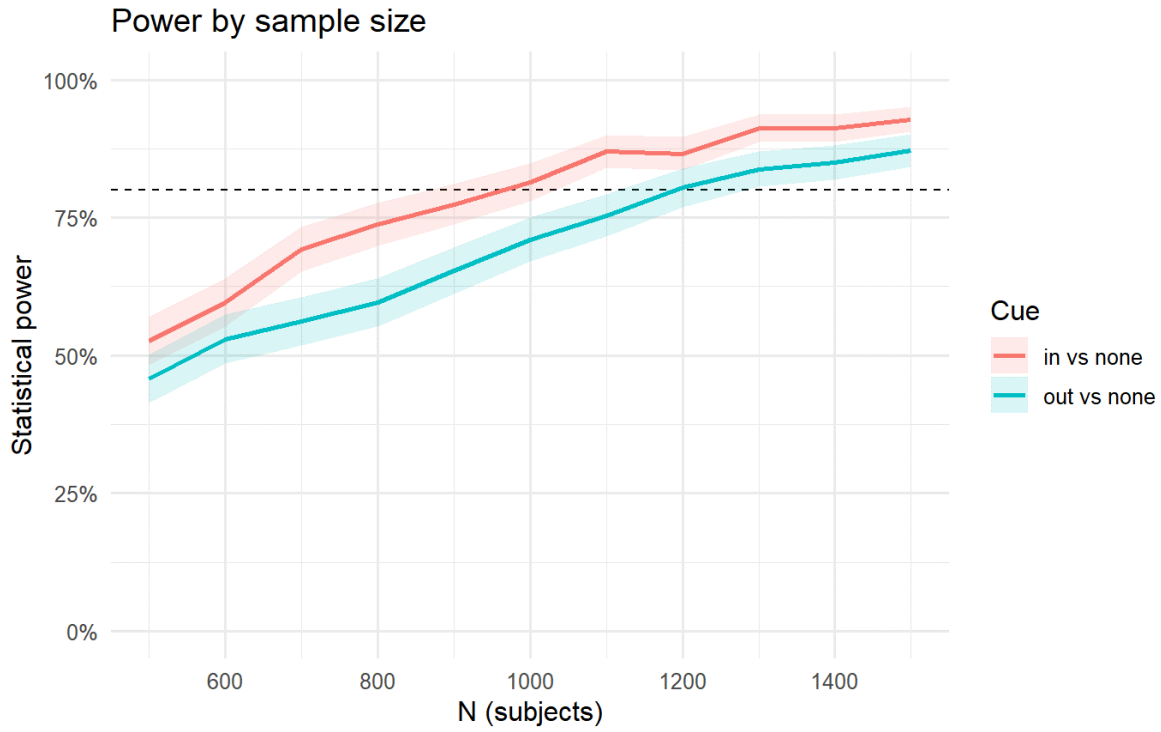


Figure 16

Power curve Study 2.



Manipulated variables

In Study 1, the manipulated variable is whether posts contain antidemocratic or democratic views on the topic of freedom of expression. These posts have been sourced from real-life posts on X, and then modified to guarantee anonymity, as well as to make sure that they sound politically ambiguous (i.e. credibly from the Spanish left or right). The posts have been pre-tested using a small convenience sample as to ensure that they sound natural, credible and politically ambiguous.

In Study 2, the manipulated variable consists of perceived political identity of the account behind the post shown to participants. Above the post, participants will either see #SiemprePodemos (radical-left), #SoloQuedaVox (radical-right), or no cue at all. These slogans are popular party slogans derived from actual headers on X. Each participant sees exactly one antidemocratic post for each condition.

Measured variables

Study 1

Predictor: A categorical variable coded as 1 if the publication contains antidemocratic

views on the topic of freedom of expression (for post, see Appendix). These posts have been sourced from real-life posts on X, and then modified to guarantee anonymity, and to make sure that they appear politically ambiguous (i.e. credibly from the Spanish left or right). The posts have been pre-tested using a small convenience sample to ensure that they seem natural, credible and politically ambiguous.

Outcome variable: a norm violation index consisting of the mean of the following three items:

- Personal normative beliefs: "It's one thing to agree with a statement. It is another to consider whether it is socially appropriate to express it to others. In your opinion, how socially appropriate or acceptable is it to express this statement to others?" (1–6 Likert scale)

- Empirical expectations (incentivised): "To what extent do you think that most survey respondents agree with this statement? Try to mark the response most often chosen by the 500 respondents in this survey." (1–6 Likert scale)

- Normative expectations (incentivised): "How socially appropriate or acceptable do you think most respondents believe this statement to be? Try to mark the response most often chosen by the 500 respondents in this survey." (1–6 Likert scale)

Attitudes will also be measured on a 1–6 Likert scale ("To what extent do you agree with this statement?") to maintain survey consistency, but will not be included in the index of norm violations. No covariates will be included, given the within-individual nature of the analysis. Nonetheless, for post-hoc exploratory checks, questions on gender, age, education, political interest (1–5 Likert scale), ideological identity (ideological self-placement on a 0–10 left/right-scale) and political identity will also be included. In both studies, political identity will be measured by the following question: "Which of the following parties do you sympathise with the most, or consider closest to your own views?"

Study 2

Predictors: Level-1 in- and out-bloc indicators, constructed from a combination of the treatment and the political identity of the respondent, measured in the same way as in Study 1. Following the strategy of Turnbull-Dugarte and Ortega-López (2025), voters of PSOE, Sumar and Podemos are conceptualised as the left-wing bloc, while the right-wing bloc consists of PP and Vox. For voters of the left-wing bloc, the in-bloc indicator has a value of 1 if the post comes with the #SiemprePodemos cue. For voters of the right-wing bloc, the in-bloc indicator has a value of 1 if the post comes with the #SoloQuedaVox cue. Conversely, the out-bloc indicator has a value of 1 if the participant belongs to the left-wing

bloc and sees the #SoloQuedaVox cue, or a value of 1 if the participant belongs to the right-wing bloc and sees the #SiemprePodemos cue. The control indicator has a value of 1 for all respondents in the case of the post without the party cue.

Outcome variable: A dichotomous variable that is coded as 1 if a respondent has marked a post as inappropriate. For more information on the instructions provided to respondents, see "Research design" in the attachments.

No covariates will be included, given the within-individual nature of the analysis. Nonetheless, for post-hoc exploratory checks, questions on gender, age, education, political interest (1–5 Likert scale), ideological identity (ideological self-placement on a 0–10 left/right-scale) and political identity will also be included. In an exploratory post-hoc analysis, it will be assessed whether the effect varies significantly between left-wing and right-wing respondents, through interactions with the left/right-wing bloc identity of respondents.

Statistical models

Study 1

Because individuals evaluate a subset from a pool of statements (crossed design), a crossed random effects model (individuals × statements) will be estimated. In this model, the antidemocratic nature of the statement is included as the independent variable, with random intercepts for participants and statements. Additionally, a random slope for the antidemocratic indicator will be included at the respondent level. In exploratory checks, the relationship studied here will also be interacted with political identity, to assess whether antidemocratic content is perceived to be more of a norm violation among specific political subgroups. The statistical notation of the model specification is provided below, with β_0 as the baseline appropriateness rating, i.e. that of democratic posts; β_1 represents the fixed effect of antidemocratic publication A_j (1 = antidemocratic); the participant random intercept is given by u_{0i}; the random slope for the indicator at the participant level by $u_{1i}A_j$; and finally, the statement random intercept by v_{0j}.

$$Y_{ij} = \beta_0 + \beta_1 A_j + u_{0i} + u_{1i}A_j + v_{0j} + \epsilon_{ij} \tag{3}$$

Study 2

Data from the behavioural experiment will be used to analyse H2 and H3, with sanctioning as a dichotomous outcome variable and cue conditions relative to respondent political identity as predictors. Norm-concordant posts with the same treatments are included in the experiment as descriptive validity checks but are excluded from the primary model. Given that the outcome variable is dichotomous, a mixed-effects logistic regression will be carried

out, with random intercepts for participants and statements. In an exploratory post-hoc analysis, an interaction term between the treatment and respondent political bloc will be added to assess whether the effect varies between left-wing and right-wing respondents. This study will rely on two-tailed hypothesis testing ($\alpha = .05$) throughout, with graphics reporting confidence intervals at 95%. The model specification for the behavioural experiment is given below, with P_{ij} as the probability that individual i sanctions antidemocratic post j. β_0 represents the baseline condition (antidemocratic content with no political cue); β_1 represents the fixed effect for the in-bloc condition; the fixed effect for the out-bloc condition is given by β_2, and the random intercepts for participants and statements by u_{0i} and v_{0j} respectively.

$$P_{ij} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \,\mathrm{InBloc}_{ij} + \beta_2 \,\mathrm{OutBloc}_{ij} + u_{0i} + v_{0j})}} \tag{4}$$

Transformations

Study 1

- The predictor, democratic status, is a dichotomous variable coded as 1 if the post contains antidemocratic content on the topic of freedom of expression, and pro-freedom of speech content as the reference category (0).
- The outcome variable (norm violation index) is computed as the unweighted mean of the three component items (personal normative beliefs, empirical expectations, normative expectations).
- For exploratory analysis, interactions will be added with political identity, recoded from a categorical choice of 5 parties into a binary variable: left-wing bloc (PSOE, Sumar, Podemos) vs. right-wing bloc (PP, Vox).

Study 2

- Predictors: This study has two predictors, namely the in-bloc indicator and the out-bloc indicator. For voters of the left-wing bloc (PSOE, Sumar, Podemos), the in-bloc indicator has a value of 1 if the post comes with the #SiemprePodemos cue. For voters of the right-wing bloc (PP, Vox), the in-bloc indicator has a value of 1 if the post comes with the #SoloQuedaVox cue. Conversely, the out-bloc indicator has a value of 1 if the participant belongs to the left-wing bloc and sees the #SoloQuedaVox cue, or a value of 1 if the participant belongs to the right-wing bloc and sees the #SiemprePodemos cue. The condition with no party cue serves as the reference category (0 for both indicators).

- The outcome variable (sanctioning behaviour) is dummy coded as 1 (sanction button pressed) and 0 (pass button pressed).
- Political identity is recoded into the same binary left/right bloc variable as in Study 1 for interaction analysis.

Inference criteria

Study 1

A moderate Cohen's d of 0.5 is defined as the smallest effect size of interest, alongside statistical significance in a two-tailed significance test ($\alpha = .05$). This relatively large effect size is chosen because in order for a behaviour to be considered a norm violation worthy of sanctioning, it should be strongly norm deviant (Wilhelm et al., 2020).

Study 2

This study will rely on two-tailed hypothesis testing ($\alpha = .05$) throughout, with graphics reporting confidence intervals at 95%.

Data exclusion

Study 1

Given the absence of a clear exclusion threshold for straightlining (Wang & Hau, 2025), Study 1 opts to exclude participants who provide the same answers to sets of questions pertaining to four consecutive publications. This threshold was chosen because participants are exposed to three publications per condition. Identical answers to items pertaining to four consecutive publications thus signify that participants do not only fail to discriminate within, but also between conditions. As to exclude response time outliers, this study opts for the transformation approach recommended by Cousineau and Chartier (2010) and tested by Berger and Kiefer (2021). This method excludes observations with a lower threshold of -2 in the z-transformation of the square root of uniformised response times.

Missing data

Study 1: Subjects who fail to complete the batteries related to all 6 posts will be excluded from the analysis.

Study 2: Subjects who fail to complete the rating task related to all 6 posts will be excluded from the analysis.

Exploratory analysis

Study 1: As described in the previous sections, an exploratory check will be carried out to assess whether antidemocratic posts are less of a norm violation among left-wing/right-wing voters.

Study 2: As described in the previous sections, an exploratory check will be carried out to assess whether shared/opposing political identity is tied more strongly to sanctioning behaviour among left-wing/right-wing voters.

Appendix E

Below, we report an anonymised version of the preregistration of the exploratory follow-up study, uploaded to OSF in July 2026, prior to fielding this study. Here, we note deviations from the preregistration, in descending order of importance:

- The lean index, corresponding to E2, was dropped as a measure to be reported. This is for two reasons. First, we did not take into account that this index automatically leans leftwards, given that out of the five parties considered in this experiment, three are left-wing. Second, the index buries the relevant observation that specifically Vox attribution is higher in the antidemocratic condition. This is because PP attribution is lower.
- Although E1 is supported by the data (dispersion was higher in the democratic condition), the difference between the two conditions is very small (Cohen's d = -0.13), and below the preregistered threshold for the power calculation. As such, we do not report this difference in the main experiment.
- The final sample consisted of 510 participants, instead of 500 participants.

Background and purpose

This exploratory check serves as an accompaniment to our main flagging experiment. Specifically, with this study we aim to shed light on how our stimuli are understood in the no-cue condition. After all, interpreting the in-bloc and out-bloc cue effect in relation to a neutral baseline necessitates knowledge on how this baseline relates to existing interpretations of partisanship. Additionally, our main design does not account for the possibility that attributions of partisanship in the baseline condition might differ between democratic and antidemocratic expressions. As such, in what follows, we will outline the design of an exploratory survey investigating whether partisan attributions of (anti)democratic expressions cluster around particular parties. This design will also investigate the possibility that the design of our experiment affected this partisan attribution. Note that expressions constitute the unit of analysis for this study; it is not sufficiently powered to detect differences in how citizens interpret antidemocratic content through a partisan lens.

Experimental design

The setup of this study is similar to that of the experiment: in both treatment arms, individuals are asked to read the same six posts (three democratic, three antidemocratic),

presented in random order. However, instead of flagging posts they consider inappropriate, participants are shown five sliders corresponding to the largest Spanish parties (PSOE, PP, Vox, Sumar, Podemos) in random order. They are then asked to rate the probability that the author of this post is a partisan of each party, ranging from 0 (not at all likely) to 100 (very likely). Respondents are forced to respond for each party by either touching the slider or selecting a "not applicable" option. In the (randomly assigned) first treatment arm, participants are shown all posts in the no-cue condition. In the second treatment arm, these expressions are accompanied by the same conditions (Vox cue / Podemos cue / no-cue) used in the main experiment. All respondents are additionally presented with an attention check: those who fail it will be excluded.

Operationalisation

The main outcome of this study is the attribution dispersion index, operationalised as the normalised Shannon entropy ($H/\log 5$) of the partisan attribution of each post, after these are normalised to sum to one. This variable will be coded such that 0 indicates that all respondents attribute the post to a single party, and 1 indicates that no party is attributed more than any other. This variable will be computed and reported in two ways: (a) the mean attribution distribution across respondents, and (b) the average of each respondent's own attribution entropy. In this way, these operationalisations capture both whether respondents agree on the partisan behind the expression, as well as how certain individual respondents are. The share of respondents answering "Not applicable" will be reported separately, but excluded from the analysis.

A second outcome of this study is the directional lean of each (type of) post. This variable, ranging from -1 to 1, will be computed by subtracting the proportion of the total likelihood attribution of the left bloc (PSOE, Sumar, Podemos) from the proportion of the total likelihood attribution of the right bloc (PP, Vox). To illustrate, 0 means that a post was equally attributed to the Spanish left and right; 1 means that it was fully attributed to the right, and -1 means that it was fully attributed to the left. In addition, we will report the mean probability attributed to each party per post, for optimal clarity.

Expectations

Given our use of a non-probability sample and the small number of posts, this study is necessarily exploratory. Nonetheless, we formulate several expectations. First, we believe that antidemocratic no-cue posts score as less dispersed in our index than their democratic counterparts. Second, we expect that antidemocratic no-cue posts show a rightward lean

in their attribution, while democratic no-cue posts do not. Third, as a manipulation check, we formulate the expectation that individuals are more likely to attribute a post to a Vox partisan if it is accompanied by a Vox cue; ditto in the case of Podemos. Fourth, we expect that exposure to the cued treatment arm is associated with higher post attribution to Podemos and Vox in the case of no-cue posts. Lastly (although we will only offer suggestive evidence on this), we expect that individuals are more likely to attribute antidemocratic posts to their out-bloc.

Analytical strategy and power calculation

The first two expectations will be tested by using respondents sorted into the first treatment arm. The indices described above are computed at the respondent-post level and regressed on content type with random intercepts for respondents in linear mixed models. Posts are not modelled as a random factor, as they describe our specific stimuli. Expectation 3 uses a similar model, except within the cued arm, regressing the cued party's attribution on the presence of said cue (versus no cue posts). Expectation 4 will be tested by regressing party attribution for no cue posts against the treatment arm (between respondents). The latter two expectations are tested using party attribution as the outcome, instead of entropy/lean. Additionally, post-level fixed effects are added for these models. All tests are one-sided.

To test these expectations, a sample of 500 participants will be recruited. This is sufficient to detect effects with a Cohen's d of 0.2 (standardised against the total SD) at 80% power with an alpha of 0.05. Figures 1 and 2 below show the power curves for each expectation, calculated with the DeclareDesign package on R with 600 simulations per curve.

Figure 17

Power curve E1 and E2.

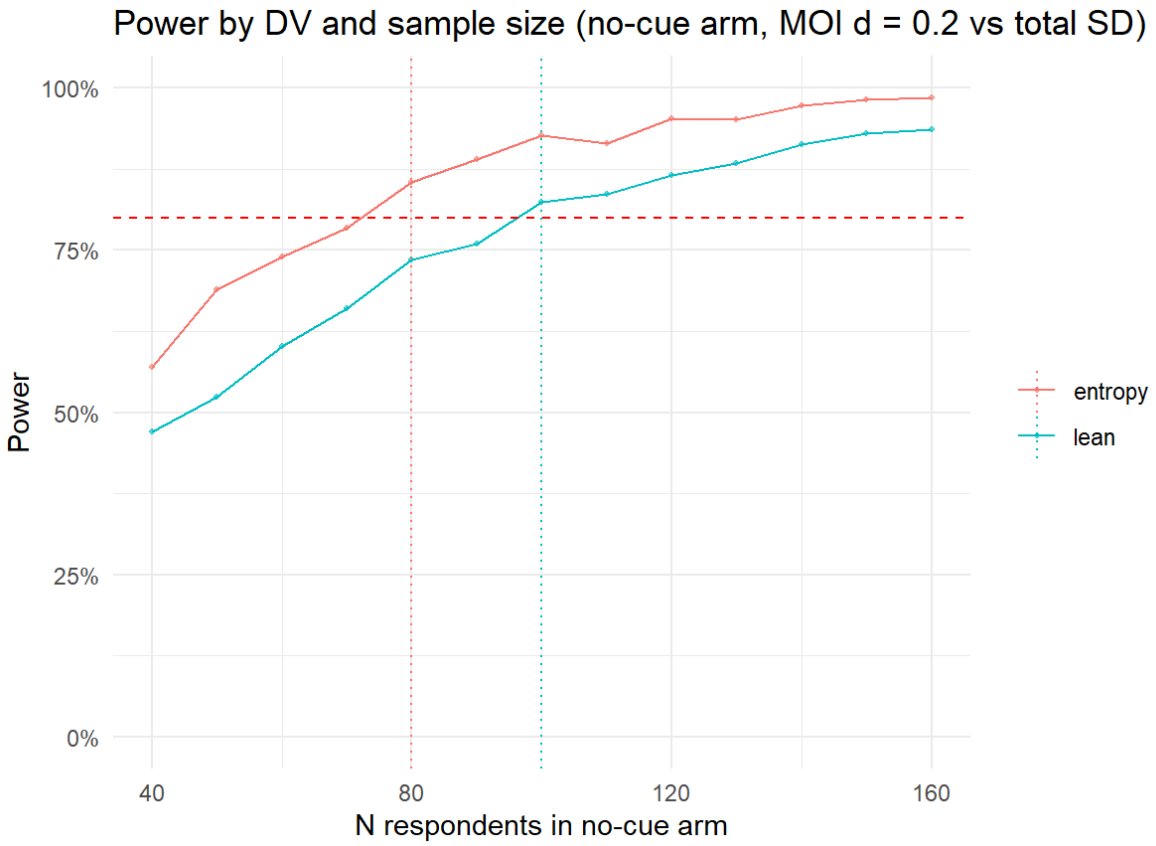
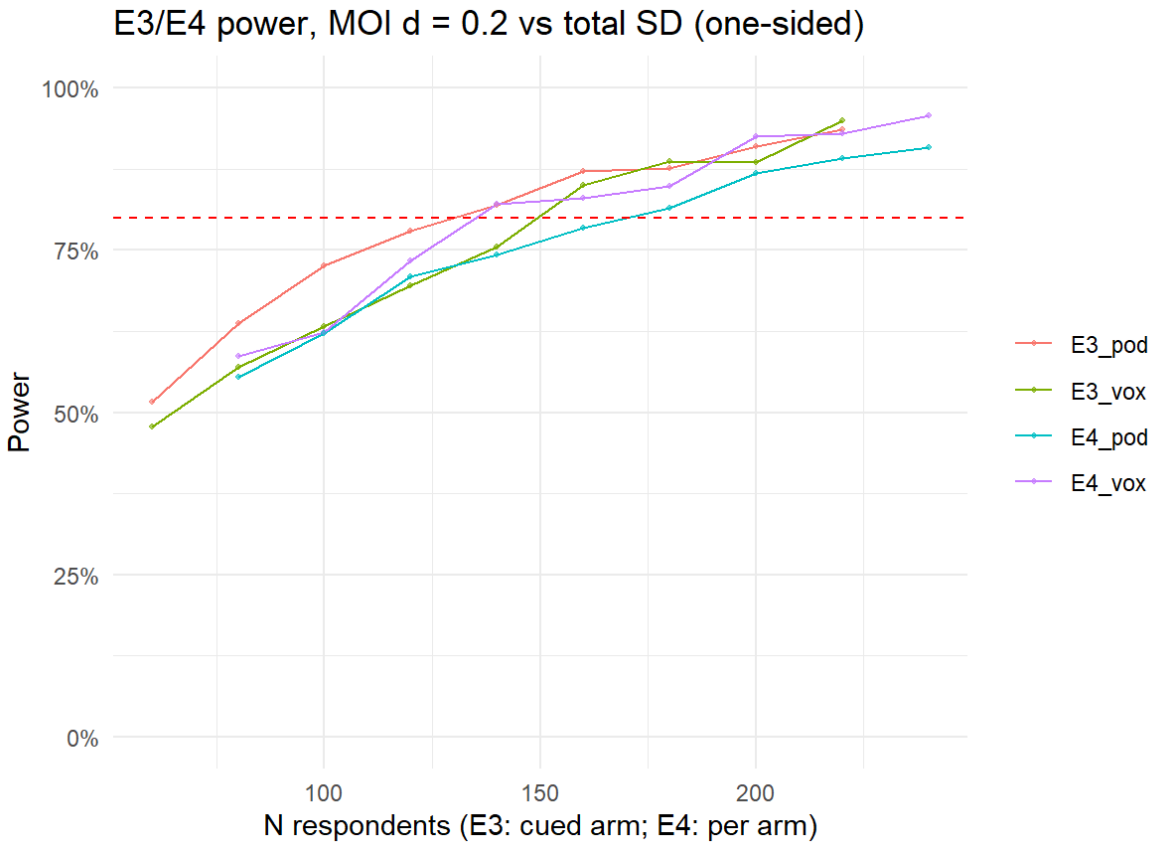


Figure 18

Power curve E3 and E4.



Limitations

Despite being sufficiently powered to detect significant differences between the democratic and antidemocratic posts, our study will be conducted without applied quotas, and using a small number of posts. We are aware that this has inherent consequences for external validity; therefore, the results of this study will be relegated to the exploratory section of our main paper.