Democratic Backsliding and Endogenous Polarization[*]

Ishita Gopal[†] Erica Frantz[‡] Caner Simsek[§] and Joseph Wright[¶]

WORK IN PROGRESS. PLEASE DO NOT CIRCULATE.

November 8, 2023

Abstract

Leading accounts of contemporary democratic decline emphasize the critical role of polarization in enabling incumbent governments to dismantle democracy from within. This study puts forth a more complicated portrait, however. In it, we suggest that polarization is not something that emerges in a vacuum, but instead is itself a product of incumbent attacks on democracy. Incumbent actions that degrade democracy give rise to polarization, in other words, in turn deepening backsliding's progression. We theorize a micro-level link between democratic backsliding events, elite opinion formation, and voter polarization to explain how incumbent efforts to subvert democracy increase negative feelings towards opposing partisans, thereby boosting polarization. Using survey experiments, survey data from electoral democracies, and expert-coded global macro-data, we find support for our argument. The central message to emerge is that political polarization is endogenous to democratic backsliding.

Keywords endogenous polarization; democratic backsliding; political parties

[*] The authors thank Killian Clarke, Scott Desposato, Andrea Kendall-Taylor, Marko Klasnja, Michael Nelson, Eric Plutzer, and Erik Voeten for helpful feedback on this project.

[†] Washington University, iug96@psu.edu

[‡] Michigan State University, frantzer@msu.edu

[§] Münster University, canersimsek@sabanciuniv.edu

[¶] Pennsylvania State University, joseph.g.wright@gmail.com.

In recent years, there has been a steady stream of reports expressing concerns with rising polarization in democracies across the globe, in contexts as diverse as Kenya, Poland, and India, in light of fears that polarization is destructive for democracy (Carothers and O'Donohue, 2019). The New York Times wrote in 2023, for example, that "bitter polarization of American politics is eroding the nation's role as the standard-bearer of freedom, democracy and human rights" (Schmemann, 2023). Existing research has supported such concerns, with a wide range of studies finding evidence that polarization is a central cause of democratic backsliding (McCoy and Somer, 2019; Svolik, 2019, 2020; Haggard and Kaufman, 2021; Chiopris, Nalepa and Vanberg, 2021).

This study presents a more complicated portrait of the relationship between polarization and democratic decline. In it, we argue that polarization does not arise in a vacuum, but rather is itself a product of incumbent attacks on democracy. We put forth that incumbent actions that degrade democracy increase polarization by stoking negative feelings (or affect) among both supporters and opponents of the incumbent towards the other (out) party. This micro-level divide among citizens results in greater macro-level polarization (Druckman et al., 2021), in turn deepening backsliding's progression. From this perspective, polarization is *endogenous* to democratic backsliding.

We define political polarization as the individual-level distance in affective attitude towards a partisan incumbent and a partisan opponent (Iyengar, Sood and Lelkes, 2012). We argue that incumbent attacks on democracy – through executive aggrandizement (Bermeo, 2016), power-grabs that violate principled democracy, and other anti-democratic actions – have a polarizing effect on voters, primarily by increasing negative sentiment towards out-partisans.[1]

It is reasonable to expect those from the opposing party of an incumbent government to increase their negative affect towards the incumbent's party should that government pursue actions that degrade democracy. In the context of the United States (U.S.), for example, this would mean that Republican voters would express greater animosity towards the Democratic party if a Democratic incumbent leader were to seek a power grab. We expect, however, that this dynamic will occur among supporters of the party of the incumbent too. Specifically, we expect that the incumbent's co-partisans will not only continue to support the incumbent's party in the face of anti-democratic actions – even when they highly value democracy – but also that they will increase their negative feelings towards the opposing party. In the U.S., this would mean that Democratic voters would express greater animosity towards the Republican party, even though the Democratic incumbent leader had pursued actions harmful to democracy (and even if they themselves value it). We therefore anticipate that both opponents and supporters of the incumbent will increase their negative affect toward the other in the face of incumbent attacks on democracy.

To explain the behavior of supporters, we emphasize the role of motivated reasoning. We suggest that motivated reasoning incentivizes the incumbent's co-partisans to seek out information and evaluate evidence that supports their existing beliefs and corresponds to their party's position. This means that rather than defecting from the party when the government pursues actions harmful to democracy, they continue to support it (having seen little sign of wrongdoing). Moreover, this disconnect between the anti-democratic actions that transpired and incumbent supporters' understanding of them will increase their negative affect towards the opposition. When opponents sound the alarm bell that the incumbent behaved in ways harmful to democracy, supporters – defiant and in disbelief that the actions indeed are harmful – ratchet up their dislike for the opposition. They respond to the opposition's accusations by engaging in a form of 'whataboutism' (Lucas, 2008), conjuring up negative examples of the other party's behavior. In this way, the opposition's criticism

[1] Polarization can increase because of increases in positive feelings for their preferred party and/or negative feelings for the party they oppose (Iyengar and Krupenkin, 2018). Research shows that in the United States, for example, rising polarization is primarily driven by the latter; positivity toward co-partisans has remained somewhat stable, while negativity toward out-partisans has grown.

of the incumbent's actions actually works to help the incumbent by rallying its supporters.

Our expectaions build on existing research emphasizing that incumbent attacks on democracy shift supporters' understanding of what acceptable behavior in democracy looks like (Grillo and Prato, 2021). And, for this reason, we often see voters who support democracy "unwittingly promote backsliding" (Chiopris, Nalepa and Vanberg, 2021, 29). Even if supporters of the incumbent value democracy in principle, they may not view their party's actions to subvert democracy as troubling.

Importantly, we argue that incumbent attacks on democracy will be most likely to intensify the negative affect of co-partisans towards out-partisans when co-partisan elites offer their endorsement. Consistent with theories of elite public opinion formation (Zaller, 1992; Kam, 2005; Gaines et al., 2007; Arceneaux, 2008), we posit that co-partisan elite support for incumbent aggrandizement provides cues to supporters that such action is justifiable and endorsing anti-democratic actions is part of the incumbent's brand and identity. Co-partisan elite endorsement therefore boosts negative feelings among incumbent supporters towards the opposition even more, producing more intense polarization. This corollary logic provides a scope condition for the macro-implications of our theory: democratic degradation will increase polarization more when ruling party elites endorse (rather than condemn) incumbent-led backsliding. This should be more likely when ruling party elites' careers are closely tied to the brand of the incumbent leader and the survival of the leader in office.

In short, we argue that incumbent behaviors that degrade democracy deepen individual-level polarization by increasing negative feelings among both incumbent opponents and supporters towards the other party. While it is natural to expect opposition voters to view the incumbent more dimly in the face of democratic subversion, we propose that incumbent supporters *also* view the opposing party more negatively at such times. We expect their animosity to be even worse when incumbent party elites endorse (rather than condemn) the incumbent's actions. In this way, we put forth that democratic backsliding is a trigger of political polarization.

To evaluate our argument, we use survey experiments, survey data from electoral democracies, and expert-coded global macro-data. The evidence we present is consistent with our expectations. A central message to emerge is that political polarization is endogenous to democratic backsliding. In this way, this study improves understanding of both the origins of political polarization and the dynamics that underlie democratic backsliding processes.

This study proceeds as follows. In the first section, we offer brief background on the literature on democratic backsliding and the role of polarization, before offering our theoretical argument. Next, we discuss our empirical approach, which involves three tests at three levels of analysis. We close offering a discussion of the study's findings.

Background

Recent treatments of democratic backsliding point to voter polarization as the source of electoral manipulation and incumbent power grabs (McCoy and Somer, 2019; Svolik, 2019, 2020; Haggard and Kaufman, 2021; Chiopris, Nalepa and Vanberg, 2021). In McCoy's (2019) account, for example, polarization weakens cross-cutting political cleavages that keep politics and the partisan coalitions that compete peacefully in democracies from devolving into opposing antagonistic groups unable to compromise and govern. Likewise, in Haggard and Kaufman's (2021) account of backsliding, political polarization is the initial exogenous condition that causes citizen distrust in institutions and government dysfunction, which would-be strongmen then exploit to capture the government and subvert democracy. And in Levitsky and Way's (2018, 155-156) explanation of how democracies die, polarization causes the erosion of social norms that underpin partisan tolerance. These arguments take polarization among citizens as the starting point and theorize how *ex ante* levels of polarization shape the strategic behavior and attitudes of leaders and elites in ways that facilitate democratic erosion.

Similarly, formal models of polarization and democratic backsliding (e.g. Svolik, 2020; Horz, 2021; Chiopris, Nalepa and Vanberg, 2021) begin with the premise that polarization among voters produces incumbent power grabs by altering the incentives of incumbent leaders to not only manipulate the rules of democracy in their favor but also (endogenously) offer increasingly extreme policy platforms. In Svolik's model, for example, some voters value both policy and principled democracy, but face a trade-off between the two. This leads strategic incumbents to leverage (exogenous) voter polarization to win over ideological voters despite their opposition to the incumbent's manipulation of the political process. As polarization increases, incumbents pursue more manipulation. The model in Chiopris, Nalepa and Vanberg (2021) builds on this intuition and introduces voter uncertainty about whether the incumbent is truly authoritarian. Here, polarization alone is not sufficient to produce backsliding; instead there must also be some chance that the challenger is a 'closet authoritarian' who prefers power to policy and thus will exploit polarization to subvert democracy once in office. Voter polarization is therefore an exogenous starting point in these models, which produces democratic backsliding.

We argue, however, that polarization is not something that emerges in a vacuum. Rather, polarization of citizens' preferences on an ideological dimension – often conceptualized on a left-right scale but increasingly drawn along identity lines – results from the actions and strategies of key political actors. Our theory builds on the political psychology literature dedicated to the influence of motivated reasoning on the formation of policy preferences and partisan attachments (e.g., Lord, Ross and Lepper, 1979, Kunda, 1990, Jerit and Barabas, 2012, Parker-Stephen, 2013), as well as the nascent literature on the psychological foundations of partisan support for anti-democratic behavior (e.g. Fishkin and Pozen, 2018, Bartels, 2020, Claassen, 2020, Touchton, Klofstad and Uscinski, 2020). It also draws from behavioral studies suggesting that elite cues – particularly cues that reveal intra-party opinion divergence – shape how voters interpret political reality (e.g. Zaller, 1992, Kam, 2005, Gaines et al., 2007, Arceneaux, 2008, Bisgaard and Slothuus, 2018), including, as our study suggests, incumbent subversion of democracy.

We discuss our theoretical argument in the section that follows.

Theoretical argument

We posit that incumbent-led democratic backsliding polarizes both opponents and supporters by increasing negative affect towards the out-party, thereby resulting in macro-polarization. It is fairly obvious why opponents of the incumbent's party would view it more negatively should the incumbent engage in efforts to subvert democracy. Such actions not only signal trouble on the horizon for the future of democracy – given that backsliding is difficult to reverse once it has begun – but also threaten the future power and influence of their favored party. It is less intuitive, however, why supporters of the incumbent's party would also have a similar response, particularly if they highly value democracy.[2]

We explain this by emphasizing the role of motivated reasoning. Because the incumbent's co-partisans are likely to seek out information and evaluate evidence that supports their existing beliefs and corresponds to their party's position, they are likely to be unconvinced of any wrongdoing following incumbent attacks on democracy (even if they highly value it) and continue to back the

[2] Incumbent supporters who only value power – and not, in principle, democracy – will remain positive towards the incumbent following anti-democratic action because it increases incumbent power, which in turn boosts the utility of voters who only value the power of the party they support. Our model, however, presumes voters value both power and principled democracy.

party. Not only are co-partisans of the incumbent unlikely to withdraw their support in such instances, but they are also more likely to increase negative affect towards out-partisans. The disconnect between how they digested the anti-democratic actions and what actually transpired works to increase their animosity. When opponents sound the alarm bell that the incumbent's actions are harmful, supporters – defiant and in disbelief that the actions indeed are harmful – intensify their dislike for the opposition.

They respond to the opposition's accusations by engaging in a form of 'whataboutism' (Lucas, 2008).[3] 'Whataboutism' entails motivated reasoning that produces concrete (real or imagined) examples of the *other* party's egregious behavior, a form of implicit counter-accusation mixed with standard "differential treatment of similar behavior" by partisans. And by conjuring negative examples of the *other*'s behavior, the motivated reasoning that underpins 'whataboutism' boosts negative attitudes towards the *other* party. Thus by increasing negative feelings towards the political opposition, incumbent supporters tolerate, excuse, and, indeed, justify incumbent anti-democratic behaviors even when it contravenes democratic principles these supporters value. In this way, the opposition's criticism of the incumbent's actions actually works to help the incumbent by rallying its supporters.

The individual-level behavioral link between incumbent-led backsliding and negative affect among partisans (i.e., polarization) that we propose is similar to a "backlash" effect where partisans – via motivated reasoning – strengthen their pre-existing beliefs when presented with evidence that is contrary to those beliefs (e.g. Taber and Lodge, 2006, Guess and Coppock, 2020). Both "backlash" theories and our argument rely on an individual's motivated reasoning. Our theory, however, focuses specifically on the issue of an incumbent leader's anti-democratic actions, which is one mechanism of democratic backsliding. While social (e.g. the death penalty), health (e.g. vaccines), or science (e.g. climate change) policy issues are only indirectly related to partisan identities, we propose a "backlash" logic that directly ties information about the *political behavior of partisan leaders towards the state* to polarization. Further, while some studies posit that voters employ motivated reasoning to accentuate opposition party elites' anti-democratic behavior but downplay their own party elites' poor behavior (e.g. Claassen and Ensley, 2016; Carey et al., 2020), we propose that "downplaying" the anti-democratic behavior of one's own party goes hand in hand with negative affect towards out-parties.

Importantly, we argue that incumbent attacks on democracy will be most likely to intensify the negative affect of co-partisans towards out-partisans when co-partisan elites offer their endorsement. Co-partisan elites, we propose, are critical in shaping the polarizing response of incumbent supporters. Building on theories of elite public opinion formation (Zaller, 1992; Kam, 2005; Gaines et al., 2007; Arceneaux, 2008), we posit that co-partisan elite support for incumbent attacks on democracy provides cues to supporters that such action is justifiable and compatible with democracy. It therefore works to increase negative affect among incumbent supporters towards the opposition even more, producing more intense polarization. Thus, those incumbent supporters who receive information about *both* the incumbent's anti-democratic actions and co-partisan elites' *endorsement* of them should be more likely to have negative affect towards the out-party. In contrast, those incumbent supporters who receive information about incumbent efforts to degrade democracy but who are exposed to elite *condemnation* of the party leader's attacks on democracy should be less likely to increase their negative affect towards the opposition.

We therefore expect that incumbent-led democratic backsliding will be most likely to endoge-

[3] In popularizing the term 'whataboutism', Lucas (2008) writes, "Soviet propagandists during the cold war were trained in a tactic that their western interlocutors nicknamed 'whataboutism'. Any criticism of the Soviet Union (Afghanistan, martial law in Poland, imprisonment of dissidents, censorship) was met with a 'What about...' (apartheid South Africa, jailed trade-unionists, the Contras in Nicaragua, and so forth)."

nously boost polarization when co-partisan elites endorse (rather than condemn) such behavior. We anticipate that this will be more likely to occur when ruling party elites are closely tied to the leader, such that their careers are dependent on staying in the leader's good favor (Samuels and Shugart, 2003; Alesina and Tabellini, 2007). In such environments, elites in the ruling party will be less less likely to disagree and criticize any anti-democratic behaviors on the part of the incumbent and more likely to endorse them.

At the micro-level, the theory therefore suggests that a leader's attacks on democracy will increase negative affect among both supporters and opponents of the leader's party. At the macro-level, the theory suggests that incumbent attacks on democracy should correlate with increased polarization. Further, in ruling parties where elites are more dependent on the leader for their careers and thus have more to lose from criticizing the incumbent, endogenous polarization in response to democratic subversion should be strongest.

How attacks on democracy produce polarization To clarify the theoretical mechanisms linking partisan attacks on the state to affective polarization, we denote the incumbent leader of an in-party (governing party) as D and an out-party (opposition party) leader as O, with partisan voters v_p being attached to either the in-party, v_D, or the out-party, v_O. Partisan voters, v_p, have an affect towards each party. We denote the affect as $A^p_{v_p}$. There are four logical combinations of affect:

1. $A^D_{v_D}$: affect among in-party supporters towards the in-party
2. $A^O_{v_D}$: affect among in-party supporters towards out-parties
3. $A^D_{v_O}$: affect among out-party supporters towards the in-party
4. $A^O_{v_O}$: affect among out-party supporters towards out-parties

Further, the in-party leader, D, can subvert democracy, which we denote as S^D. We propose the following micro-level empirical expectations *when D subverts democracy*:

- $S^D \Rightarrow \downarrow A^D_{v_O}$: out-party supporters increase negative affect towards the in-party
- $S^D \Rightarrow \downarrow A^O_{v_D}$: in-party supporters increase negative affect towards the out-party

The first empirical expectation comes naturally that an opposition supporter will assess incumbent attacks on democracy negatively, increasing dislike of the incumbent government. We call this mechanism *disgust'*: out-party partisan dislike the incumbent more when incumbent attacks the state to undermine democracy. However, this *disgust* effect might be attenuated if opposition supporters already have a highly negative view of the incumbent because the marginal effect of additional negative information might be quite small: pre-existing highly negative views of the incumbent have no farther drop. In a society that is already polarized, the *disgust* mechanism may be small.

The second empirical expectation reflects incumbent supporters' defiant reactions to the alarm raised by the opposition about threats to democracy and their increasing negative affect towards the opposition. Because this often entails reliance on forms of 'whataboutism,' we call this mechanism the *whataboutism* effect.

Within this framework partisan supporters of an incumbent who aggrandizes could increase or decrease their affect towards the incumbent, even while these same supporters of the incumbent increase negative affect towards the opposition. We can use this framework to define *tolerance* and *intolerance*:

Tolerance. If the response of the incumbent's supporters is to double down because they are appalled that the opposition would view the actions of their leader negatively, polarization will increase because incumbent aggrandizement both boosts incumbent supporters' affect towards the incumbent ($S^D \Rightarrow\uparrow A^D_{v_D}$) and decreases their affect towards the opposition via *whataboutism*: ($S^D \Rightarrow\downarrow A^O_{v_D}$). In this scenario, incumbent supporters' *tolerance* of aggrandizement, operationalized as $\uparrow A^D_{v_D}$, contributes to affective polarization.

Intolerance. In contrast, if incumbent supporters recognize aggrandizement as bad for democracy and hence reduce their affect towards the incumbent ($S^D \Rightarrow\downarrow A^D_{v_D}$), we might still observe an increase in polarization. In this scenario, aggrandizement might even *reduce* affect towards the incumbent among incumbent co-partisans, which we might interpret as evidence of *intolerance* for backsliding, while simultaneously, via *whataboutism*, reducing affect towards opposition parties among incumbent co-partisans ($S^D \Rightarrow\downarrow A^O_{v_D}$). Polarization still increases in this scenario when incumbent co-partisans reduce affect for opponents more than they reduce affect towards the incumbent when presented with evidence of incumbent aggrandizement ($S^D \Rightarrow |\downarrow A^O_{v_D}| > |\downarrow A^D_{v_D}|$). That is, the *whataboutism* channel outweighs *intolerance.*

This framework proposes the logical pathways through which partisan leader attacks on democracy could shape partisan affect and hence political polarization. In what follows, we focus on what we believe is the novel theoretical contribution of this framework: the *whataboutism* effect. We propose that incumbent-led attacks on democracy polarize society by boosting in-party members' negative affect towards the out-party. Importantly, this mechanism for producing polarization does not rely on citizens de-valuing democracy (Claassen, 2020); nor does it rely on a close kin of the devaluing democracy theory, namely that citizens like their own party leader more when the leader attacks the foundations of democratic rule. Indeed much of the literature that points to populism as a source of democratic backsliding implicitly presumes that citizens reward politicians for attacking the "corrupt elite" (Mudde and Rovira Kaltwasser, 2018); and if these elites comprise the democratic system of government, then populism may simply entail voters rewarding their partisan leader for attacking democracy. Instead of focusing on whether voters devalue democracy or reward a populist leader when they attack democracy, we propose that partisan voters rationalize their support for a backsliding leader by hating the other party more.

Co-partisan elite responses We argue that the incumbent attacks on democracy are more likely to be polarizing when co-partisan elites sanction them. To understand how different kinds of parties shape this dynamic, we propose that party elites can have one of three responses to their party leader's attack on the democracy: endorse the attack, condemn the attack, or remain silent. In the U.S. Republican party, for example, Senator Mitt Romney condemned President Trump's claim of election fraud after the 2020 election. Meanwhile, other Republican elites, such as Congressman Paul Gosar, endorsed this claim. Meanwhile, many elected Republican elites refused to comment, a response we denote as acquiescing to the attack on democracy by remaining silent.

As such, there exist partisan elites, e_O and e_D, who may either endorse or condemn incumbent attacks on democracy or remain silent. We posit that opposition elites always publicly condemn an incumbent leader's attacks on democracy and that this condemnation increases v_O's negative affect towards D, assuming existing polarization is not already high. In contrast, incumbent party elites, e_O, choose to publicly *endorse* or *condemn* incumbent attacks on democracy or stay *silent.*

We put forth that co-partisan elite endorsement when D subverts democracy increases supporters' negative affective towards the opposition – via the *whataboutism* channel:

- S^D increases $A^O_{v_D}$ more when e_D *endorses* attacks on democracy than when e_D *condemns* them.

Finally, elite silence in the face of their leader's attacks on democracy may, logically, either indicate implicit endorsement or implicit condemnation. However, because most voters are unlikely to pay close attention to the *absence* of an elite cue, we posit that silence implies endorsement. Thus, co-partisan elite silence when D subverts democracy increases supporters' negative affective towards the opposition – via the *whataboutism* channel:

- S^D increases $A^O_{v_D}$ more when e_D remains *silent* about attacks on democracy than when e_D *condemns* them.

We suggest co-partisan elites will be more likely to endorse a leader's anti-democracy actions (either overtly or implicitly) when their careers are dependent on the leader. Building on existing research on party personalism (Kostadinova and Levitt, 2014; Frantz et al., 2021), we put forth that this environment is more likely when the ruling party is *personalist*. In personalist parties, elites have political careers closely tied to the political fate of the party leader because the latter tends to control party funding and nominations (Frantz, Kendall-Taylor and Wright, 2024).[4] Therefore, in personalist parties elites will be more likely to endorse the leader's attacks on democracy or stay silent; they will be less likely to condemn them. There will be more elites like Paul Gosar in personalist parties and fewer elites like Mitt Romney. In non-personalist parties, by contrast, party elites will be more likely to condemn incumbent attacks on democracy because their careers are not closely tied to the political fate of the incumbent leader (Frantz, Kendall-Taylor and Wright, 2024). They will be less likely to endorse them or stay silent.

We can apply these expectations about party elites' behavior to the micro-level behavioral implications discussed above by operationalizing the concept of a personalist party as co-partisan elite endorsement of (or, as we explain below, silence about) the party leader's attacks on democracy. Similarly, a non-personalist party can be operationalized as co-partisan elite condemnation.

Empirical tests

To examine implications of the theory, we conducts three tests at three levels of analysis. The first test is an online experiment conducted in the U.S. in which partisan respondents are treated with a written vignette describing their party leader's verbal attack on democracy (control) as well as a co-partisan elite condemnation (endorsement, silence) of the leader's attack. We then measure partisan affect towards their own party and the other party. Thus Republican respondents are treated with former President Donald Trump's attack on democracy, while Democratic respondents are treated with President Joe Biden's attack on democracy. This test examines the main behavioral implication of our theory. While we might expect partisans to tolerate their own leaders' attack on democracy or even increase their affect towards their party leader, we focus on the more novel implication of the theoretical framework: *whataboutism.* We expect partisans treated with their own party leader's attack on democracy to have more negative affect towards the *other* party than partisans provided with the control. In this test we operationalize the concept of *personalist parties* as elite endorsement, silence, or condemnation, with the assumption that personalist parties have

[4] This logic helps us understand why ruling parties where elites' careers are closely tied to the political fate of the incumbent leader (as is often the case in personalist parties), are the source of endogenous polarization and not necessarily populist parties that arise with the decline of traditional parties (e.g. Berman and Snegovaya, 2019; Benedetto, Hix and Mastrorocco, 2020) or parties where political outsiders take control (e.g. Barr, 2009; Carreras, 2012; Buisseret and Van Weelden, 2020).

elites who are more likely to endorse or remain silent about their party leaders' attack on democracy than condemn it. We expect party leader attacks on democracy to decrease affect towards the out-party when in-party elites are either silent in the face of the attack or endorse it.

The second test uses aggregated survey data on out-party and in-party affect from the Comparative Study of Election Systems (CSES). While the standard macro measurement of affective polarization is simply the aggregate of out-party and in-party affect (Gidron, Adams and Horne, 2020), using CSES data allows us to test whether attacks on democracy influence out-party and in-party affect separately (Reiljan et al., 2023). This is essential because our theory posits that the *whataboutism* channel produces increases in out-party negative affect. Thus an aggregate measure of polarization that encompasses both in-party and out-party affect (i.e., one that does not distinguish between the two) cannot isolate the *whataboutism* mechanism. We expect that incumbent leader attacks on democracy will decrease out-party affect and that this decrease will be larger when the ruling incumbent party is more personalist. This expectation stems from the assumption that personalist parties have elites who are more likely to endorse or remain silent about their party leaders' attack on democracy than condemn it.

The final test is a global analysis of affective polarization for all democracies from 1991 to 2020. Again, we expect that incumbent leader attacks on democracy will boost polarization and that this increase will be larger when the ruling incumbent party is more personalist. We use the Varieties of Democracy data on affective polarization, which is based on expert judgments, and an objective measure of ruling party personalism from Frantz, Kendall-Taylor and Wright (2024). In this test, we use a dynamic panel model to leverage changing levels of macro polarization in response to incumbent attacks on democracy.

Each empirical test has strengths and drawbacks. The survey experiment: (a) leverages random treatment assignment; (b) utilizes a narrow treatment that identifies both the party leader who attacks democracy and the elite party members who endorse the attacks; and (c) precisely measures individual-level partisan affect towards the in-party and the out-party. The experiment was conducted online in the United States in 2023, however, and thus lacks generalizability. The CSES survey again enables us to test how incumbent government attacks on democracy shape affect towards both the in-party and out-party, but the outcome is an aggregate country-election year average measure of partisan affect. The data span 39 countries during the period from 1996 to 2019 but the analysis relies on an imprecise measure of incumbent government attacks on democracy based on expert opinions, as we explain below. Finally, while the analysis of macro-polarization data from the V-Dem project provides the best temporal and geographic coverage (democracies in 100 countries from 1991 to 2020), it does not allow us to test how attacks on democracy separately influence attitudes towards the in-party and the out-party; and again, this analysis uses an imprecise macro measure of attacks on democracy. That said, taken together the results of the three tests provide substantial evidence consistent with the proposition that attacks on democracy increase polarization and do so by increasing negative affect towards the out-party.

We note that our approach differs from Albertus and Grossman (2021), who ask questions about incumbent power grabs in *other* countries to assess citizens' tolerance for executive aggrandizement in their own country, because we examine how incumbent attacks on democratic institutions *in the respondent's country* shape the components of affective polarization. Further, our framework allows us to unpack the separate effects of *tolerance* – when a co-partisan of the incumbent does not alter their affect towards the incumbent in the face of attacks democracy – and *polarization* – when there is a change in the relative affect for incumbents and opposition parties – by estimating changes in affect towards the incumbent and opposition among both incumbent and opposition co-partisans.

Conceptualizing 'attacks on democracy' as attacks on an independent judiciary Attacks on democracy can take many forms: jailing opposition candidates, subduing independent media, restricting voting rights, gerrymandering electoral districts, violating executive term limits, and directly curbing the power of institutional constraints, such as the legislature and judiciary. As importantly, many incumbent-led attempts to undermine democracy are framed by the incumbent as necessary to improve or defend democracy (Levitsky and Ziblatt, 2018).

To operationalize attacks on the democracy, we focus on the judiciary. We do so for three reasons. First, the judiciary is a state institutional body that has the potential to constrain executive behavior, particularly when the leader attempts to undermine democracy (Larkins, 1996; Gibler and Randazzo, 2011; Reenock, Staton and Radean, 2013; Blauberger and Kelemen, 2017). Strong and politically autonomous judicial institutions are a quintessential element of checks and balances on executive power (e.g. North and Weingast, 1989). Tasked with the responsibility of interpreting constitutions and laws, courts can issue rulings that place limits on executive actions, including behaviors that undermine democracy. Because court appointments do not necessarily change when the partisan composition of government changes, judicial constraint serves as an intertemporal check on the executive even when partisan-controlled legislative institutions do not.

Second, for incumbent attacks on democracy to polarize voters, citizens must observe these attacks. That is, individuals must have information that the leader has, in fact, done something. Incumbent attacks on the judiciary tend to be newsworthy and thus highly visible to the public. For example, leaders' attempts to undermine judicial independence in Pakistan (2007), Poland (2017-2019), and Israel (2023) have not only been highly visible but have generated mass protests, magnifying the issue in public discourse. And, far from concealing his attack on the court, Salvadoran President Najib Bukele broadcast to his Twitter audience his purge of five incumbent Supreme Court justices in 2021. In short, incumbent attacks on the judiciary are highly visible, not hidden from the public.

Finally, the behavior of the leader towards the courts can typically be attributed to the leader – and not to some other actor – making the judiciary a good venue to measure leaders' attacks on democracy. While legislatures are often the institution that passes laws to curb judicial power, these direct attacks on the courts rarely happen without executives – or heads of government – leading the way. And even when the legislature legally curbs the independent power of the judiciary to benefit the executive, voters interpret the action as the responsibility of the ruling party and its leader. That is, executives are clearly responsible, in the minds of voters, for attacks on the judiciary.

Incumbent attacks on the judiciary and its independence can take many forms, including verbal assaults and violent threats on the institution or its justices (e.g. in the U.S.); attempts to legally circumscribe the jurisdiction of the court (e.g., Israel); packing or purging the court (e.g., El Salvador, Philippines, Poland); and even abolishing the court altogether. Researchers operationalize attempts to 'curb the court' in many of these same ways: "do away with" the court; "reduce the issues" under court jurisdiction; make the court "less independent" of the executive; and "remove justices" from or "expand the size" of the court (Bartels and Johnston, 2020; Driscoll and Nelson, 2022). Others measure verbal attacks on the court, such as a leader disparaging the court with comments such as, "justices are really nothing more than politicians in robes" (Nelson and Gibson, 2019).

A substantial literature suggests that voters are unlikely to approve of executive attacks on the judiciary because the courts tend to enjoy diffuse support from a broad cross-section of the public (Caldeira and Gibson, 1992; Vanberg, 2001, 2004; Staton, 2006, 2010). But more recent experimental evidence from the U.S. suggests that partisan voters may tolerate – or even approve of – incumbent attacks on the judiciary when they have confidence in the source of those attacks (Armaly, 2018;

Nelson and Gibson, 2019; Driscoll and Nelson, 2022). Meanwhile, Bartels and Johnston (2020) show that polarized partisan attempts to 'curb the courts' affect public perceptions of court legitimacy; and Armaly and Enders (2022) demonstrates that polarization influences support for the U.S. courts. These studies, however, look at court legitimacy, operationalized as support for the institution of the judiciary among the mass public. That is, this literature has not thus far examined how curbing the court shapes voter polarization, as our study does.[5]

By operationalizing the concept of leader 'attacks on democracy' as an attack on the judiciary we can use this concept at multiple levels of analysis, as outlined above. In the survey experiment, we employ a precise treatment that operationalizes attacks on the judiciary as a party leader's verbal attack on the judiciary and a stated intent to "fire judges or expand the number of [court] justices" to improve the chances the court issues rulings favorable to the leader. In the analysis of the CSES survey data on in-party and out-party affect and the global data on macro-polarization, we measure government attacks on the judiciary by aggregating information from three expert-coded variables from the V-Dem project that attempt to capture: whether the government verbally attacked the judiciary; whether the government purged the high court; and whether the government packed the high court. While these three concepts match the operationalization of attacks on the judiciary we use in the survey experiment treatment, the actor in the observational variables is simply the "government", which we interpret as the ruling party and its leader even though the leader and ruling party's name is not actually used in the variable question. Nonetheless, both ways we operationalize 'attacks on the judiciary' capture the same three concepts: verbally attacking the court, packing the court and purging the court.

Study 1: Testing the micro-behavioral implications

We conduct a nationally-representative survey experiment in the U.S. in June 2023. This case allows us to test how party leader attacks on democracy influence in-party and out-party affect among partisans who support the ruling party (Democrat) and the opposition party (Republican). Importantly, while it easy to identify the leader of the ruling party, partisans of opposition parties may not know the de facto leader of their party when it stands in the opposition. In the U.S., for example, there is typically not a clear party leader once a presidential candidate loses an election – at least until the primary campaign during the next election cycle selects a new party leader. In the U.S. case in 2023, however, the partisan name recognition of the opposition party leader is very high; all partisans know who Donald Trump is and view him as the de facto party leader. Secondly, the U.S. case provides variation in the level of party personalism for the ruling (low personalism) and opposition (high personalism) parties.[6]

Our theory suggests that ruling party leader (Biden) attacks will boost negative affect toward the opposition party (Republican) among incumbent party partisans (Democrat identifiers). Similarly, opposition party leader (Trump) attacks on democracy should increase negative affect toward the ruling party (Democrat party) among opposition partisans (Republican identifiers). While party leader (Biden) attacks on democracy might produce negative affect towards the attacker's party (Democrat) because the institutional target of the attack (Supreme Court) enjoys broad support,

[5] One study tests whether partisan contests over judicial nominations polarize citizens' perceptions of the court nominees and the court itself (Rogowski and Stone, 2021). This test, however, does not examine how attempts to curb the court shape affective partisan polarization.

[6] During the Trump presidency, the Republican party had moderate to high party personalism scores of 0.7 (0-1 scale) and -0.3 (-2.9 to 3.8 scale), according to Frantz et al. (2022) and the Varieties of Party Identity and Organization data sets. In contrast, the Democratic Party would have a very low score of 0.0 (0-1 scale) for the Democratic party during Biden's presidency and had a score of -2.3 (-2.9 to 3.8 scale) in 2018 using the Varieties of Party Identity and Organization coding.

or legitimacy, among the public, including incumbent co-partisans (Democrat identifiers), it is also possible that (Democrat) supporters of the ruling party may view such attacks positively, particularly if a leader they trust justifies the attacks in partisan terms (Armaly, 2018; Nelson and Gibson, 2019; Driscoll and Nelson, 2022).

Design Because our theory has different empirical expectations for assessments of different parties among pre-existing groups of partisans, we treat partisan respondents with a statement by their own party leader attacking the U.S Supreme Court. We use a control scenario that entails a statement by the party leader that primes respondents to think about the institution of the court without the leader attacking the court. Thus similar to the treatment arms, the control scenario introduces the *saliency* of the court, allowing us to isolate the effect of an attack on the institution in the treatment from the saliency of the institution, in the treatment and control. Second, to test the conditioning effect of co-partisan elite endorsement, respondents are treated, using a cross-subject conjoint design (Bansak et al., 2021), with: (a) the treatment only; (b) a treatment plus co-partisan elite *endorsement* of attacks on democracy; or (c) the treatment plus co-partisan elite *condemnation*. This yields the conditions shown in Table 1 that are randomly assigned within self-identified Democrats and Republicans:

Table 1: Survey treatment arms

		Co-partisan party elite		
		Endorsement	Condemnation	Ignore
Party leader	Judicial aggrandizement statement	Treatment + Endorse	Treatment + Condemn	Treatment
	Judicial vacation statement	Empty	Empty	Control

Treatment We want treatment conditions and an associated control that: (a) closely match the concept of incumbent attacks on the judiciary; (b) resonate across different countries with distinct legal systems and legacies of executive-court relations; and (c) isolate the treatment – incumbent attack – from the saliency of the attack's target, in this case the judiciary. By testing the treatment relative to a "control" that also primes respondents to think about a policy change regarding the judiciary, the design holds constant the issue area across treatment and control to isolate the treatment effect of an incumbent attack on democracy. We operationalize a treatment in which the incumbent executive verbally attacks the court in response to an adverse decision; the incumbent leader's attack on the court takes the form of suggesting that the executive should be able to purge or pack the court. The control condition also includes a statement by the party leader about a change to the court. However, the control statement does not contain a leader attack on the independence of the court and instead concerns a proposal to expand judicial vacation time.

Incumbent attacks on the state rarely occur outside of a partisan political context. We therefore frame the attack as a response to a court decision that adversely affects the executive or their party's position on an issue. The treatment does not reference a specific court decision or a specific issue area (e.g. health care policy, human rights, electoral conduct, or corruption) about which the court has ruled because providing this information might contaminate the treatment by prompting respondents to think about their position on a specific policy issue.

Further, incumbent attacks rarely occur without the attacker providing some rationale or justification for the attack. In the context of democratic backsliding, in fact, incumbents often justify their attempts to undermine executive constraints by framing their attack as a defense of democracy. We operationalize this point with an elite endorsement or condemnation that explicitly raises the issue of democracy. An elite endorsement thus justifies the leader's attack on the judiciary as good for democracy, while elite condemnation does the opposite. In both cases, democracy is the term used to justify co-partisan elites' position, holding constant the rationale in the justification across the endorsement and condemnation conditions. We employ the following vignettes:

- *Judicial aggrandizement treatment*: "Incumbent [executive title] [executive name] responded to an adverse [court name] ruling by suggesting that the [executive office] should be able to fire judges or expand the number of justices to get more judges on the court who agree with [him/her]."

- *Judicial vacation control*: "Incumbent [executive title] [executive name] responded to the justice commission's finding that justices are overworked by suggesting that the [court name]'s justices should be eligible for expanded paid vacation."

- *Co-partisan elite endorsement*: The senior legislative leader of [executive party name] Party responded to [executive name]'s proposal to change the court composition to make it more friendly to [leader's name] by insisting that these changes be made immediately. "This is the only sensible path forward for our democracy. The [executive title] is the elected leader; the [court name] should not be making policy. The [executive title] should." [name of partisan legislative leader] said.

- *Co-partisan elite condemnation*: The senior legislative leader of [executive party name] Party strongly condemned [executive name]'s proposal to change the court composition to make it more friendly to [leader's name]. "Allowing the [executive title] to arbitrarily change the composition of the [court name] will pose a threat to our democracy, if not now then in the future" [name of partisan legislative leader] said.

Outcome Post-treatment, respondents answer two questions that measure negative partisan identity, using a five-point Likert scale (Bankert, 2021):

- "When people criticize this party, it makes me feel good"'

- "When I meet someone who supports this party, I feel disconnected"

This outcome, which captures negative partisan identity, should measure a feature of individual respondents' "negational categorization", a process by which individuals form an identity in opposition to another group, often an of out-group (Zhong et al., 2008; Lee et al., 2022). We combine the two ordinal variables separately for Republicans and Democrats using polychoric PCA.[7]

Second, respondents mark feeling thermometers, using a `0` to `100` scale, about the opposition leader (name) and the ruling party leader (name). The feeling thermometer outcome builds on

[7] See Appendix A. Reproduction files show the main result holds irrespective of the aggregation method (polychoric PCA, linear combination, or graded-response IRT).

research that demonstrates survey respondents (in the U.S.) identify the generic party label with party elites, especially the party leader (Lelkes and Westwood, 2017; Druckman and Levendusky, 2019) and not with partisan voters. This measure specifically identifies the (ruling and opposition) party leaders to isolate opinion about elites from opinions about mass supporters of different parties. Further, feeling thermometer measures are highly correlated with party trait ratings and trust measures and match cross-national surveys that contain party feeling thermometers questions (e.g. Comparative Study of Election Campaigns, Boxell, Gentzkow and Shapiro, 2020).

Additional items The survey asks pre-treatment questions to measure self-identified partisanship, which we use as a moderator (Sheagley and Clifford, 2023).[8] This allows the survey to randomize treatment (control) conditions within partisan groups. Post-treatment, we ask respondents demographic information about sex, age, education level, rural (urban), and ethnic group/race.

We also ask pre-treatment questions about diffuse support for democracy and respondent attitudes towards the target of the attack, the judiciary. We ask four questions about support for democracy, which we then use to construct a latent measure of diffuse support for democracy. These questions correspond to diffuse democracy survey items on standard cross-national surveys (e.g. World Values Survey). Importantly, these questions do *not* force respondents to choose between partisan interests and democratic principles (Graham and Svolik, 2020). However, survey items that measure respondents' support for democracy relative to their partisanship implicitly assume these constructs are exogenous and independent from one another. The point of our study, in contrast, is to examine whether the intensity of partisanship is endogenous to incumbent attacks on democracy. We find two dimensions in these items – one associated with diffuse support from democracy and the other related to support for strongman rule.

Finally, the pre-treatment items include questions about respondents' knowledge of the judiciary.[9] Including these items allows for testing whether the treatment effects vary by respondents' political knowledge of the court and its role as a key institution for democracy (Zaller, 1991, 1992; Driscoll and Nelson, 2022). We would not expect respondents who have no factual knowledge of the courts to interpret a party leader's attack on the court as an attack on a democratic institution.

Analysis We drop respondents who speed, defined as those who spent less than half the median time to complete the survey. In Appendix A we show that speeding is highly correlated with: (a) lack of basic, factual knowledge about the Supreme Court; (b) failing the treatment information check; (c) skipping the feeling thermometer questions; and (c) indicating that they feel strongly positive towards *both* Trump and Biden. We find it unlikely that partisan respondents genuinely have highly positive feeling towards both leaders.[10] This suggests that speeders are not paying attention to the survey questions; thus we drop them from the analysis. Appendix A also shows that the reported results do not change at any points near the threshold for demarcating speeders.

Second, in the reported results we only look at respondents who correctly answer at least one of three factual questions about basic Supreme Court knowledge. Less knowledgeable respondents are unlikely to view attacks on the judiciary – irrespective of the identity of the leader – as problematic if they know little about the court; they should thus be less likely to react in a polarizing way

[8] Self-identified independents who typically vote for same party are counted as partisans.

[9] The questions, shown in Appendix B, include: How many justices are usually on the Supreme Court; Who appoints members of the Supreme Court; and If the President and the Supreme Court differ on whether an action by the government is constitutional, who has the final responsibility for determining if the action is constitutional.

[10] Strongly liking both is coded as rating them each above 80 on a 0-100 scale. Conversely there are a substantial number of respondents who do not like either leader, which is consistent with polling in 2023 (Epstein, Igielnik and Baker, 2023; Loffman, 2023).

to attacks on the courts. Central to our core theoretical claim is the contention that partisans recognize an attack on democracy and, in a vacuum, would not like the attack. But to justify continued support for their party leader, voters boost their negative affect towards the out-party. Restricting analysis to respondents who know something about court independence in the first place helps ensure respondents correctly interpret the treatment as an attack on democracy.

We estimate OLS regressions, adjusting the data with demographic controls (age, male, college education, rural, and white) as well as scales for diffuse support for democracy and support for strongman rule.[11] We randomize the four treatment conditions within each partisan group (Republican id and Democratic id): control, treatment, treatment + endorse, and treatment + condemn.

Results Recall that the whataboutism pathway for boosting affective polarization entails partisan perceptions of other parties, not voters' perceptions of their own party. The outcome we examine here is scaled such that higher values indicate *negative* affect towards the *out* party. For self-identified Republican respondents, the outcome is thus negative affect towards Democrats; and for Democratic respondents, the outcome is negative affect towards Republicans.

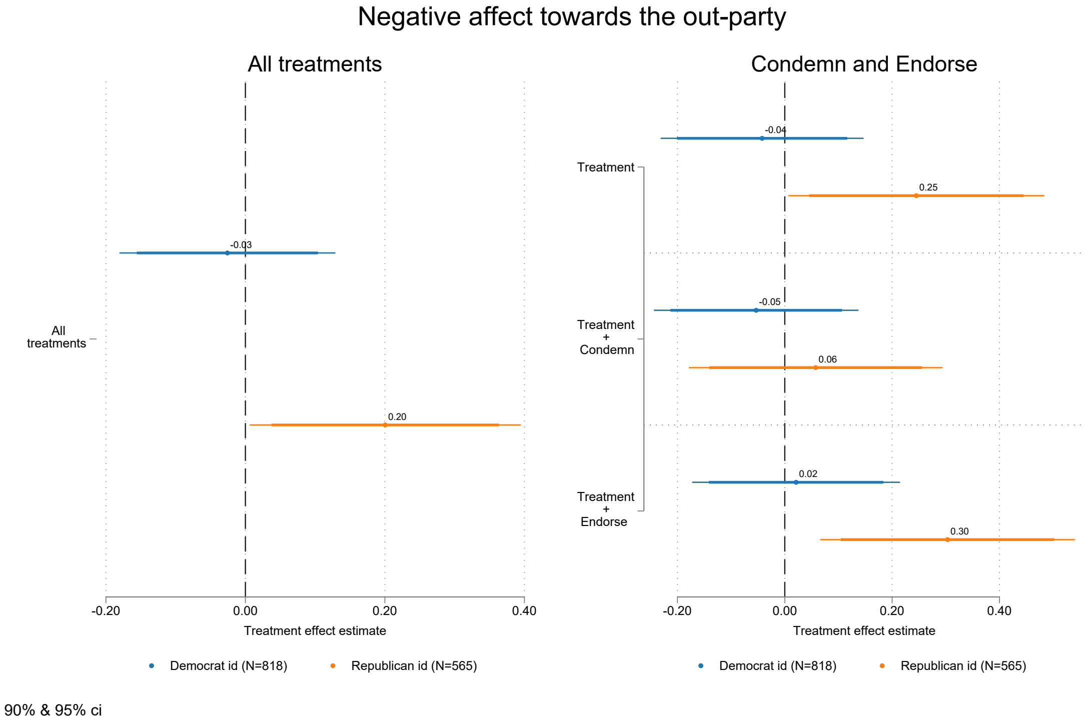


Figure 1: Negative affects towards the *out* party

The first analysis, shown in left panel of Figure 1, groups three treatments together and compares these three arms to the control group.[12] We show the average treatment effect for negative out-party affect for each group of partisans, Democratic identifiers and Republic identifiers. The top estimate, for Democratic self-identifiers, is almost exactly zero, indicating that when treated with a Biden message attacking the judiciary, these respondents do not increase their negative affect

[11] See Appendix A for balance tests and estimates from unadjusted regressions.

[12] Estimates from a linear regression with treatment conditions plus covariate adjustment, with separate regressions for Democratic self-identifiers and for Republican self-identifiers.

towards Republicans. However, among Republican respondents, the estimate for the average of all three treatments is 0.20 standard deviations of negative affect towards Democrats. This estimate is significant at the 0.05 level. Thus only the estimate for Republican identifiers provides evidence consistent with the whataboutism mechanism.

However, combining all three treatment conditions (treatment only; treatment + condemn; and treatment + endorse) obscures a key mechanism that links antidemocratic leader behavior to polarization: elite endorsement of (or silence about) the antidemocratic behavior. We expect treatment to be strongest when elites either endorse the leader's behavior or are silent about it. In the right panel of Figure 1 we therefore report the estimates for each of these treatment arms relative to the control group.

The top set of estimates – for the Treatment only – show that among Democratic respondents there is no treatment effect (-0.04). However, for Republican respondents the treatment effect is 0.25 standard deviations and significant. This indicates that Republican respondents treated with a Trump attack on democracy increase negative affect towards Democrats.

Next, the middle set of estimates – for Treatment + Condemn – are both close to zero. This suggests that when partisans learn about their leader's attack on democracy but an elite member of their party condemns this attack, partisans do *not* increase negative affect towards the out-party.

Finally, the bottom two estimates in the right panel of Figure 1 show how respondents react to the treatment plus an elite member of their own party *endorsing* the leader's attack on democracy. For Democratic respondents, the estimate is 0.02 – again a null finding. For Republicans, however, this estimate is 0.30. This result indicates that being treated with a Trump attack on democracy plus elite endorsement increases negative affect towards Democrats by nearly a third of one standard deviation.

How should we interpret the size of these estimates? To provide some context, it is useful to know how respondent demographic characteristics shape negative affect. As we show in Appendix A, Democrats have higher negative affect than Republicans by about 0.19 standard deviations. And men (0.11) and those who have an affinity for strongman rule (0.12) have more negative affect than women and respondents who are averse to strongman rule, respectively. This suggests that a Trump attack on democracy increases negative affect among Republicans by roughly twice as much as the negative affect difference between men and women. However, this effect only manifests when Republican elites are either silent in the face of the attack or endorse the attack. When elites condemn the attack Republican respondents do not increase their negative affect towards Democrats.

Appendix A provides evidence that the main pattern in Figure 1 remains when we do not adjust for covariates or only adjust for covariates that are correlated with treatment conditions. Further, we show that treatments to do not influence either party's respondents' positive affect towards their own party. There is thus no evidence that party leaders' attacks on democracy boost polarization by increasing affect towards one's own party.

These results suggest two conclusions. First, we find no evidence that party leader attacks on democracy boost negative affect among partisans in a non-personalized party (Democratic party). But we find evidence for the whataboutism channel of polarization in the more personalized party (Republican). Second, party leader attacks on democracy only boost negative affect towards the out-party when partisan elites either endorse the attack or remain silent about it. When elites condemn the leader's attack, the antidemocratic behavior does not polarize that party's voters.

We do not interpret this evidence to suggest that partisans in a more personalized party are more likely to be persuaded by their leader's attacks on democracy. Indeed, we do not test how these attacks on democracy influence partisans' attitudes towards democracy. Instead, our evidence suggests that partisans in a personalized party may justify their support for their party when their leader attacks democracy by disliking the out-party more. Furthermore, we suggest that personalized parties are less likely than institutionalized ones to have elite members who either endorse their leader's attacks on democracy or remain silent about them. Thus personalized political parties may shape affective polarization through two distinct mechanisms: partisan voters boost negative affect towards the out-party (whataboutism) and elites in personalized parties fail to condemn such attacks.

Decomposing polarization Recall that we also measured respondents' post-treatment attitudes towards the party leaders, Joe Biden and Donald Trump, using standard feeling thermometers. We therefore check whether the main patterns of affective polarization remain with respect to partisans' attitudes towards the party leaders.

Figure 2 reports the estimates for the treatment effect among Republican respondents only.[13] The left panel estimates combine all treatment groups together (treatment only; treatment + endorse; and treatment + condemn) and compares these treatment conditions to the control group.[14] The outcomes are threefold: how much respondents "like" Biden, "like" Trump, and a measure of polarization that is simply the difference between the two. This polarization measure is scaled such as that, for Republican respondents, positive values indicate how much they like Trump less how much they like Biden. While the aggregate treatment both increases how much respondents like Trump (2.93 points on a 100 point scale) and decreases how much they like Biden (3.85), the size of the estimate of negative affect towards Biden is larger (in absolute size) than the positive affect estimate towards Trump. Putting these two outcomes together, we see that treatment increases polarization (1.60). None of these results, however, are close to standard statistical significance.

The right panel of Figure 2 shows results that combine the two treatment conditions we expect to boost polarization: treatment only and treatment plus endorsement. That is, we exclude treatment + elite condemnation. We find a similar, stronger pattern: Trump's attack on democracy increases polarization mostly by increasing negative affect among Republicans towards Biden. Again, the estimates are not statistically significant, with one exception: the negative affect (like Biden) estimate is significant at the 0.10 level.

In short, when we measure polarization using party leader feeling thermometers we find that personalist party leader attacks on democracy boost polarization by increasing negative affect towards the out-party's leader. However, these estimates are quite imprecise, which may reflect the fact that many respondents have very low fixed – and thus immovable – attitudes about out-party leaders; indeed nearly half (47 percent) of the Republican respondents in the sample gave Biden a 0 rating on a 100-point scale.[15]

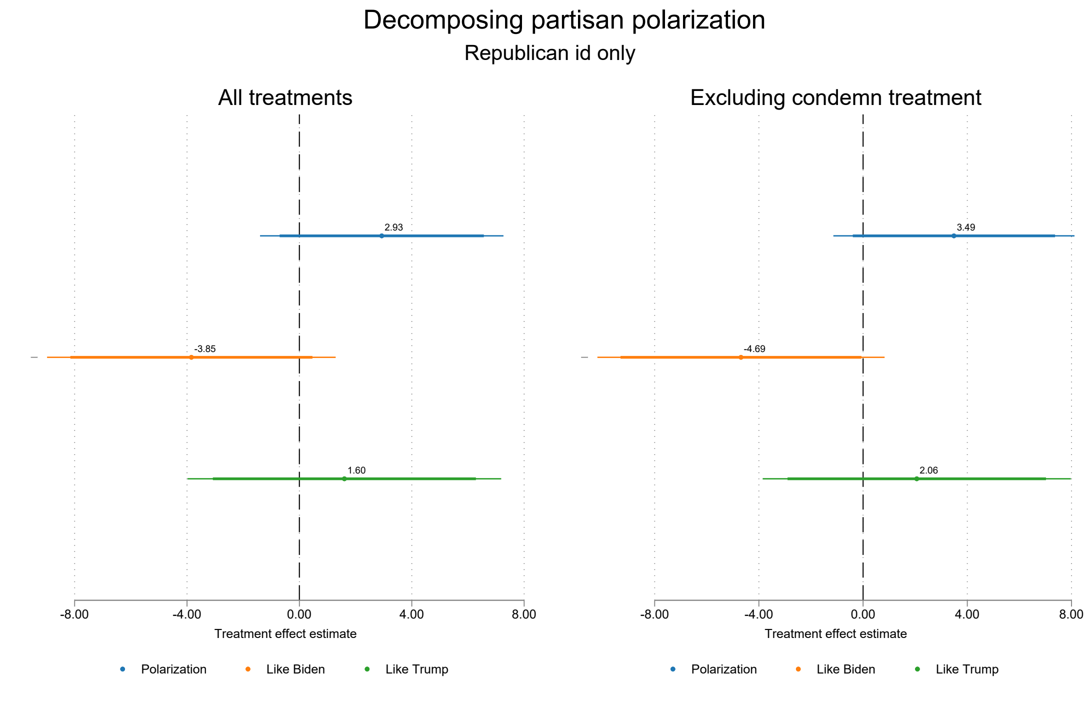
Study 2: Survey data on in-party and out-party affect

This test examines survey data from the CSES, compiled by Reiljan et al. (2023). The data contain measures of in-party and out-party affect for each election in over three dozen democracies. For each election, respondents provide information on their affective evaluations of each major party; this yields measures of in-party and, more importantly, out-party affect averaged across all out-

[13] As shown in Appendix A, we find no evidence that treatments shift Democratic respondents' attitudes towards the two party leaders.

[14] Again we estimate OLS and adjust for five demographic variables as well as indices of support for democracy and strongman rule.

[15] The distribution of affect towards the outparty leader (1.46) is much more skewed than the distribution of negative affect (0.32).



90% & 95% ci; Only respondents who answered at least 1 of 3 court information checks correctly

Figure 2: Decomposing polarization

parties contesting the election.[16] Further, Reiljan et al. (2023) calculate the average affect across respondents for each country-election year among partisans and for all voters.

Our theory posits that incumbent government attacks on democracy should increase polarization by boosting negative out-party affect; and this effect should be stronger when the ruling party is more personalist. This latter modifying condition captures the incentives of partisan elites to either endorse leader attacks on democracy or remain silent about them rather than condemning them, as we demonstrated in the behavioral experiment in Study 1. Finally, we expect evidence for the out-party channel of polarization to be stronger for partisan respondents than for all voters because partisans should be more inclined to accept partisan elite cues than the group of all voters that includes both partisans and independents.

We therefore test whether incumbent attacks on democracy decrease out-party affect and whether the effect is stronger when the ruling party is more personalist. We operationalize incumbent attacks on democracy as an aggregated measure of government verbal attacks on the judiciary as well as government purges of court justices and government court-packing.[17] While we present results from tests for partisan respondents in this section, additional tests show that the polarization patterns for all voters are considerably weaker than those for partisans only, as expected.

[16] The measure is the "average divergence of partisan affective evaluations between in-party and out-parties, weighted by the electoral size of the parties" (Reiljan et al., 2023, 8).

[17] We use the linear combination of three expert-coded variables from the Varieties of Democracy project: government purges of the courts (`jupurge`); government court packing (`jupack`); and government verbal attacks on the judiciary (`jupoatck`).

Analysis We utilize the main empirical specification proposed in Reiljan et al. (2023).[18] However, the government effectiveness variable used in Reiljan et al. (2023) is highly collinear with attacks on the judiciary, in part because it overlaps conceptually with attempts to undermine the judiciary. We replace this variable with measures of public corruption and GDP per capita (Reiljan et al., 2023). Predictors that explain affective polarization and its constituent parts, in-party affect and out-party affect, include: party id; left-right polarization; effective number of parties; presidential (parliamentary) system; government effectiveness; corruption; GDP per capita; and a time trend. We add four variables to this specification: government attacks on the judiciary; ruling party personalism; democracy age; and initial level of democracy in year in which each new leader is selected into power. Attacks on the judiciary is the main treatment variable and ruling party personalism is the hypothesized moderator.[19] We include age of democracy and the initial level of democracy in the leader selection year to account for the fact that personalist ruling parties are more common in newer, less consolidated democracies (Frantz et al., 2022; Frantz, Kendall-Taylor and Wright, 2024). This ensures that factors that cause selection into ruling party personalism do not bias the results.[20]

We examine three related outcomes. The first is the aggregate measure of affective polarization among partisan respondents.[21] While attacks on the judiciary may be associated with *more* polarization, our core theoretical expectation is that attacks should increase polarization more when the ruling party is more personalist than when it is less personalist.

The second is in-party affect and the third is out-party affect. The in-party affect measure combines affect from respondents who vote for the ruling party and for those who vote for opposition parties. This means we do not have a precise measure of in-party affect for respondents who voted for the ruling party that controls the government and is thus responsible for the attack democracy, which is the main treatment. If attacks on democracy increase in-party affect, we would only expect this to manifest for partisans of the ruling party that carries out these attacks.

The third outcome is out-party affect, measured as (vote-share) weighted average of out-party affect for partisans of the ruling and opposition parties. Similar to the in-party affect variable, the out-party affect captures both: (a) opposition party supporters' negative affect towards the ruling party, which we would expect to increase when the government attacks democracy; and (b) ruling party supporters' negative affect towards opposition parties, which should also increase when the government attacks democracy. This latter channel is the whataboutism effect theorized earlier and which we highlighted in Study 1. Given the out-party affect measure cannot isolate affect among ruling party partisans, we cannot precisely identify the whataboutism mechanism. That said, theoretically both supporters – via whataboutism – and opponents of the government

[18] Our sample differs slightly from the main sample in Reiljan et al. (2023) because we use elections in democracies with an elected chief executive. The ruling party personalism variable from Frantz et al. (2022) is coded for all democracies (1991-2020) in countries with more than 1 million population and a directly elected leader. This excludes Czech Republic (2010) and Greece (2012) because they have appointed interim leaders, not elected chief executives. Mexico (1997) and Taiwan (1996) are not included because these elections occurred prior to democratic transitions in these countries in 2000. Turkey (2018) is excluded because it is not a democracy after the 2016 failed coup attempt. Montenegro has less than 1 million population and Switzerland has no elected chief executives.

[19] Frantz et al. (2022) measure ruling party personalism using objective information about the chief executive and the party that supports the executive. Importantly, the measure only draws information from prior to the leader being selected into the chief executive position and thus contains only information that is exogenous to leaders' attempts to undermine democracy, including leaders' attacks on the judiciary.

[20] In reproduction files we show that results are robust to additional potential confounders that might cause selection into ruling party personalism: party system institutionalism; ruling party populism; initial level of judicial independence; and initial levels of macro-polarization. We also show the result is robust to changes in the specification by dropping covariates and testing kernel regressions that relax specification assumptions.

[21] Partisanship is calculated using vote choice responses rather than partisan identification (Reiljan et al., 2023, 8).

responsible for attacks on democracy should view out-parties more negatively when the government attacks the judiciary. Thus we except attacks to decrease out-party affect and that this pattern should be strongest when the ruling party is more personalist.

We report estimates from linear regressions with cluster-robust standard errors but confirm the patterns using nonparametric and kernel estimators. For each of the three outcomes we report results from two specifications: one *without* an interaction between attacks on the judiciary and ruling party personalism; and one *with the interaction.* The specification with the interaction is:

$$Affect_{i,t} = \beta_1 Attack_{i,t} + \beta_2 PersParty_{i,t} + \beta_3 (Attack_{i,t} \times PersParty_{i,t}) + \beta X_{i,t} + \epsilon_{i,t} \qquad (1)$$

We are interested primarily in the interaction coefficient estimate, β_3. In models of polarization we expect this to be positively signed, indicating that an attack increases polarization more when the ruling party is highly personalist than when the ruling party is less personalist. In contrast, for models of out-party affect we expect a negative estimate for the interaction term, indicating that attacks *reduce* out-party affect more when the ruling party is more personalist.

Results The first two column of Table 2 report results for partisan affective polarization. The estimate for *Attacks on judiciary* is positive and significant (at the 0.10 level) in the first column. The second column reports the interaction specification, with a positive and significant (at the 0.10 level) estimate for the interaction between *Attacks on judiciary* and *Ruling party personalism.* This suggests that attacks increase polarization and this pattern is strongest when the ruling party is more personalist, consistent with the theoretical expectations.

The next two columns report results for in-party affect. We find no substantive results linking attacks on the judiciary to in-party affect – either an average effect or one moderated by ruling party personalism.

The final two columns of Table 2 report results for out-party affect. The estimate for *Attacks on judiciary* is negative and significant in column (5), indicating that attacks on the judiciary decrease out-party affect. The estimate for the interaction effect, reported in column (6), is also negative and significant, suggesting that the negative relationship between attacks and out-party affect is strongest when the ruling party is more personalist.

Table 2: Government attacks on democracy and micro-polarization

	Polarization		In-party affect		Out-party affect	
	(1)	(2)	(3)	(4)	(5)	(6)
Attack on judiciary	1.151	-1.209	-0.745	-0.956	-2.074*	0.519
	(0.622)	(1.265)	(0.577)	(0.990)	(0.654)	(0.896)
Ruling party personalism	0.618	-1.836	0.248	0.017	-0.273	2.425*
	(0.449)	(1.209)	(0.219)	(0.745)	(0.444)	(0.906)
Attack on jud. × Ruling party pers.		6.177		0.566		-6.787*
		(3.153)		(1.677)		(2.308)
Covariates	✓	✓	✓	✓	✓	✓
# of elections	82	82	89	89	82	82
# of countries	36	36	39	39	36	36

* indicates statistical significance at the 0.05 level. Cluster robust standard errors. Covariates include: party id; left-right polarization; effective number of parties; presidential system; corruption; GDP per capita; a time trend; democracy age; and initial level of democracy when the leader is selected into power.

Next we check the interaction effects using kernel regression and plot the results in Figure 3.[22] The left plot shows the substantive result for affective polarization; the average level of polarization in the sample is about 4.5 on a 10-point scale and the standard deviation is just under one point. At low levels of ruling party personalism, attacks have no effect on polarization, but the marginal effect of attacks increases as ruling party personalism increases. At the highest ruling party personalism levels, attacks on the judiciary boost polarization by as much as 3 points, which is an effect size similar to the difference between polarization in the U.K. in 1997 (4) and polarization in Turkey in before the 2016 coup attempt (7). The middle plot shows that government attacks on the judiciary have no influence on in-party affect, across the full range of ruling party personalism.

The right plot of Figure 3 shows that attacks decrease out-party affect but only at high levels of ruling party personalism. Substantively, these attacks as associated with as much as a four-point decline in out-party affect, which is the difference between out-party affect in the Netherlands (5) in 2006 and Hungary (1) in 2018.

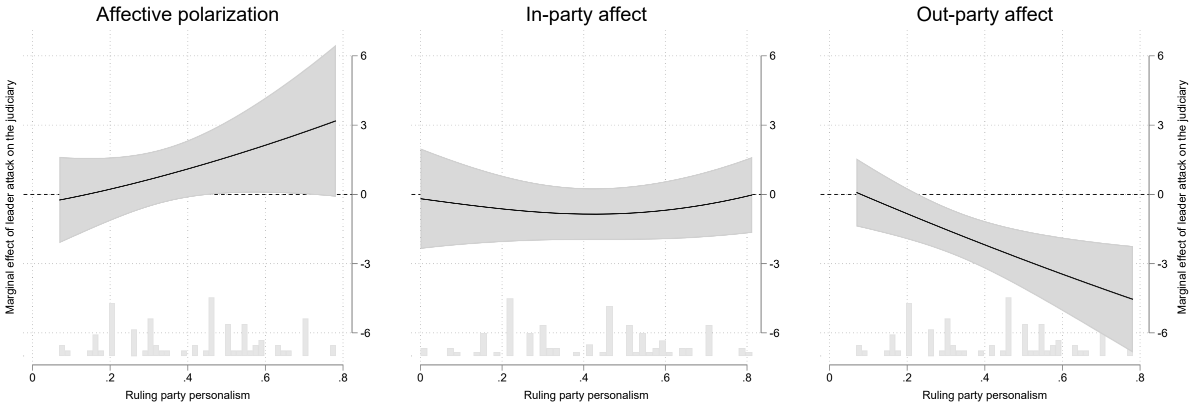


Figure 3: Government attacks on democracy and affective polarization

In reproduction files we show these patterns remain when we alter the specification and are, in fact, slightly stronger when we measure polarization and out-party affect using leaders' names instead of party labels. We also test whether results hold when using measures of affect for all voters and not just for partisans. While the direction of the patterns is the same, the estimates are substantively smaller and generally not statistically significant. These weaker results are consistent with our theory because partisan voters should be more receptive to elite cues than non-partisan voters when interpreting and justifying party leaders' behavior that undermines democracy.

Overall, the patterns in the CSES data suggest that government attacks on democracy boost polarization when ruling parties are personalist. Further, the channel through which this occurs is out-party affect, not in-party affect. However, given the aggregate nature of the in-party and out-party affect measures we cannot distinguish between out-party affect among partisans who support the ruling party and those who voted for opposition parties. The whataboutism channel of polarization we identify in the survey experiment reflects the attitudes of partisans who back a party that undermines democracy. Thus with this data we would need to measure out-party affect separately for ruling party partisans and opposition party partisans instead of lumping them

[22] Kernel regression flexibly estimates the functional form of the marginal effect of the treatment across values of the moderator. This approach also relaxes linear functional form assumptions for covariate marginal effect estimates, protecting against misspecification bias.

together: the whataboutism channel would entail government attacks on democracy decreasing out-party affect for ruling party – but not necessarily opposition – party partisans. That said, we might still expect opposition party supporters to like the out-party less when that party is the ruling party attacks democracy.

Study 3: Global data on polarization in democracies

The final test examines the macro-implications of the theory with expert-coded observational data on polarization and designs that identify causal effects by addressing time-varying confounding using generalized difference-in-difference estimators (Bai, 2009; Xu, 2017; Athey et al., 2021; Liu, Wang and Xu, 2021). We again operationalize government attacks on democracy using the data on incumbent attacks on the judiciary described in the prior section. Data on macro polarization is from the Varieties of Democracy data set; and we measure ruling party personalism with original data from Frantz et al. (2022).[23] This project collects objective information on ruling party personalism from prior to the leader taking executive office; the variable is therefore *exogenous* to the leader's strategic behavior in office towards the party (and party elites) that may shape attacks on democracy and political polarization.

Design We test the main expectation – that attacks on democracy boost polarization and that this effect is stronger when ruling party personalism is higher – with a series of two-way fixed effects models. The global sample encompasses 103 countries with democracies over the three decades from 1991 to 2020. The baseline model with country- and year-fixed effects is a generalized difference-in-difference (DiD) that accounts from all global time trends in the data as well as all time-invariant features of different countries, including electoral rules, presidentialism, the stock of historical democracy, and autocratic legacies that shape party systems – all factors that either change very slowly over time or are fixed for each country. We first test a set of models – without an interaction between attacks on democracy and party personalism and one with this interaction – for the baseline two-way FE model.

Next, we adjust for the initial level of polarization in each country when the leader of the country is first selected into power. For example, since U.S. President Donald Trump's was selected into office in 2016 the measure of polarization in this specification is for the year 2015. Similarly, Israeli Prime Minister was selected into power for his second stint in power in 2009; so the initial level of polarization is 2008. By adjusting for initial polarization we block the channel by which prior polarization or trends in polarization that cause selection into ruling party personalism confound the estimate for ruling party attacks on democracy.

Sticking with the model that adjusts for selection-year polarization we adjust for potential confounders: initial democracy level, democracy age, election year, and the initial level of judicial independence when the leader is first elected to office. The first two confounders are proxies for democratic consolidation that can vary over time within countries; and polarization increases in election years. Further, by adjusting for initial levels of judicial independence we ensure that estimates for attacks on democracy, which we measure as government attacks on the judiciary, are not confounded by selection into different levels of judicial strength that might influence both incumbent attempts to undermine judicial independence and might be greater in weaker, more polarized democracies.

Finally, we estimate dynamic panel models that substitute the lagged outcome for the initial level of polarization. This transforms the outcome to year-to-year changes in polarization $(Y_{i,t} - Y_{i,t-1})$

[23] See Haggard and Kaufman (2021), Bryan (2023), Piazza (2023), and Treisman (2023) for studies of democratic backsliding that use this V-Dem polarization variable.

and blocks the causal pathway by which past attacks on the judiciary boost prior polarization.[24]

Results The estimates in Table 3 for attacks on the judiciary are all positive and significant in the even numbered columns, indicating that the average treatment effect is positive. Note that, as expected, the dynamic treatment estimate in (7) is substantively smaller than the others because it only captures year-on-year changes in polarization and not the cumulative effect over multiple years of a leader's tenure in power. The interaction models reported in even-numbered columns all have a positive and significant estimate for the interaction between attacks on the judiciary and ruling party personalism. This indicates that personalism moderates the treatment effect: government attacks on the judiciary are substantially higher when the ruling party is more personalist. This is consistent with our theoretical expectation because elites in personalist ruling parties, we posit, are less likely than elites in non-personalist parties, to condemn government attacks on democracy, amplifying the polarizing effect of these attacks.

Table 3: Government attacks on democracy and macro-polarization

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Attack on judiciary	2.248*	1.324*	1.577*	0.892*	1.585*	0.907*	0.480*	0.231
	(0.283)	(0.462)	(0.214)	(0.329)	(0.220)	(0.329)	(0.104)	(0.157)
Ruling party personalism	0.104	-0.517*	0.014	-0.443*	0.021	-0.439*	0.054	-0.117
	(0.097)	(0.257)	(0.075)	(0.191)	(0.077)	(0.189)	(0.033)	(0.081)
Attack on jud. × Pers.		1.454*		1.068*		1.074*		0.399*
		(0.596)		(0.429)		(0.426)		(0.186)
Country FE	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
$Y_{t=0}$			✓	✓	✓	✓		
Covariates					✓	✓	✓	✓
Y_{t-1}							✓	✓
N × T	2359	2359	2302	2302	2302	2302	2302	2302
# Leaders	583	583	567	567	567	567	567	567

* indicates statistical significance at the 0.05 level. Standard errors clustered on leader. Covariates include: initial level of judicial independence in leader selection year; initial level of democracy in leader selection year; election year; and democracy age (log). $Y_{t=0}$ is the level of polarization in the leader selection year and Y_{t-1} is the level of polarization in the prior year.

Figure 4 shows the substantive result of the moderating effect of ruling party personalism from the dynamic panel model in (8).[25] At low levels of ruling party personalism (0.3 on the horizontal axis) the estimated effect of attacks on democracy on polarization is 0.2, or one-fifth of one-standard deviation. However, at high levels of personalism (0.75 on the horizontal axis) the marginal effect is more than three times this size (over 0.6) and statistically significant. Further, the estimates indicate that the interactive effect is roughly linear.

These findings are robust, as reported in reproduction files. We adjust for additional confounders, both in isolation and interacted with the treatment variable: ruling party populism, democratic consolidation, initial judicial independence, party system institutionalization, presidentialism, and ruling party seat share. To ensure the dynamic panel model provides a causal estimate, we check that the dynamic panel result holds when adding additional lags of the outcome and when we estimate

[24] This estimator also purges the model of serial correlation; nonetheless we report cluster robust errors that allow for within-panel serial correlation.

[25] We use a kernel estimator that allows for potential nonlinear interaction effects.

an interactive fixed effects model that allows for time-invariate factors captured in the fixed effects to vary across time periods. Finally, we test a counterfactual fixed effects estimator that assesses the assumption of no pre-treatment trend in the outcome (Liu, Wang and Xu, 2021). Again, we find robust results and demonstrate using a placebo test that the identifying causal assumptions are plausible.

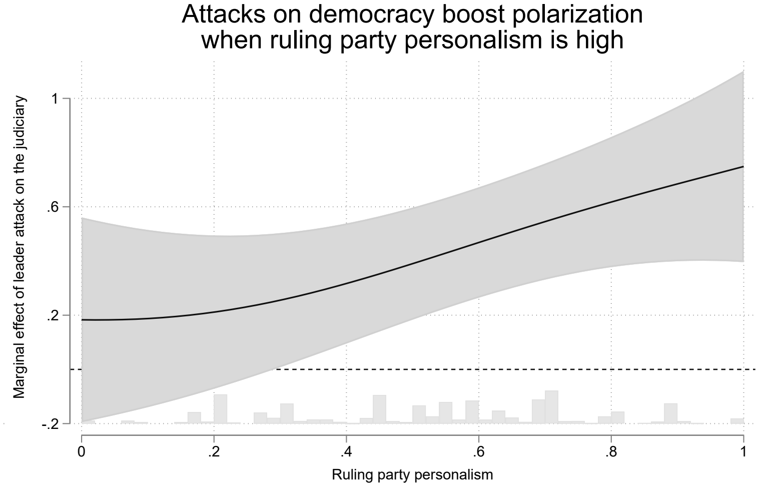


Figure 4: Government attacks on democracy and polarization

The findings are consistent with our expectation: incumbent attacks on the judiciary increase polarization in democracies and ruling party personalism amplifies the polarizing effect of these attacks. The estimated causal effect is not conditional on the level of democratic consolidation, presidentialism, party system institutionalization, ruling party seat share, or even populism, suggesting the macro-relationships are consistent across various types of democracies. Further, we establish that prior trends in polarization that might cause selection into ruling party personalism or incentivize the government to attack democracy do not account for the result.

Discussion

References

Albertus, Michael and Guy Grossman. 2021. "The Americas: When Do Voters Support Power Grabs?" *Journal of Democracy* 32(2):116–131.

Alesina, Alberto and Guido Tabellini. 2007. "Bureaucrats or politicians? Part I: a single policy task." *American Economic Review* 97(1):169–179.

Arceneaux, Kevin. 2008. "Can partisan cues diminish democratic accountability?" *Political Behavior* 30:139–160.

Armaly, Miles T. 2018. "Extra-judicial actor induced change in Supreme Court legitimacy." *Political Research Quarterly* 71(3):600–613.

Armaly, Miles T and Adam M Enders. 2022. "Affective polarization and support for the US Supreme court." *Political Research Quarterly* 75(2):409–424.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens and Khashayar Khosravi. 2021. "Matrix completion methods for causal panel data models." *Journal of the American Statistical Association* pp. 1–15.

Bai, Jushan. 2009. "Panel data models with interactive fixed effects." *Econometrica* 77(4):1229–1279.

Bankert, Alexa. 2021. "Negative and positive partisanship in the 2016 US presidential elections." *Political Behavior* 43(4):1467–1485.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins, Teppei Yamamoto, James N Druckman and Donald P Green. 2021. "Conjoint survey experiments." *Advances in Experimental Political Science* 19.

Barr, Robert R. 2009. "Populists, outsiders and anti-establishment politics." *Party Politics* 15(1):29–48.

Bartels, Brandon L and Christopher D Johnston. 2020. *Curbing the court: Why the public constrains judicial independence*. Cambridge University Press.

Bartels, Larry M. 2020. "Ethnic antagonism erodes Republicans' commitment to democracy." *Proceedings of the National Academy of Sciences* 117(37):22752–22759.

Benedetto, Giacomo, Simon Hix and Nicola Mastrorocco. 2020. "The rise and fall of social democracy, 1918–2017." *American Political Science Review* 114(3):928–939.

Berman, Sheri and Maria Snegovaya. 2019. "Populism and the decline of social democracy." *Journal of Democracy* 30(3):5–19.

Bermeo, Nancy. 2016. "On democratic backsliding." *Journal of Democracy* 27(1):5–19.

Bisgaard, Martin and Rune Slothuus. 2018. "Partisan elites as culprits? How party cues shape partisan perceptual gaps." *American Journal of Political Science* 62(2):456–469.

Blauberger, Michael and R Daniel Kelemen. 2017. "Can courts rescue national democracy? Judicial safeguards against democratic backsliding in the EU." *Journal of European Public Policy* 24(3):321–336.

Boxell, Levi, Matthew Gentzkow and Jesse M Shapiro. 2020. Cross-country trends in affective polarization. Technical report National Bureau of Economic Research.

Bryan, James D. 2023. "What Kind of Democracy Do We All Support? How Partisan Interest Impacts a Citizen's Conceptualization of Democracy." *Comparative Political Studies* p. 00104140231152784.

Buisseret, Peter and Richard Van Weelden. 2020. "Crashing the party? Elites, outsiders, and elections." *American Journal of Political Science* 64(2):356–370.

Caldeira, Gregory A and James L Gibson. 1992. "The etiology of public support for the Supreme Court." *American journal of political science* pp. 635–664.

Carey, John, Katherine Clayton, Gretchen Helmke, Brendan Nyhan, Mitchell Sanders and Susan Stokes. 2020. "Who will defend democracy? Evaluating tradeoffs in candidate support among partisan donors and voters." *Journal of Elections, Public Opinion and Parties* pp. 1–16.

Carothers, Thomas and Andrew O'Donohue. 2019. *Democracies Divided: The Global Challenge of Political Polarization.* Brookings Institutions Press.

Carreras, Miguel. 2012. "The rise of outsiders in Latin America, 1980–2010: An institutionalist perspective." *Comparative Political Studies* 45(12):1451–1482.

Chiopris, Caterina, Monika Nalepa and Georg Vanberg. 2021. A Wolf in Sheep's Clothing: Citizen Uncertainty and Democratic Backsliding. Technical report University of Chicago Working Paper.

Claassen, Christopher. 2020. "Does public support help democracy survive?" *American Journal of Political Science* 64(1):118–134.

Claassen, Ryan L and Michael J Ensley. 2016. "Motivated reasoning and yard-sign-stealing partisans: Mine is a likable rogue, yours is a degenerate criminal." *Political Behavior* 38(2):317–335.

Driscoll, Amanda and Michael J Nelson. 2022. "The Costs of Court Curbing: Evidence from the United States." *Journal of Politics* 85(1).

Druckman, James N and Matthew S Levendusky. 2019. "What do we measure when we measure affective polarization?" *Public Opinion Quarterly* 83(1):114–122.

Druckman, James N, Samara Klar, Yanna Krupnikov, Matthew Levendusky and John Barry Ryan. 2021. "How affective polarization shapes Americans' political beliefs: A study of response to the COVID-19 pandemic." *Journal of Experimental Political Science* 8(3):223–234.

Epstein, Reid J., Ruth Igielnik and Camille Baker. 2023. "These new poll numbers show why Biden and Trump are stuck in a 2024 dead heat." PBS News.

Fishkin, Joseph and David E Pozen. 2018. "Asymmetric constitutional hardball." *Columbia Law Review* 118(3):915–982.

Frantz, Erica, Andrea Kendall-Taylor, Carisa Nietsche and Joseph Wright. 2021. "How Personalist Politics Is Changing Democracies." *Journal of Democracy* 32(3):94–108.

Frantz, Erica, Andrea Kendall-Taylor, Jia Li and Joseph Wright. 2022. "Personalist Ruling Parties in Democracies." *Democratization* 29(5):918–938.

Frantz, Erica, Andrea Kendall-Taylor and Joseph Wright. 2024. *The Origins of Elected Strongment: How Personalist Parties Undermine Democracy from Within*. Oxford University Press.

Gaines, Brian J, James H Kuklinski, Paul J Quirk, Buddy Peyton and Jay Verkuilen. 2007. "Same facts, different interpretations: Partisan motivation and opinion on Iraq." *The Journal of Politics* 69(4):957–974.

Gibler, Douglas M and Kirk A Randazzo. 2011. "Testing the effects of independent judiciaries on the likelihood of democratic backsliding." *American Journal of Political Science* 55(3):696–709.

Gidron, Noam, James Adams and Will Horne. 2020. *American affective polarization in comparative perspective*. Cambridge University Press.

Graham, Matthew H and Milan W Svolik. 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.

Grillo, Edoardo and Carlo Prato. 2021. "Reference Points and Democratic Backsliding." *American Journal of Political Science* https://doi.org/10.1111/ajps.12672.

Guess, Andrew and Alexander Coppock. 2020. "Does counter-attitudinal information cause backlash? Results from three large survey experiments." *British Journal of Political Science* 50(4):1497–1515.

Haggard, Stephan and Robert Kaufman. 2021. *Backsliding: Democratic Regress in the Contemporary World*. Cambridge University Press.

Horz, Carlo M. 2021. "Electoral Manipulation in Polarized Societies." *The Journal of Politics* 83(2):483–497.

Iyengar, Shanto, Gaurav Sood and Yphtach Lelkes. 2012. "Affect, not ideology: a social identity perspective on polarization." *Public opinion quarterly* 76(3):405–431.

Iyengar, Shanto and Masha Krupenkin. 2018. "Partisanship as Social Identity; Implications for the Study of Party Polarization." *The Forum* 16(1):23–45.

Jerit, Jennifer and Jason Barabas. 2012. "Partisan perceptual bias and the information environment." *The Journal of Politics* 74(3):672–684.

Kam, Cindy D. 2005. "Who toes the party line? Cues, values, and individual differences." *Political behavior* 27:163–182.

Kostadinova, Tatiana and Barry Levitt. 2014. "Toward a theory of personalist parties: Concept formation and theory building." *Politics & Policy* 42(4):490–512.

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological Bulletin* 108(3):480.

Larkins, Christopher M. 1996. "Judicial Independence and Democratiziation: a Theoritical and conceptual analysis." *American Journal of Comparative Law* 44:605.

Lee, Amber Hye-Yon, Yphtach Lelkes, Carlee B Hawkins and Alexander G Theodoridis. 2022. "Negative partisanship is not more prevalent than positive partisanship." *Nature Human Behaviour* pp. 1–13.

Lelkes, Yphtach and Sean J Westwood. 2017. “The limits of partisan prejudice.” *The Journal of Politics* 79(2):485–501.

Levitsky, Steven and Daniel Ziblatt. 2018. *How democracies die.* Crown.

Liu, Licheng, Ye Wang and Yiqing Xu. 2021. “A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data.” *arXiv preprint arXiv:2107.00856* .

Loffman, Matt. 2023. “Voters Are Dreading a Trump-Biden Rematch. Enter R.F.K. Jr.” New York Times.

Lord, Charles G, Lee Ross and Mark R Lepper. 1979. “Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence.” *Journal of Personality and Social Psychology* 37(11):2098.

Lucas, Edward. 2008. “Whataboutism.” *The Economist* 31 January 2008.

McCoy, Jennifer and Murat Somer. 2019. “Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies.” *The Annals of the American Academy of Political and Social Science* 681(1):234–271.

Mudde, Cas and Cristóbal Rovira Kaltwasser. 2018. “Studying Populism in Comparative Perspective: Reflections on the Contemporary and Future Research Agenda.” *Comparative Political Studies* 51(13):1667–1693.

Nelson, Michael J and James L Gibson. 2019. “How does hyperpoliticized rhetoric affect the US Supreme court’s legitimacy?” *The Journal of Politics* 81(4):1512–1516.

North, Douglass C and Barry R Weingast. 1989. “Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England.” *The Journal of Economic History* 49(4):803–832.

Parker-Stephen, Evan. 2013. “Tides of disagreement: How reality facilitates (and inhibits) partisan public opinion.” *The Journal of Politics* 75(4):1077–1088.

Piazza, James A. 2023. “Political polarization and political violence.” *Security Studies* .

Reenock, Christopher, Jeffrey K Staton and Marius Radean. 2013. “Legal institutions and democratic survival.” *The Journal of Politics* 75(2):491–505.

Reiljan, Andres, Diego Garzia, Frederico Ferreira Da Silva and Alexander H Trechsel. 2023. “Patterns of Affective Polarization toward Parties and Leaders across the Democratic World.” *American Political Science Review* pp. 1–17.

Rogowski, Jon C and Andrew R Stone. 2021. “How political contestation over judicial nominations polarizes Americans’ attitudes toward the Supreme Court.” *British Journal of Political Science* 51(3):1251–1269.

Samuels, David J and Matthew Soberg Shugart. 2003. “Presidentialism, elections and representation.” *Journal of Theoretical Politics* 15(1):33–60.

Schmemann, Serge. 2023. “Democracy May Be Struggling but the Fight Is Far From Over.” The New York Times.

Sheagley, Geoffrey and Scott Clifford. 2023. "No Evidence that Measuring Moderators Alters Treatment Effects." *American Journal of Political Science* .

Staton, Jeffrey K. 2006. "Constitutional review and the selective promotion of case results." *American Journal of Political Science* 50(1):98–112.

Staton, Jeffrey K. 2010. *Judicial power and strategic communication in Mexico.* Cambridge University Press.

Svolik, Milan W. 2019. "Polarization versus democracy." *Journal of Democracy* 30(3):20–32.

Svolik, Milan W. 2020. "When Polarization Trumps Civic Virtue: Partisan Conflict and the Subversion of Democracy by Incumbents." *Quarterly Journal of Political Science* 15(1):3–31.

Taber, Charles S and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American Journal of Political Science* 50(3):755–769.

Touchton, Michael, Casey Klofstad and Joseph Uscinski. 2020. "Does partisanship promote anti-democratic impulses? Evidence from a survey experiment." *Journal of Elections, Public Opinion and Parties* pp. 1–13.

Treisman, Daniel. 2023. "How great is the current danger to democracy? Assessing the risk with historical data." *Comparative Political Studies* p. 00104140231168363.

Vanberg, Georg. 2001. "Legislative-judicial relations: A game-theoretic approach to constitutional review." *American journal of political science* pp. 346–361.

Vanberg, Georg. 2004. *The politics of constitutional review in Germany.* Cambridge University Press.

Xu, Yiqing. 2017. "Generalized synthetic control method: Causal inference with interactive fixed effects models." *Political Analysis* 25(1):57–76.

Zaller, John. 1991. "Information, values, and opinion." *American Political Science Review* 85(4):1215–1237.

Zaller, John R. 1992. *The nature and origins of mass opinion.* Cambridge University Press.

Zhong, Chen-Bo, Katherine W Phillips, Geoffrey J Leonardelli and Adam D Galinsky. 2008. "Negational categorization and intergroup behavior." *Personality and Social Psychology Bulletin* 34(6):793–806.

Appendix A: Additional analysis for Study 1

Affective polarization We use two questions, each with a five-ordered responses, to measure the outcome, *affective polarization*. We combine the ordinal scales for the two variables separately for each group of partisan identifiers using polychoric correlation instead of Pearson correlation because the former relaxes the assumption of normally distributed data and the Likert-scale data is ordinal not normal. Thus we obtain a measure among Republican id respondents towards Republicans; a measure among Republican id respondents towards Democrats (negative affect); a measure among Democratic id respondents towards Republicans (negative affect); and measure among Democratic id respondents towards Democrats. Table A-2 shows the correlations for these four measures. The correlations for in-party affect (R id → R and D id → D) are higher than for out-party affect (R id → D and D id → R). This suggests that, with these respondents, the survey instruments probably are probably a better measure in-party than out-party effect.

Table A-2: Polychoric correlations for partisan affect

Affect		ρ	Eigenvalue	% explained
In-party	R id → R	0.71	1.71	85
Out-party	R id → D	0.61	1.61	81
Out-party	D id → R	0.60	1.60	80
In-party	D id → D	0.75	1.75	87

Speeders The median value of the time to completion for the survey questionairre (see Appendix B) was 325 seconds. We code speeders as respondents who finished the survey in less than half that time (163 seconds). An initial look at the speeding is shown in Figure A-2. Here we examine the extent to which speeding is correlated with three factors: the share of correct answers to three factual questions about the Supreme Court; the likelihood of passing a basic information check about the treatment; and whether the respondent marked that they liked both Trump and Biden (i.e., greater than 80 points on a 0-100 scale). In each plot in Figure A-2 the horizontal scale is the log of time to completion for each respondent. The density plot in each is the distribution of log time to completion and the vertical dotted line at 163 seconds marked the cutpoint for delineating "speeders" (i.e., 1/2 to median time to completion).

The left plot depicts the share of correct answers to factual Supreme Court questions. The upward-sloping red line an associated confidence interval is the nonlinear fit between the time to completion and share of correct answers. Respondents who completed in the survey in less than 30 seconds did not answer any of the questions correctly, but the share of correct answers rises to over 40 percent for non-speeders. Across all respondents, the share of correct answers is 42 percent; for speeders it is 15 percent and for non-speeders, it is 48 percent.

The middle plot depicts the likelihood of passing the information check. None of the respondents who take 30 seconds or less to complete the survey pass the information check. But as time to completion increases past 163 seconds, the share of respondents who pass the information check rises to over 80 percent. For speeders, on average, 45 percent of respondents pass the check; for non-speeders 92 percent pass the information check.

Finally, the right plot in Figure A-2 depicts the likelihood of liking both Trump and Biden. We believe this outcome simply reflects respondents' inattention to the survey because so few American voters like both politicians. All the respondents who take less than 30 seconds like both leaders. But as the time to completion increases past 163 seconds, the likelihood of liking both leaders drops to less than 15 percent. Among speeders, 52 percent of respondents like both; but for non-speeders only 5 percent of respondents like both.

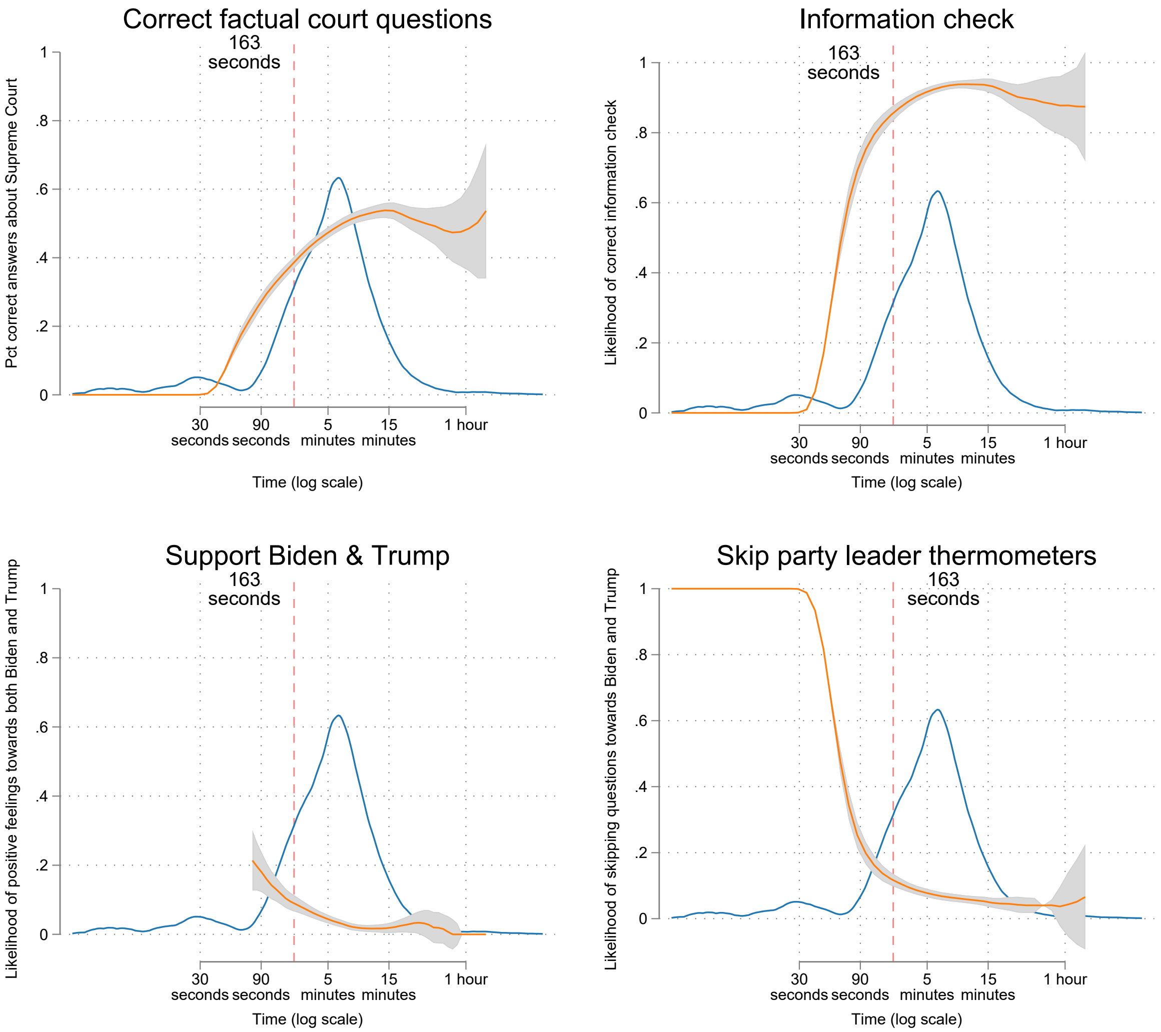


Figure A-2: Speeders and information accuracy

These correlations show that speeding is associated with a large degree of inaccuracy in answering survey questions correctly. These respondents are simply not paying attention to the survey. Thus the conventional cut point of 1/2 the median time to completion has support in information responses indicating that speeders should be dropped from the analysis.

Next we examine the main findings reported in the main text but vary the cut point for speeding. This analysis ensures that the convention of 1/2 the median time to completion is not, by chance, picking up a point in the distribution of time to completion that pushes the findings in one direction or another. The cutpoint should be arbitrary with respect to the estimated treatment effect.

Recall that the main findings for partisan affect reported in the main text (Figure 1) are the following: (a) the combined treatment effect (i.e., all three treatment conditions relative to the control condition) for Republican id respondents on negative outparty affect is positive and significant at the 0.10 level; (b) the estimated effects for the treatment condition alone and the treatment combined with endorsement are positive and statistically significant. Here we show how these two estimates – (a) and (b) – vary as we change the threshold for excluded speeders. The left plot in Figure A-3 shows the estimates for combined treatment effect (a). The horizontal scale depicts the time to completion and the histogram displays the distribution of time to completion. The gray part of the distribution marks respondents we exclude as speeders in the main reported results (i.e. time to completion less than 163 seconds). The green part of the distribution marks non-speeding respondents. The vertical scale measures the estimated treatment effect. Lowering the threshold for speeding produces almost the exact same result as the reported one. While increasing the threshold yields a stronger treatment effect – up to about 7 minutes. We are thus confident that the cut point of 163 is not driving the reported result.

The right plot of Figure A-3 shows the estimated treatment effect for (b): treatment condition alone plus treatment combined with endorsement, relative to the control condition. Again, lowering the speeding threshold yields almost the same sized effect as the reported one and increasing the threshold makes the estimate much stronger. We are thus confident that the cut point of 163 is not driving the reported result.

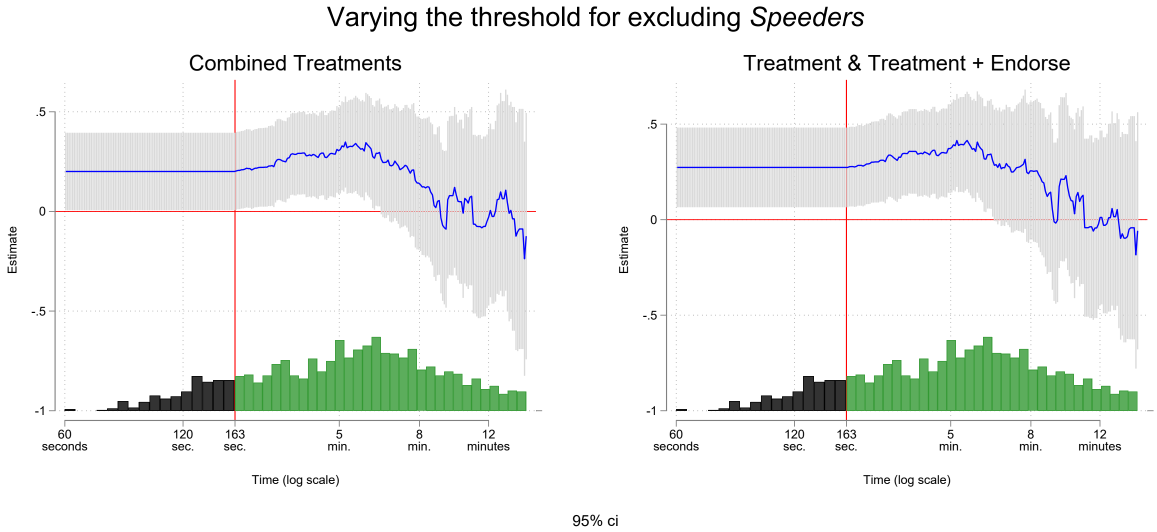


Figure A-3: Estimates by time to completion

Balance tests Table A-3 shows the result of covariate balance tests. For each covariate, we conduct four t-tests. The first is a test of the difference of means between the combined treatment group and the control group. The second is a t-test comparing the treatment only group to the control group. And the third and fourth compare the control to the treatment + condemn and treatment + endorse groups. We report p-values for one-tailed tests in the direction of the difference in means. Two covariates fail the test, with some p-values less than 0.10 or 0.05: strongman values and college education. In the reported tests, we adjust for all seven covariates.

Table A-4 compares the main reported result that adjusts for all covariates with estimates that adjust only for the strongman index and college. The regression only use a sample of Republican id respondents; and omits the comparison of the treatment + condemn condition. The first column indicates that the treatment boosts negative affect by roungly 20 percent relative to the control; this estimate is significant at the 0.05 level. The second column is the same regression but only adjusts

Table A-3: Covariate t-tests

Covariate	Combined treatment	Treatment only	Treatment + condemn	Treatment + endorse
Strongman values	-0.10*	-0.10	-0.08	-0.11*
	(0.04)	(0.07)	(0.10)	(0.04)
Democratic values	0.02	0.02	0.04	-0.01
	(0.38)	(0.38)	(0.30)	(0.46)
Male	- 0.01	0.02	0.00	-0.05
	(0.37)	(0.26)	(0.44)	(0.06)
Age	0.28	0.56	-0.27	0.00
	(0.37)	(0.30)	(0.39)	(0.50)
Rural	0.00	0.01	-0.00	-0.00
	(0.43)	(0.34)	(0.49)	(0.47)
College	0.03	0.06*	0.05*	-0.02
	(0.10)	(0.03)	(0.04)	(0.32)
White	0.00	0.01	0.01	-0.03
	(0.45)	(0.36)	(0.35)	(0.15)

* indicates statistical significance at the 0.05 level. p-values reported in parentheses.

for two covariates with imbalance. The estimate is rougly 19 percent; with no adjustment (column 3) the treatment estimate again is roughly 19 percent. The latter two estimates, however, are significant at the 0.10 levels. In short, changing the specification does not alter the main reported estimate by much.

Positive affect In the main text we reported results for negative partisan affect to examine the whataboutism channel linking leader attacks on democracy to affective partisanship and thus polarization. Here we report results for positive partisan affect. This outcome reflects whether partisans respond to leader attacks on democracy by increasing their in-party affect. If respondents embrace their partisan leader because the leader is a "fighter" who attacks government institutions that constrain the leader, then partisans may increase their in-party affect in response to an attack on democracy. In short, there is no evid

Figure A-4 reports the results for positive partisan affect. In the left plot we reported the combined treatment effect for each group of partisan respondents. The estimate for Democratic party id respondents is positive (0.09) but not significant at the 0.10 level. For Republican respondents, the estimated treatment effect is almost zero. Disaggregating the treatment effects, reported in the right plot, shows that the treatment + condemn condition has the strongest effect (0.12) on positive affect for Democratic id respondents. However, again this estimate is not statistically significant. Even if the estimate for the treatment + condemn were positive and signficant, such as estimate would be inconsistent with our theory, which suggests that this estimate should be zero. In short, there is no evidence that treatment influences positive partisan affect in ways that are consistent with theoretical expectations.

Combined results for covariates and affect In the section we report results from regressions that combine all respondents in the same sample. This allows us to see how partisanship and the covariates influence negative and positive affect, on average. These tests are *NOT* tests of the theory but rather provide some substantive context for interpreting the tests of the experimental treatments.

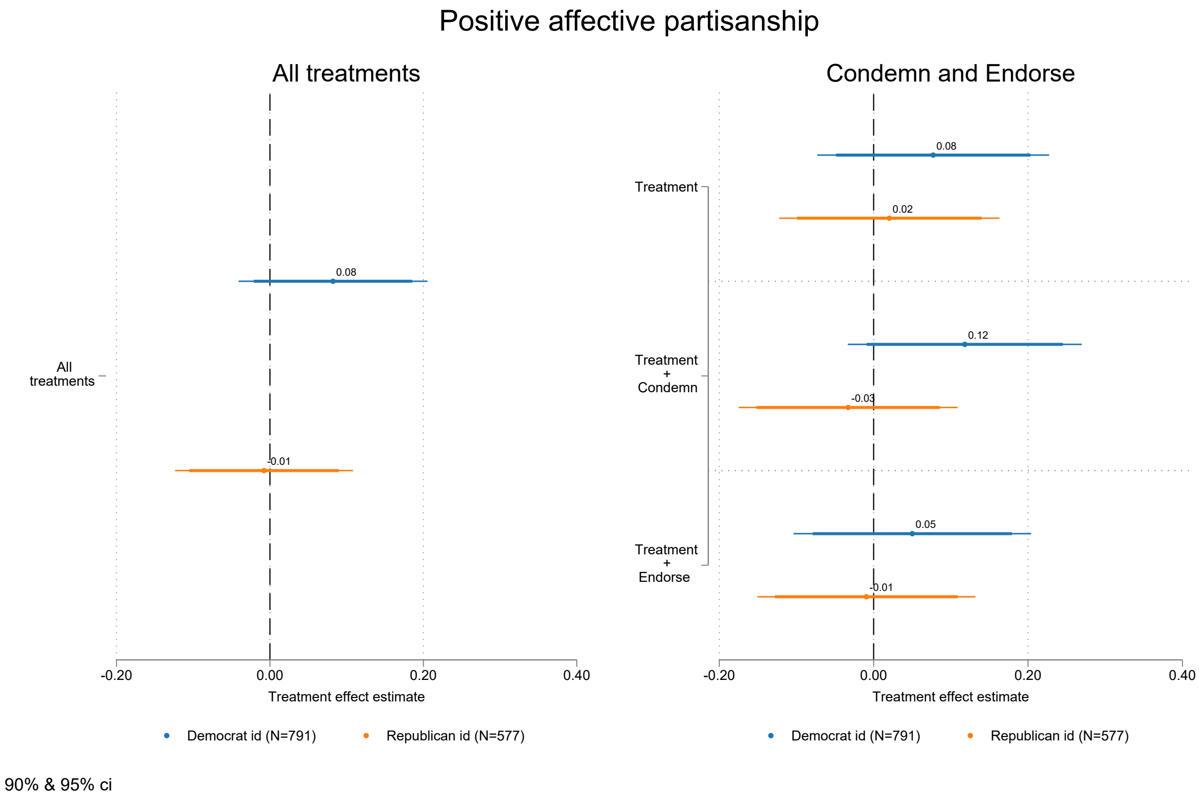


Figure A-4: Estimates by time to completion

Table A-4: Robustness tests for negative partisan affect

	Reported result	Adjust for 2 covariates	No covariate adjustment
	(1)	(2)	(3)
Treatment	0.273*	0.253*	0.257*
	(0.102)	(0.103)	(0.103)
College	0.060	0.005	
	(0.106)	(0.105)	
Strongman	0.056	0.050	
	(0.064)	(0.053)	
White	-0.050		
	(0.163)		
Male	0.305*		
	(0.108)		
Rural	0.119		
	(0.102)		
Age	0.005		
	(0.003)		
Democracy	0.078		
	(0.047)		
(Intercept)	-11.506	-1.196*	-1.202*
	(6.515)	(0.092)	(0.082)

* indicates statistical significance at the 0.05 level. Republican id respondents only. Treatment includes: treatment only & treatment + endorse. Treatment + condemn group dropped from sample.

The left plot of Figure A-5 shows the correlations for negative affect – or the affect that partisans have towards the out-party. The estimate for Democratic id, male and strongman are all positive and statistically significant. This indicates that, on average, Democratic have 0.13 standard deviations more negative affect than Republicans. Similarly, males have 0.13 more negative affect than non-males; and those with a strong affinity for strongman rule have more negative affect (0.11) than those with low affinity for strongman rule. We use the size of these estimates to better interpret and contextualize the treatment effects estimates reported in the main text.

Decomposing polarization: Democratic id only

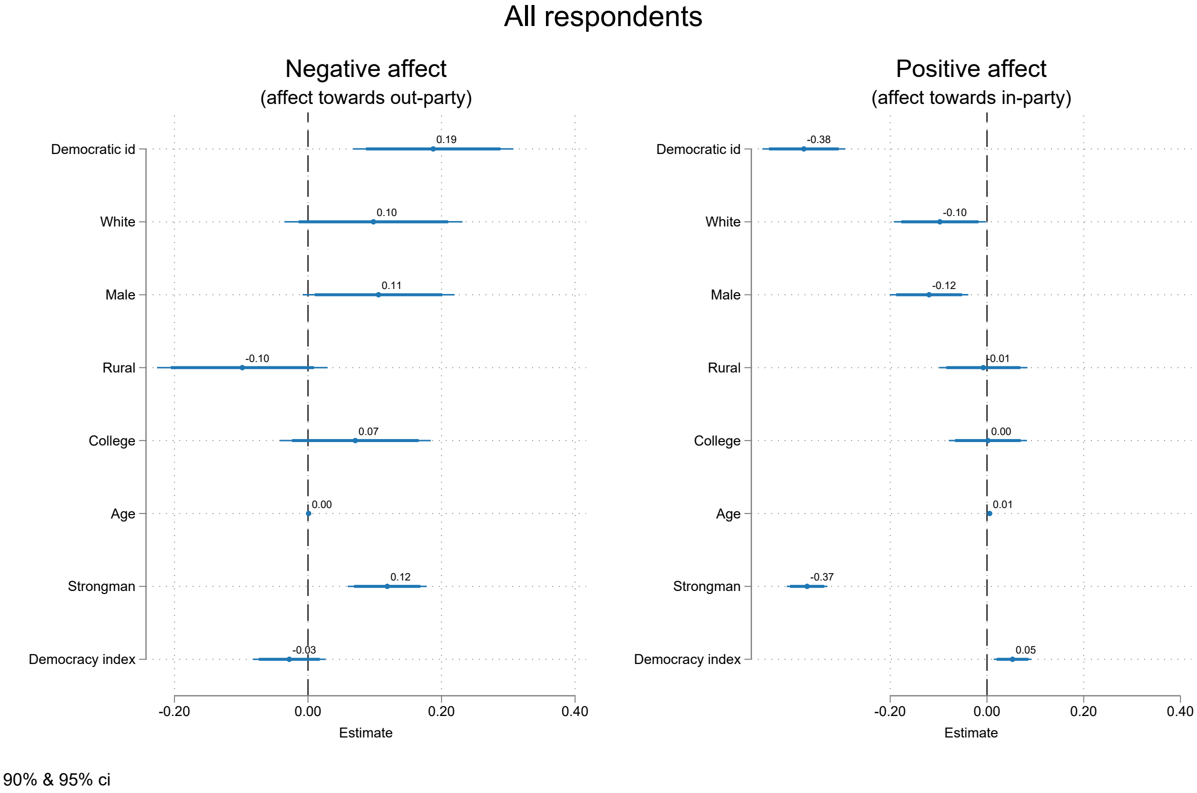


Figure A-5: Combined Republican id and Democratic id results for affect

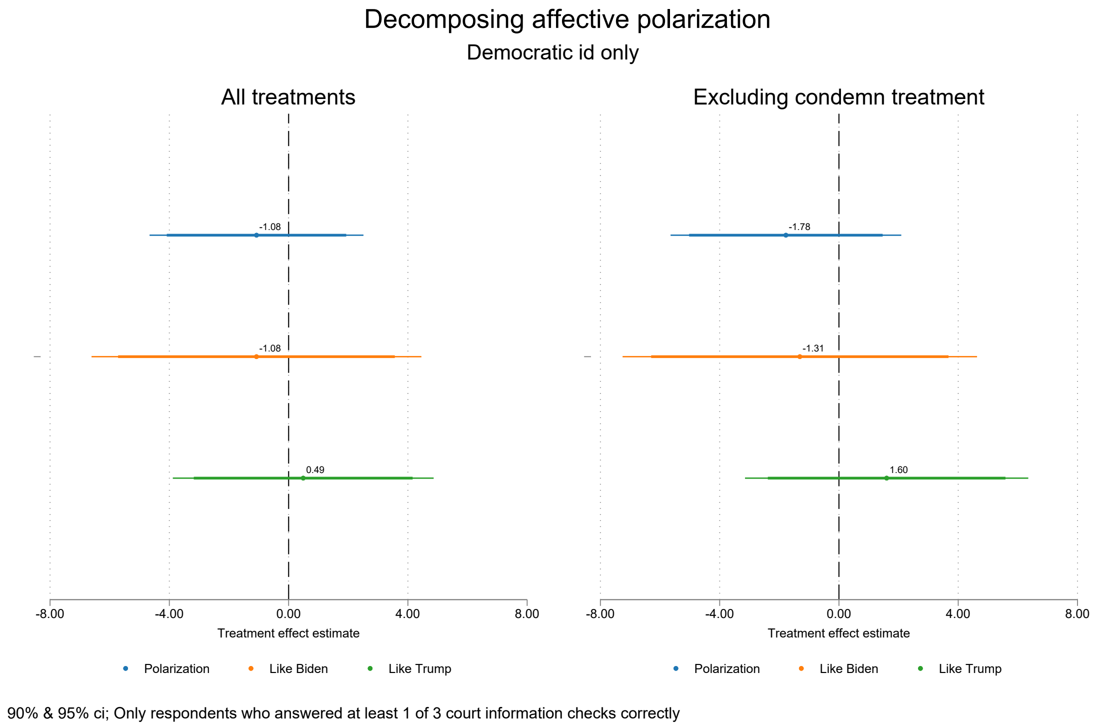


Figure A-6: Decomposing polarization: Democratic id only

Appendix B: Survey questionnaire

Questionnaire

This study is being conducted by Erica Frantz and Joseph Wright, who are professors at Pennsylvania State University (Wright) and Michigan State University (Frantz). We ask for your attention for a few minutes and we thank you for your attention and your responses. Your participation is voluntary and you may decline the survey or withdraw at any time. There are no negative consequences if you don't want to take it or if you withdraw once you have started. No information that identifies you will be collected or retained by the researchers, and all of the information we collect will be stored securely. However, any online interaction carries some risk of being accessed. Please contact the IRB at the Pennsylvania State University with any questions or concerns.

➤ Do you consent to participate in the survey?

- □ Yes
- □ No
- □ Skipped

➤ In what year were you born?

- □ min: 1900
- □ max: 2005
- □ Don't know [8]
- □ Skipped [9]

➤ Generally speaking, do you think of yourself as a:

- □ Democratic
- □ Republican
- □ Independent
- □ Other
- □ Not sure
- □ Skipped [9]

➤ Generally speaking, do you think of yourself as a:

- □ Strong Democratic [if Democrat]
- □ Not very strong Democratic [if Democrat]
- □ Strong Republican [if Republican]
- □ Not very strong Republican [if Republican]
- □ Democratic Party [if Independent, Other, Not Sure]
- □ Republican Party [if Independent, Other, Not Sure]
- □ Neither [if Independent, Other, Not Sure]
- □ Not sure [if Independent, Other, Not Sure]

☐ Don't know [8] [if Independent, Other, Not Sure]

☐ Skipped [9]

Now we're going to ask you some questions about democracy. [Randomize democracy questions order.]

➤ Democracy may have problems, but it is the best system of government.

☐ Strongly agree

☐ Somewhat agree

☐ Do not agree/disagree

☐ Somewhat disagree

☐ Strongly disagree

☐ Don't know [8]

☐ Skipped [9]

➤ Which of the following statements comes closest to your own opinion?

☐ For people like me, it does not matter whether we have a democracy

☐ Under some circumstances, an authoritarian government can be preferable

☐ Democracy is always preferable to any other kind of government

☐ Don't know [8]

☐ Skipped [9]

➤ Governance by a powerful leader without the restriction of a legislature or elections.

☐ Very good

☐ Fairly good

☐ Neither good nor bad

☐ Fairly bad

☐ Very bad

☐ Don't know [8]

☐ Skipped [9]

➤ Best to get rid of Congress and elections and have a strong leader who can quickly decide everything.

☐ Strongly agree

☐ Somewhat agree

☐ Do not agree/disagree

☐ Somewhat disagree

☐ Strongly disagree

☐ Don't know [8]

☐ Skipped [9]

Now we're going to ask you some questions about the U.S. Supreme Court. [Randomize Courts questions order.]

➤ If the U.S. Supreme Court starts making a lot of decisions that most people disagree with, it might be better to do away with the Supreme Court altogether.

□ Strongly agree
□ Somewhat agree
□ Do not agree/disagree
□ Somewhat disagree
□ Strongly disagree
□ Don't know [8]
□ Skipped [9]

➤ The U.S. Supreme Court gets too mixed up in politics.

□ Strongly agree
□ Somewhat agree
□ Do not agree/disagree
□ Somewhat disagree
□ Strongly disagree
□ Don't know [8]
□ Skipped [9]

➤ The U.S. Supreme Court should have the right to say what the Constitution means, even when the majority of the people disagree with the Court's decision.

□ Strongly agree
□ Somewhat agree
□ Do not agree/disagree
□ Somewhat disagree
□ Strongly disagree
□ Don't know [8]
□ Skipped [9]

➤ The U.S. Supreme Court can usually be trusted to make decisions that are right for the country as a whole.

□ Strongly agree
□ Somewhat agree
□ Do not agree/disagree
□ Somewhat disagree
□ Strongly disagree

□ Don't know [8]

□ Skipped [9]

➤ How many justices are usually on the U.S. Supreme Court?

□ Less than five

□ Between five and ten

□ More than ten

□ Don't know [8]

□ Skipped [9]

➤ Who appoints members of the U.S. Supreme Court?

□ State legislatures

□ The Judicial Committee

□ The President

□ The U.S. Supreme Court

□ Don't know [8]

□ Skipped [9]

➤ If the President and Supreme Court differ on whether an action by the president is constitutional, who has the final responsibility for determining if the action is constitutional?

□ The Senate

□ The House of Representatives

□ The President

□ The U.S. Supreme Court

□ Don't know [8]

□ Skipped [9]

Next we're going to provide information about what some politicians have said about the U.S. Supreme Court. First, though, it is important to know some basic facts about the court. In the United States, the President appoints Supreme Court justices when there is a vacancy on the court; and the Senate has to confirm the President's nominee before someone new joins the court. The court has a vacancy either when a current justice dies or a justice voluntarily retires from the position. In the past few years, the workload of the U.S. Supreme Court has increased substantially and the justices work many long hours. In fact, a recent Presidential commission found that Supreme Court Justices work longer hours than most Americans and get very little vacation time.

[For the pilot study, we want to maximize size and thus only test **how self-identified partisans respond to aggrandizement by the leader of their own party** ($S^D \Rightarrow\downarrow A^O_{v_D}$: leader's supporters increase negative affect towards the other party when the leader aggrandizes). This means that we treat Joe Biden as the incumbent leader for self-identified Democrats and Donald Trump as the incumbent leader for self-identified Republicans. The survey *randomly assigns one of four* *statements to a respondent*:]

Now we're going to provide information about what some politicians have said about the U.S. Supreme Court.

[For respondents who identify as **Democrats**]:

❶ *Treatment*: President Joe Biden recently responded to a Supreme Court ruling by suggesting that the President should be able to fire justices or expand the number of justices on the court to get more people on the court who agree with him.

❷ *Treatment + Endorsement*: President Joe Biden recently responded to a Supreme Court ruling by suggesting that the President should be able to fire Supreme Court justices or expand the number of justices on the court to get more people on the court who agree with him. Senior Democratic leader Chuck Schumer, head of the Senate, backed Biden's proposal to change the court composition to make it more friendly to President Biden. "This is the only sensible path forward for our democracy. The President is the elected leader. The Supreme Court should not be making policy; Joe Biden should," Schumer said.

❸ *Treatment + Condemnation*: President Joe Biden recently responded to a Supreme Court ruling by suggesting that the President should be able to fire Supreme Court justices or expand the number of justices to get more people on the court who agree with him. Senior Democratic leader Chuck Schumer, head of the Senate, strongly condemned Biden's proposal to change the court composition to make it more friendly to President Biden. "Allowing the President to change the composition of the Supreme Court when he wants will pose a threat to our democracy, if not now then in the future," Schumer said.

❹ *Control*: President Joe Biden recently responded to the Presidential commission's finding that Supreme Court justices are overworked by suggesting that justices should be eligible for increased vacation time.

[For respondents who identify as **Republicans**]:

❶ *Treatment*: When Donald Trump was President, he responded to a Supreme Court ruling by suggesting that the President should be able to fire Supreme Court justices or expand the number of justices on the court to get more people on the court who agree with him.

❷ *Treatment + Endorsement*: When Donald Trump was President, he responded to a Supreme Court ruling by suggesting that the President should be able to fire Supreme Court judges or expand the number of justices on the court to get more people on the court who agree with him. Senior Republican leader Mitch McConnell, head of the Senate, backed Trump's proposal to change the court composition to make it more friendly to President Trump. "This is the only sensible path forward for our democracy. The President is the elected leader. The Supreme Court should not be making policy; Donald Trump should," McConnell said.

❸ *Treatment + Condemnation*: When Donald Trump was President, he responded to a Supreme Court ruling by suggesting that the President should be able to fire Supreme Court judges or expand the number of justices on the court to get more people on the court who agree with him. Senior Republican leader Mitch McConnell, head of the Senate, strongly condemned Trump's proposal to change the court composition to make it more friendly to President

Trump. "Allowing the President to change the composition of the Supreme Court when he wants will pose a threat to our democracy, if not now then in the future," McConnell said.

❹ *Control*: When Donald Trump was President, he responded to the Presidential commission's finding that Supreme Court justices are overworked by suggesting that justices should be eligible for increased vacation time.

Next we would like to get your feelings toward some of our political leaders and political parties. You will see the name of a person and we would like you to rate that person using something we call the feeling thermometer. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the person and that you don't care too much for that person. You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person. [Randomize leader questions order.]

➤ How do you rate Joe Biden?

☐ Joe Biden [0-100]
☐ Prefer not to answer [7]
☐ Don't know [8]
☐ Skipped [9]

➤ How do you rate Donald Trump?

☐ Donald Trump [0-100]
☐ Prefer not to answer [7]
☐ Don't know [8]
☐ Skipped [9]

➤ Does a higher rating, one close to 100, mean you liked or disliked the politician?

☐ Disliked
☐ Lied
☐ Don't know [8]
☐ Skipped [9]

Now we're going to ask you some questions about political parties. First, we're going to ask about the Republican Party. [Randomize party questions order.]

➤ When people criticize the Republican Party, it makes me feel good"'

☐ Strongly agree
☐ Somewhat agree
☐ Do not agree/disagree
☐ Somewhat disagree
☐ Strongly disagree

□ Prefer not to answer [7]

□ Don't know [8]

□ Skipped [9]

➤ When I meet someone who supports the Republican Party, I feel disconnected.

□ Strongly agree

□ Somewhat agree

□ Do not agree/disagree

□ Somewhat disagree

□ Strongly disagree

□ Prefer not to answer [7]

□ Don't know [8]

□ Skipped [9]

Now we're going to ask you some questions about the Democratic Party.

➤ When people criticize the Democratic Party, it makes me feel good.

□ Strongly agree

□ Somewhat agree

□ Do not agree/disagree

□ Somewhat disagree

□ Strongly disagree

□ Prefer not to answer [7]

□ Don't know [8]

□ Skipped [9]

➤ When I meet someone who supports the Democratic Party, I feel disconnected.

□ Strongly agree

□ Somewhat agree

□ Do not agree/disagree

□ Somewhat disagree

□ Strongly disagree

□ Prefer not to answer [7]

□ Don't know [8]

□ Skipped [9]

Now we're going to ask you some questions about yourself.

➤ What is your gender?

- □ Female
- □ Male
- □ Other
- □ Prefer not to answer [7]
- □ Skipped [9]

➤ What is the highest degree or level of school you have completed?

- □ Did not graduate from high school
- □ High school diploma or the equivalent (GED)
- □ Some college
- □ Bachelor's degree
- □ Graduate degree: Masters degree, Professional degree or Doctorate degree
- □ Prefer not to answer [7]
- □ Skipped [9]

➤ How do you mainly spend your time? Are you currently:

- □ Working, full time?
- □ Working, part time?
- □ Not working, but have a job?
- □ Actively looking for a job?
- □ A student?
- □ Retired, a pensioner or permanently disabled to work?
- □ Not working and not looking for a job?
- □ Prefer not to answer [7]
- □ Skipped [9]

➤ Do you live in:

- □ A city?
- □ On the outskirts or surroundings of a city/suburbs?
- □ In a town near a rural area/zone?
- □ Prefer not to answer [7]
- □ Skipped [9]

➤ Do you belong to a religion or religious denomination? If yes, which one?

- □ No: do not belong to a religious denomination [0]
- □ Yes: Roman Catholic
- □ Yes: Protestant
- □ Yes: Orthodox (Russian/Greek/etc.)
- □ Yes: Jew

☐ Yes: Muslim

☐ Yes: Hindu

☐ Yes: Buddhist

☐ Yes: Other

☐ Prefer not to answer [7]

☐ Skipped [9]

➤ What racial or ethnic group best describes you?

☐ White

☐ Black

☐ Hispanic or Latino

☐ Asian

☐ Native American

☐ Middle Eastern

☐ Two or more races

☐ Other

☐ Prefer not to answer [7]

☐ Skipped [9]

Thanks for answering all of these questions. Now we're going to ask you one last question about the survey you just completed.

➤ Who do you think paid for and conducted this survey?

☐ Republican party

☐ Democratic party

☐ University researchers

☐ The government

☐ A news or media organization

☐ Prefer not to answer [7]

☐ Don't know [8]

☐ Skipped [9]

Earlier, the survey provided you with statements that politicians made about the U.S. Supreme Court. It is important that you know that these statement are **fictitious**. We have no evidence that these exact statements can be attributed to these politicians. That said, Joe Biden and Donald Trump have both verbally criticized the U.S. Supreme Court. For example, President Joe Biden said the following in the context of the court ruling that overturned *Roe vs. Wade* (this ruling, *Dobbs*, allowed states to make abortion access illegal): "The Supreme Court is more of an advocacy group these days than it is ... evenhanded about it."[26] Donald Trump criticized the Supreme Court after it rejected his request to block Congress from obtaining his tax records: "The Supreme Court has lost its honor, prestige, and standing, & has become nothing more than a political body, with our Country paying the price."[27]

[26] "Biden says Supreme Court is 'more of an advocacy group' than 'evenhanded' ". Rebecca Shabad. *NBC News* online. 12 October 2022).

[27] "Trump rips the Supreme Court..." Kelsey Vlamis. *Business Insider* online. 23 November 2022.

Appendix C: CSES analysis

Table C-1 shows results from a reproduction and an extension of the main modeled of party affective polarization, presented in Reiljan et al. (2023). The first column is an exact production of the result in column 1 of their Table 2. The second column omits Government effectiveness and substitutes measures of related concepts, GDP per capita and public sector corruption. We change the specification in this way for our tests because government effective is conceptually and empirically highly correlated with government attacks on the judiciary. Thus, for a very ordinary technical reason – collinearity – we make this substitution. However, in Figure **??** we show that main finding does not change appreciably when include Government effectiveness in the specification. The final column uses the same specification as in (2) but alters the sample to match the sample we use in our analysis (see main text).

The results indicate that altering the specification or the sample slightly has no material effect on the results reported in column (1). When change the specification as a result of collinearity issues that arise in our analysis and change the sample due to data availability, we do not alter any of the findings for the covariates that **?** hypothesize are related to party affective polarization.

Table C-1: Verification and extension of Reiljan et al. (2023)

	Table 2, column 1 Party affective polarization (PAP)		
	Original estimate (1)	Adjust covariates (2)	Adjust sample (3)
Party id	1.524*	1.232*	1.310*
	(0.412)	(0.437)	(0.426)
L-R polarization	0.184*	0.251*	0.282*
	(0.080)	(0.062)	(0.062)
Effective # parties	-0.149*	-0.167*	-0.153*
	(0.039)	(0.041)	(0.043)
Presidential	-0.899*	-0.700*	-0.826*
	(0.202)	(0.219)	(0.179)
Government effectiveness	-0.762*		
	(0.130)		
Public sector corruption index		1.533*	1.502*
		(0.573)	(0.551)
GDP pc (log)		-0.521*	-0.537*
		(0.174)	(0.185)
Year	0.025*	0.036*	0.034*
	(0.008)	(0.009)	(0.009)
(Intercept)	6.134*	8.246*	8.205*
	(0.642)	(1.763)	(1.852)
N × T	102	93	100
# Countries	40	37	39

* indicates statistical significance at the 0.05 level. Standard errors clustered on country.

Tables C-2 and C-3 show results from robustness tests for the Affective polarization index (C-2) and Out-party affect (C-2). Each table highlights two rows. First, we highlight the estimates for interaction between an attack on democracy and ruling party personalism, as this variable is the test of the main theoretical prediction. Second, we highlight a row at the bottom of each table, which reports the estimate for $\beta_{Attack|HighPers}$. This estimate is the linear combination of the treatment plus the interaction when the moderator has a high value (one standard deviation above the sample mean). Substantively, this is the estimated marginal effect of an attack on democracy when ruling party personalism is high.

The first column in each table re-reports the results from the main text, Table 2 columns (2) and (6). These tests measure polarization and out-party affect among partisans and the referent in the *party.* Next, in each table, we alter the outcome variable such that outcomes polarization and out-party affect with the *leader* as the referent. In both cases– polarization in Table C-2 and out-party affect in Table C-3 – the result in column (1) is slight stronger than the result reported in the main text. This makes sense because leaders may be more polarizing that party affect.

The next two columns – (2) and (3) – in Tables C-2 and C-3 show results when we measure the outcomes among all voters – not just among partisans. In both cases, the results are weaker. Again, this makes sense because the polarizing effect of government attacks on democracy (amplified by party personalism) should be weaker among self-identified non-partisans than among partisans; and measuring the outcomes among all voters combines both of these groups together.

The next three columns – (4) - (6) – in Tables C-2 and C-3 show results when we omit covariates from the specification. First we drop all covariates in (4). Next we drop covariates in the specification offered by **?** in (5). Last, we drop covariates related to selection into ruling party personalism: initial democracy level when the leader is selected into power and the age of the democracy, in (6). Substantively, the main results for the interaction between attacks on democracy and ruling party personalism remain in columns (4) and (5). However in (6), the main results point in the correct direction but are weaker. Even in these tests in column (6) – which omit variables related to selection into ruling party personalism – the effect of an attack on democracy boosts polarization and reduces out-party affect remains statistically significant when ruling party personalism high but not when it is low.

The last four columns in each table – (7) - (10) – add a variables, one at a time, to the main specification. Theses variables help us isolate the effect of ruling party personalism from other party-related concepts. We test four: initial level of *polarization* in society in the year in which each leader is selected into power; initial level of *judicial independence*; initial level of *party system institutionalization*; and ruling party *populism.* Adding these to the specification does not alter the main results and, in fact, makes them slightly stronger in some cases.

Figure C-1 shows the results for in-party and out-party affect when we include *Government effectiveness* in the specification. The main results hold: attacks on democracy decrease out-party affect when the ruling party is more personalism.

Table C-2: Affective polarization index

	Party polariz. among partisans	Leader polariz. among partisans	Party polariz. among voters	Leader polariz. among voters	Party polarization among partisans, columns (5)-(11)						
	Reported	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Attack on judiciary	-1.209	-1.031	-0.242	0.326	0.313	0.086	-0.691	-0.656	-1.560	-1.386	-1.153
	(1.265)	(1.483)	(1.150)	(1.340)	(1.275)	(1.295)	(1.189)	(1.574)	(1.153)	(1.449)	(1.215)
Ruling party persl.	-1.836	-2.009	-0.607	-0.228	-2.209	-3.198*	-1.020	-2.077	-2.227	-1.982	-1.836
	(1.209)	(1.388)	(1.107)	(1.128)	(1.098)	(1.461)	(0.939)	(1.624)	(1.135)	(1.228)	(1.156)
Attack × pers.	6.177	7.171	2.228	1.898	6.687*	8.376*	4.038	6.153	7.094*	6.538*	6.289*
	(3.153)	(3.745)	(2.517)	(2.729)	(2.703)	(3.254)	(2.198)	(4.196)	(3.018)	(3.213)	(2.934)
Initial democracy	3.287	5.915	0.611	2.383		2.834		3.059	2.424	3.397	6.069*
	(3.196)	(4.559)	(1.984)	(2.755)		(2.468)		(3.527)	(3.031)	(3.149)	(2.675)
Democracy age	0.006	0.049	0.030	0.082		-0.106		-0.001	-0.013	0.016	-0.022
	(0.058)	(0.106)	(0.061)	(0.087)		(0.098)		(0.059)	(0.064)	(0.061)	(0.053)
Time trend	0.034*	0.039*	0.031*	0.031*			0.032*	0.033*	0.037*	0.034*	0.030*
	(0.012)	(0.013)	(0.010)	(0.011)			(0.013)	(0.013)	(0.013)	(0.013)	(0.012)
Party id	0.780	0.581	1.310*	1.506*			0.864	0.768	0.187	0.792	0.836
	(0.521)	(0.647)	(0.438)	(0.490)			(0.562)	(0.535)	(0.578)	(0.548)	(0.487)
L-R polarization	0.162	0.132	0.242*	0.223*			0.206*	0.165	0.172*	0.184	0.129
	(0.084)	(0.109)	(0.067)	(0.089)			(0.064)	(0.097)	(0.081)	(0.095)	(0.069)
Effective # parties	-0.178*	-0.227*	-0.152*	-0.180*			-0.182*	-0.186*	-0.202*	-0.184*	-0.177*
	(0.059)	(0.081)	(0.049)	(0.061)			(0.062)	(0.073)	(0.060)	(0.074)	(0.081)
Presidential	-0.487*	-0.333	-0.811*	-0.649			-0.577*	-0.482	-0.369	-0.472*	-0.417*
	(0.209)	(0.447)	(0.188)	(0.352)			(0.233)	(0.241)	(0.206)	(0.204)	(0.168)
Public sector corrupt.	1.058	1.887	1.386	2.036			0.518	0.841	1.353	0.991	1.336
	(0.852)	(1.310)	(0.867)	(1.129)			(0.821)	(0.821)	(0.823)	(0.888)	(0.941)
GDP pc (log)	-0.713*	-0.780	-0.496	-0.535			-0.649	-0.674	-0.690	-0.652	-0.504
	(0.340)	(0.452)	(0.259)	(0.316)			(0.363)	(0.347)	(0.343)	(0.402)	(0.331)
Polarization$_{t=0}$								-0.056			
								(0.070)			
Judical indep.$_{t=0}$									0.276*		
									(0.082)		
Party system inst.$_{t=0}$										-0.737	
										(1.375)	
Ruling party popul.											0.520
											(0.452)
(Intercept)	8.521*	6.341	7.347	5.280	4.238*	2.266	10.422*	8.247	8.899*	8.433*	3.935
	(4.115)	(5.921)	(3.686)	(4.794)	(0.361)	(2.111)	(4.014)	(4.087)	(4.065)	(4.086)	(3.863)
$\beta_{Attack\|LowPers}$	0.26	0.40	0.20*	0.71	1.65	1.76*	0.12	0.57	-0.14	-0.08	0.10
	(0.77)	(0.95)	(0.75)	(0.93)	(0.84)	(0.78)	(0.81)	(0.87)	(0.67)	(0.96)	(0.77)
$\beta_{Attack\|HighPers}$	2.50*	3.27*	1.10	1.46	4.32*	5.11*	1.73*	3.04*	2.70*	2.53*	2.62*
	(0.98)	(1.29)	(0.75)	(0.87)	(0.77)	(1.00)	(0.56)	(1.24)	(0.94)	(1.01)	(0.93)
Elections	82	82	100	100	84	84	82	77	80	80	77
Countries	36	36	39	39	36	36	36	35	36	36	33

[*] indicates statistical significance at the 0.05 level. Standard errors clustered on country. Estimates for β_{Attack} at High and Low ruling party Personalism set personalism as one standard deviation below the Personalism mean (0.2) and one standard deviation about the mean (0.6).
$\beta_{Attack|LowPers} \equiv \beta_{Attack} + (\beta_{Attack \times personalism} \times 0.2)$; $\beta_{Attack|HighPers} \equiv \beta_{Attack} + (\beta_{Attack \times personalism} \times 0.6)$.

Table C-3: Out-party affect

	Party affect among partisans	Leader affect among partisans	Party affect among voters	Leader affect among voters			Out-party affect among partisans, columns (5)-(11)				
	Reported	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Attack on judiciary	0.519	1.363	-0.533	-0.324	-1.124	-1.068	0.060	0.173	0.266	1.002	0.658
	(0.896)	(1.323)	(0.805)	(1.253)	(1.096)	(1.149)	(0.787)	(1.192)	(0.873)	(0.881)	(1.016)
Ruling party pers.	2.425*	2.594*	1.489	1.245	1.135	1.828	1.349*	2.418	2.332*	2.497*	2.654*
	(0.906)	(1.195)	(0.768)	(1.014)	(0.902)	(1.165)	(0.592)	(1.329)	(0.918)	(0.844)	(0.844)
Attack × Pers.	-6.787*	-8.427*	-4.064*	-4.267	-3.669	-5.076	-4.398*	-6.356	-6.388*	-7.322*	-7.543*
	(2.308)	(3.109)	(1.818)	(2.541)	(2.250)	(2.571)	(1.255)	(3.379)	(2.333)	(2.072)	(2.031)
Initial democracy	-3.675	-6.634	-2.003	-3.895		-2.693		-3.374	-4.198	-4.126	-6.332*
	(2.536)	(3.612)	(1.618)	(2.259)		(1.867)		(2.866)	(2.551)	(2.287)	(1.457)
Democracy age	0.070	0.013	0.073	0.007		0.046		0.079	0.036	0.079	0.091
	(0.075)	(0.095)	(0.071)	(0.082)		(0.087)		(0.077)	(0.066)	(0.078)	(0.079)
Time trend	-0.029*	-0.038*	-0.028*	-0.034*			-0.027*	-0.027*	-0.032*	-0.024	-0.027*
	(0.012)	(0.014)	(0.010)	(0.012)			(0.013)	(0.013)	(0.012)	(0.013)	(0.011)
Party id	0.242	0.291	0.215	0.018			0.217	0.273	-0.235	0.300	0.155
	(0.463)	(0.621)	(0.411)	(0.534)			(0.505)	(0.445)	(0.466)	(0.453)	(0.441)
L-R polarization	-0.065	-0.041	-0.067	-0.056			-0.119*	-0.077	-0.062	-0.116	-0.028
	(0.065)	(0.097)	(0.056)	(0.085)			(0.045)	(0.078)	(0.064)	(0.074)	(0.044)
Effective # parties	0.272*	0.334*	0.260*	0.306*			0.267*	0.277*	0.247*	0.303*	0.279*
	(0.050)	(0.069)	(0.040)	(0.055)			(0.057)	(0.061)	(0.050)	(0.056)	(0.066)
Presidential	0.114	0.109	0.419*	0.430			0.175	0.141	0.166	0.130	0.047
	(0.188)	(0.315)	(0.201)	(0.294)			(0.220)	(0.208)	(0.197)	(0.193)	(0.157)
Public sector corrupt.	0.099	-1.047	0.417	-0.460			0.559	0.256	0.445	0.056	-0.092
	(0.623)	(1.121)	(0.602)	(0.993)			(0.662)	(0.661)	(0.692)	(0.660)	(0.738)
GDP pc (log)	0.428	0.446	0.470	0.533			0.462	0.396	0.518*	0.134	0.300
	(0.270)	(0.348)	(0.242)	(0.306)			(0.290)	(0.290)	(0.252)	(0.343)	(0.279)
Polarization$_{t=0}$								0.003			
								(0.069)			
Judical indep.$_{t=0}$									0.197*		
									(0.069)		
Party system inst.$_{t=0}$										2.010	
										(1.221)	
Populism											-0.285
											(0.381)
(Intercept)	1.350	3.786	-0.159	1.235	4.078*	6.204*	-1.628	1.411	0.939	2.767	4.863
	(3.339)	(5.163)	(2.988)	(4.398)	(0.324)	(1.625)	(3.107)	(3.396)	(3.315)	(3.371)	(3.274)
$\beta_{Attack\|LowPers}$	-0.84	-0.32	-1.35*	-1.18	-1.86*	-2.08*	-0.82	-1.10	-1.01	-0.46	-0.85
	(0.57)	(0.91)	(0.55)	(0.86)	(0.72)	(0.74)	(0.60)	(0.69)	(0.53)	(0.61)	(0.67)
$\beta_{Attack\|HighPers}$	-3.55*	-3.69*	-2.97*	-2.88*	-3.33*	-4.11*	-2.58*	-3.64*	-3.57*	-3.39*	-3.87*
	(0.80)	(1.13)	(0.62)	(0.79)	(0.60)	(0.76)	(0.47)	(1.14)	(0.79)	(0.77)	(0.51)
Elections	82	82	100	100	84	84	82	77	80	80	77
Countries	36	36	39	39	36	36	36	35	36	36	33

* indicates statistical significance at the 0.05 level. Standard errors clustered on country. Estimates for β_{Attack} at High and Low ruling party Personalism set personalism as one standard deviation below the Personalism mean (0.2) and one standard deviation about the mean (0.6).
$\beta_{Attack|LowPers} \equiv \beta_{Attack} + (\beta_{Attack \times personalism} \times 0.2)$; $\beta_{Attack|HighPers} \equiv \beta_{Attack} + (\beta_{Attack \times personalism} \times 0.6)$.

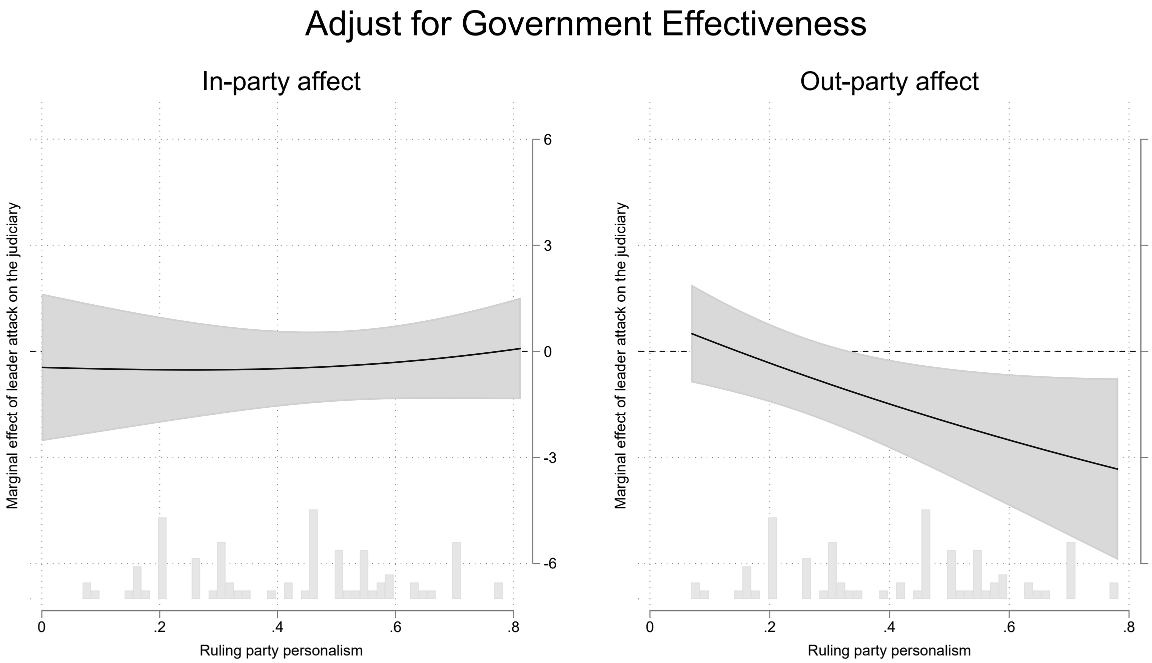


Figure C-1: Adjusting for Government Effectiveness in tests of government attacks on democracy and affective polarization

Appendix D: Macro-polarization analysis

In Table 3 in the main text, we reported results from a test of macro-polarzation with a specification that included an interaction term between *Attacks on democracy* and ruling party *Personalism*. The linear model in those regression assumes a linear interaction. In Figure 4 we reported the substantive effect of the interaction terms with a kernel regression estimator that relaxes the linear interaction assumption for the dynamic panel model (i.e. two-way FE + lagged outcome). Here, in Figure D-1 we report the binning estimates for the interactive effect for all four models reported in Figure 3. While the interaction effect is not perfectly linear, the monotonically increasing marginal effect of *Attacks* as *Personalism* increases is consistent with the main hypothesis.

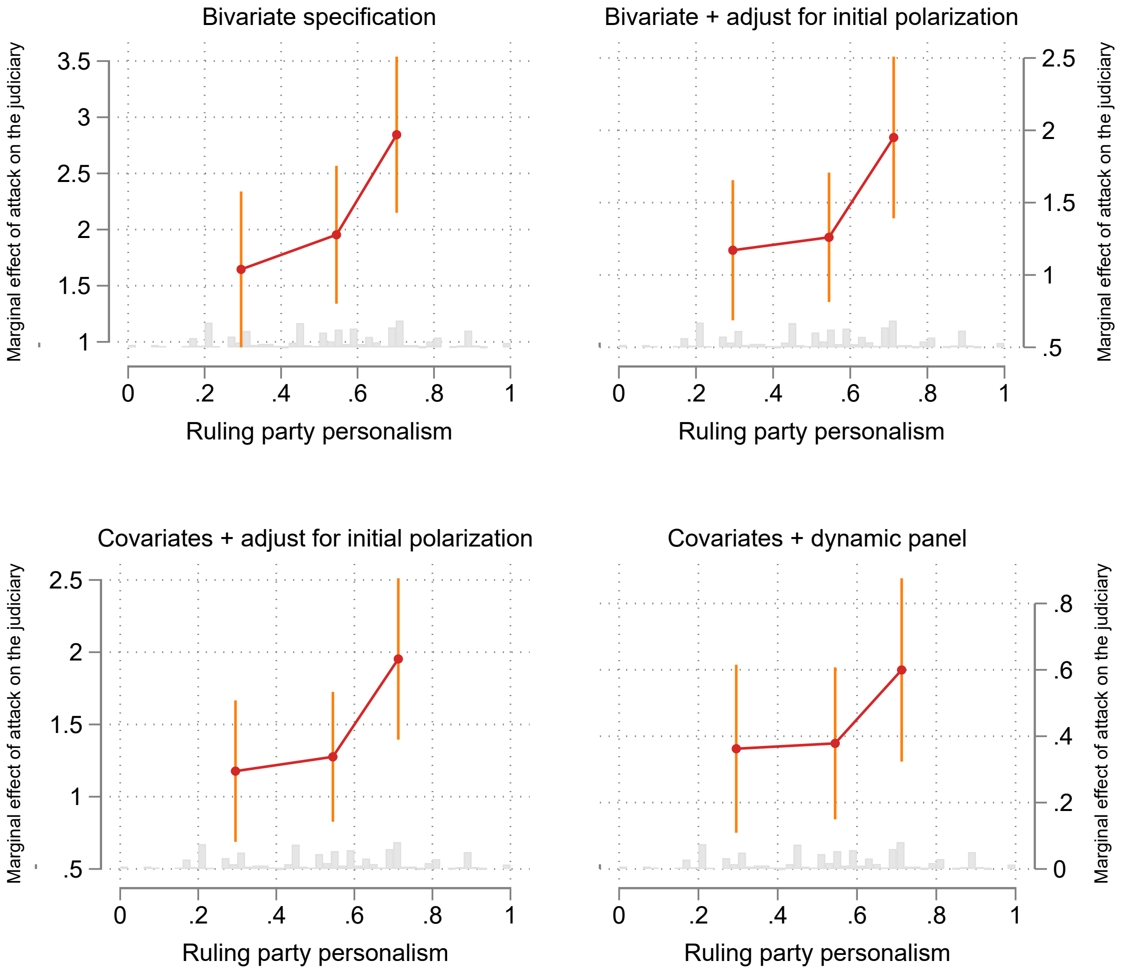


Figure D-1: Government attacks on the judiciary boost polarization when the ruling party is personalist, binning estimates

Table D-1 reports results from additional tests of macro-polarization: models with additional lags of the outcome variable and interactive fixed effect models. Testing models with additional outcome lags ensures that the model captures any prior trends – potentially not captured in the one-year lag – in the outcome that might cause selection into the treatment. The interactive fixed effects models are an extension of two-way fixed effects models. The standard 2-way FE models all time-invariant heterogeniety between countries and all common time trends. The IFE, in addition, allows for the possibility that time shocks that have heterogeneous – rather than a common – effect on different units (Bai, 2009). For example, the 2008 Great Recession may have produced different shocks to long-lived democracies than in new democracies.

The results in Table D-1 show that the main results hold. The theoretical interaction term in column (2) is positive and signficant, while the split sample analysis in (4) and (5) shows that attacks on democracy only boost polarization when ruling party personalism is high.[28]

Table D-1: Macro-polarization, additional lags and IFE

	Additional lags (1)-(2)		Interactive FE (3)-(5)		
	No interaction	Interaction	Full sample	Low ruling party personalism	High ruling party personalism
	(1)	(2)	(3)	(4)	(5)
Attack on democracy	0.488*	0.233	0.477*	0.080	0.666*
	(0.104)	(0.155)	(0.105)	(0.142)	(0.189)
Attack × personalism		0.414*			
		(0.186)			
Ruling party personalism	0.051	-0.126	0.028	0.033	0.088
	(0.033)	(0.081)	(0.033)	(0.041)	(0.109)
Judicial independence$_{t=0}$	0.079	0.085	0.057	0.078	0.056
	(0.090)	(0.090)	(0.088)	(0.126)	(0.146)
Democracy level$_{t=0}$	-0.006	-0.005	-0.004	-0.011	-0.006
	(0.010)	(0.011)	(0.009)	(0.017)	(0.020)
Democracy age	0.019	0.019	0.017	0.050*	-0.015
	(0.013)	(0.013)	(0.012)	(0.019)	(0.026)
Election year	0.034*	0.034*	0.036*	0.031*	0.038*
	(0.009)	(0.009)	(0.009)	(0.011)	(0.015)
Polarization$_{t-1}$	0.863*	0.860*	0.847*	0.882*	0.768*
	(0.041)	(0.041)	(0.023)	(0.022)	(0.047)
Polarization$_{t-2}$	-0.021	-0.021			
	(0.024)	(0.024)			
Polarization$_{t-3}$	-0.008	-0.007			
	(0.022)	(0.022)			
Polarization$_{t-4}$	0.010	0.010			
	(0.018)	(0.018)			
(Intercept)	-0.316*	-0.222*	-0.281*	-0.226	-0.274
	(0.104)	(0.112)	(0.098)	(0.152)	(0.181)
N × T	2285	2285	2302	1077	1222
# Leaders	564	564	567	259	304
# Countries			100	73	84
Country FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Additional lags	✓	✓			
IFE			✓	✓	✓

* indicates statistical significance at the 0.05 level. Standard errors clustered on leader.

[28] For the split-sample analysis, which is akin to binning estimates, we split the sample of leaders at the median value of ruling party personalism.

Figure D-2 reports estimates for the interaction term of interest (*Attacks* × *Personalism*) for specifications that add covariates, one at a time. The left plot shows the estimate for *Attacks* × *Personalism* when we add covariates: ruling party Populism, Presidential system, Party system institutionalization in the year the leader is selected into power, judicial independence in the year the leader is selected into power, polarization in the year the leader is selected into power, a measure of democratic consolidation that combines information on democracy age and democracy level in the year the leader is selected into power; and ruling party seat share. The vertical line just below 0.40 is the estimate of the interaction term from the dynamic panel model reported in column (8) of Table 3. We want to see how interaction estimate changes from the main reported estimate as we add covariates.

The right plot reports similar estimates for the main interaction term. However, in these specification we add both the additional covariate and an interaction between the treatment (*Attacks*) and the covariate. This ensures that the interaction effect we test is not simply picking up the interaction between the added covariate and the treatment.

In all changes to the specification, we find that estimate for the key interaction term is either the same or larger than the estimate of the interaction term from the dynamic panel model reported in column (8) of Table 3.

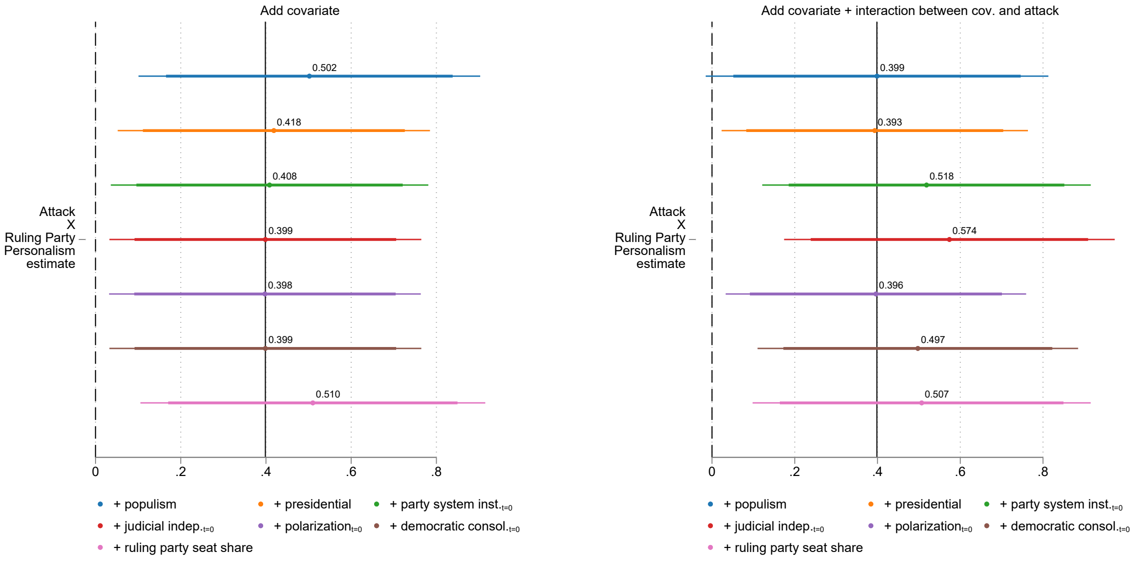


Figure D-2: Interaction term estimates, added covariates

A dynamic treatment effects approach corrects biases induced by treatment effect heterogeneity. We estimate a two-way fixed effects counterfactual model using software provided by Liu, Wang and Xu (2021). The specificiation adjusts for: Polarization$_{t=0}$; Democracy level$_{t=0}$; Judicial independence$_{t=0}$; Election year and Democracy age. By including Polarization$_{t=0}$ in the specification, the outcome is transformed into the extent to which polarization has *changed* since the leader was first selected into power.

The *attacks on democracy* variable is continuous but the treatment effects estimators require a time-varying binary treatment variable. So we dichotomize the treatment variable at its median; this yields 42 of 102 units (41 percent) with time-varying treatment status.[29] To test the interaction effect, we then split the sample into two bins, one for leaders with ruling party personalism at or below the median value and another bin for leaders with party personalism values higher than the median. 42 percent of units in the high personalism bin have time-varying treatment status while 40 percent of units do in the low personalism bin.

Figure D-4 reports the treatment estimate for each of the two bins, high- and low-ruling party personalism. For high personalism the average treatment effect on the treated is just over 0.3 and statistically signficant. For low ruling party personalism cases, the ATT estimate is 0.1. This suggests that the ATT estimate is three times larger in when ruling party personalism is high relative to low.

The latter two estimates shown in Figure D-4 are the estimates for placebo tests. These tests use six pre-treatment periods as a placebo to see if the pre-treatment ATT in this range is statistically significantly different from zero. Both of these estimates are close to zero and *not* statistically significant. This suggests that pre-treatment trends in the outcome are *not* causing selection into treatment.

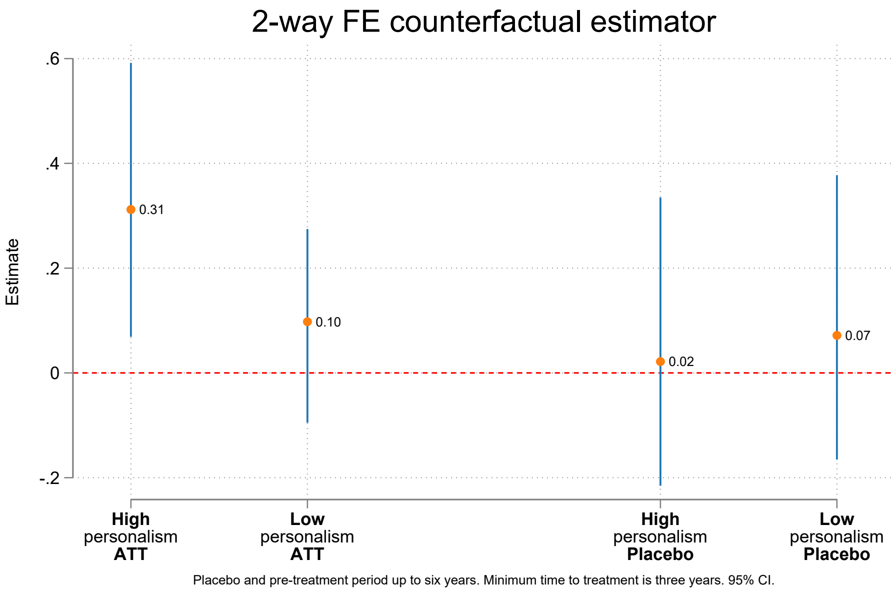


Figure D-3: Counter-factual fixed effects estimates, high and low ruling party personalism

[29] In these tests we set the minimum period that a unit must under the control condition to three years to be included in the analysis.

Finally, Figure D-4 shows the pre-treatment ATT's and the post-treatment ATT's for the bin of high ruling party personalism cases – for eight pre- and post-periods, respectively. In the pre-treatment periods, the ATT's are not consistently pointing in one direction and none are statistically different from zero. In contrast, the post-treatment ATT's are all positive and most are significant.

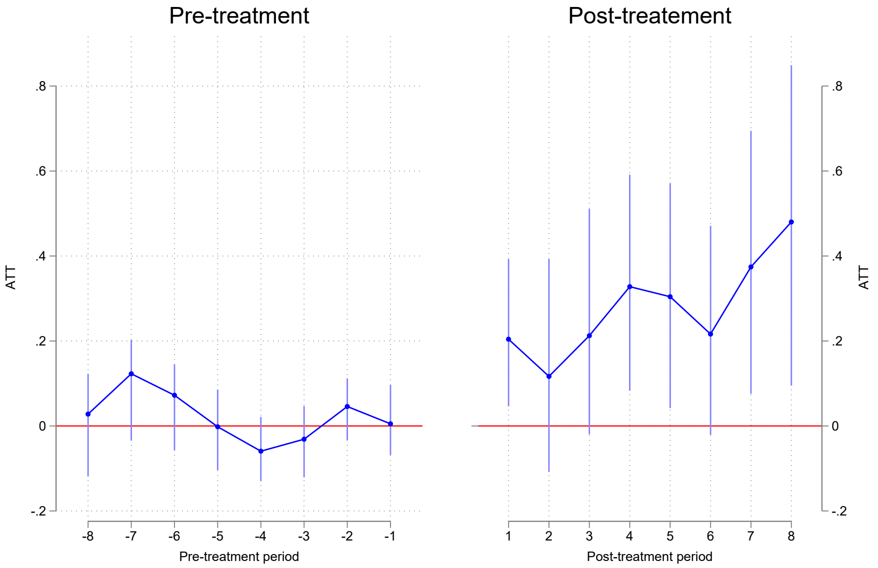


Figure D-4: Counter-factual fixed effects estimates, high ruling party personalism