Social Desirability and Affective Polarization*

Elizabeth C. Connors, Assistant Professor, University of South Carolina

Abstract. Media coverage of affective polarization—partisans disliking and distrusting out-partisans while liking and trusting in-partisans—is abundant, both creating and reflecting a belief among the public that partisans are more affectively polarized than they are. These trends suggest that affective polarization among partisans could be viewed as socially desirable, which may then shape partisans' expressed attitudes and behavior. To examine this, I run four original surveys and study two broad research questions: 1) does this social desirability exist; and 2) can it influence partisans' expressed affective polarization. I find that affective polarization among partisans is indeed socially desirable and that, largely motivated by self-presentation desires, this social desirability can shape partisans' expressed affective polarization. However, my results also suggest that affective polarization responses are rather ingrained in partisans, and that while partisans are aware of this social desirability and its effect on their behavior, small changes in survey context do not necessarily produce large changes in affective polarization responses. Overall, the results offer necessary nuance to our understanding of affective polarization, implying that social desirability—which can be shifted by contexts—can alter how affectively polarized people act.

Key Words: social desirability; social influence; affective polarization

**Paper accepted at Public Opinion Quarterly*

Research was partly funded by Time-Sharing Experiments for the Social Sciences (TESS) and was partly internally funded. It was conducted under the approval of the University of South Carolina Institutional Review Board. Previous versions of this manuscript have received helpful feedback from Yanna Krupnikov, John Barry Ryan, Jennifer Jerit, Peter DeScioli, Dave Darmofal, Matthew Levendusky, participants and discussants at MPSA 2019, the Toronto Political Behavior Workshop 2019, APSA 2019 and 2020, SPSA 2020, Hot Politics 2021, Gothenburg University's Party Research Seminar 2022, and various anonymous reviewers. Replication data and documentation are available at https://doi.org/10.7910/DVN/2WHXXI.

Certain trends suggest that affective polarization—partisans disliking and distrusting out-partisans while liking and trusting in-partisans[1]—could be socially desirable among partisans[2] and that this could shape expressed[3] affective polarization. In particular, the rise in media coverage and overestimation of affective polarization by the public suggest that partisans could believe that to impress other partisans they should act affectively polarized, leading some partisans to do so. This would suggest that certain contexts could shape affective polarization reports and behavior, implying a social dimension to polarization that is currently missing from the literature. In this piece, I thus test if partisans believe affective polarization is socially desirable and if this can influence their expressed affective polarization.

Affective Polarization

Affective polarization[4] (heretofore: polarization) concerns researchers and practitioners alike. Polarization has been found to impact our social lives, lower interpersonal trust (Lee 2022), and lead to consequential life choices such as who to talk to (Mutz 2002; Barbera 2014), where to live (Bishop 2008), who to marry (Iyengar, Konitzer, and Tedin 2018), and preferences about who one's child should marry (Iyengar and Westwood 2015; see Iyengar et al. 2019 for review). Although researchers have found somewhat limited *political* ramifications of polarization, findings do suggest that it can influence political beliefs (Druckman et al. 2021), participation (Iyengar and Krupenkin 2018), and trust (Hetherington and Rudolph 2015), as well as lead to anti-democratic attitudes (Kingzette et al. 2021; but see Broockman, Kalla, and Westwood 2022 and Voelkel et al. ND) and out-party dehumanization (Cassese 2021; Martherus et al. 2021).

While researchers are hesitant to draw conclusions about how dangerous polarization is, the media does not follow this same hesitancy. In fact, media coverage of polarization generally has increased dramatically since the 1990s (Levendusky 2009; see also Klar and Krupnikov 2016 and Robison and Mullinix 2016). As Levendusky and Malhotra (2016b) explain, "The mass media depict polarization as widespread, occurring across many issues, and accompanied by incivility and dislike of the opposition" (pg. 286).

This focus on polarization by researchers and the media is a natural outcome of the steep rise in polarization over the past 40 years (Iyengar et al. 2019), which is believed to be due to a myriad of trends, including the increasing importance of partisanship as a social identity (e.g., Tajfel and Turner 1979; Green, Palmquist, and Schickler 2002; Dias and Lelkes 2021; but see West and Iyengar 2022). Other explanations range from similar identity explanations—such as increased sorting along ideological (Levendusky 2009) and demographic (Mason 2018) lines—to changing issue attitudes (Webster and Abramowitz 2017; Orr and Huber 2020), media coverage of elites (Huddy and Yair 2021) and partisans (Levendusky and Malhotra 2016b), and polarized networks (Butters and Hare 2022).

[1] Affective polarization refers to the gap between feelings toward in-party and out-party members. Thus, when I reference people's reports of polarization I am referring to the gap between their reports of feelings toward in-party versus out-party members.

[2] Given my focus on polarization, my theory and empirics are limited to partisans.

[3] I use "expressed" to refer to reported attitudes and behavior.

[4] I narrowly focus on affective polarization, although there is an extensive literature on other polarization types (some of which co-occur with affective polarization).

For my purposes, the reason for this rise is tangential to the fact that it has led to a slew of media coverage on polarization.[5] Importantly, this coverage coincides with the public's exaggerated perceptions of polarization among the electorate (Ahler 2014; Levendusky and Malhotra 2016a; Ahler and Sood 2018). Druckman et al. (2021), for example, find that people (wrongly) think the modal partisan is extreme (see also Krupnikov and Ryan 2022). This co-occurrence of media coverage and polarization misperceptions may be because the media reflects these perceptions, but also because they *guide* these perceptions (Ross and Dumetrescu 2019; see also Arias 2018). Indeed, Levendusky and Malhotra (2016b) find that media coverage leads people to believe the public is more polarized than they actually are.

Social Desirability

Thus, the media's narrative and public perceptions imply that polarization is viewed as commonplace among partisans (i.e. a descriptive norm), suggesting that it could also be socially desirable among partisans (i.e. an injunctive norm), as these often (but not always) work together (Cialdini, Reno, and Kallgren 1990; Rimal and Real 2003, 2005; Karp and Brockington 2005). Descriptive norms refer to what people *do* and injunctive norms refer to what people *should do*—or what is socially desirable (Cialdini et al. 1990; Gerber and Rogers 2009). A descriptive norm could be that most people read books, while an injunctive norm could be that people *should* read books. In the context of my research, people viewing the typical partisan as polarized is a descriptive norm—one that past research demonstrates—and people viewing a *good* partisan as polarized is an injunctive norm—one that I examine here (and, since people often use "norm" to refer to descriptive norms, I rely on the term social desirability: the belief that an attitude or behavior is seen as positive by others). As Iyengar et al. (2019) explain, "the rhetoric and actions of political leaders demonstrate that hostility directed at the opposition is acceptable and often appropriate" (pg. 133).[6] This thus leads to my first hypothesis:

Hypothesis 1: Partisans believe it is socially desirable for partisans to be polarized.

Importantly, social desirability can shape reported attitudes and behavior because people have an innate desire to impress others (Goffman 1955, 1967; Cosmides and Tooby 1992) and work at "self-presentation" almost constantly (Holtgraves 1992, 2004). The tendency to misrepresent oneself to others based on what is socially desirable varies by individual and can be measured with self-monitoring (Snyder 1974, 1979; Gangestad and Snyder 2000; Berinsky 2004; Berinsky and Lavine 2007, 2012).[7] Self-monitoring ranges from low to high, where higher levels indicate greater

[5] Further, I cannot differentiate how much of this rise is due to "true" polarization versus socially desirable responding. I do, however, believe that the former has increased and that this has led to the latter. Nonetheless, as much work has been done on the former aspect, my aim is to highlight this latter aspect.

[6] It is noteworthy that partisans likely receive conflicting cues here: the media seem to portray polarization as widespread but negative, while elites (and potentially peers) portray it as desirable. Thus, Hypothesis 1 examines which of these cues partisans internalize: polarization's desirability or *un*desirability.

[7] Other measures can also assess one's likelihood of socially desirable responding, but incorporate one's likelihood of doing so based on *self-deception* in addition to *impression management* (Paulhus 1991; Paulhus et al. 2003). Using self-monitoring—which measures one's likelihood of socially desirable responding based on *impression management* desires specifically—is more suitable for this particular

tendency to change oneself to appease others. Note that self-monitoring does not necessarily measure *perceptions* of social desirability—both low and high self-monitors can perceive desirable behavior, but the latter will respond by acquiescing and the former by potentially doubling down on their "authentic selves" (Premeaux and Bedeian 2003; see also Banaji and Prentice 1994). These differential reactions indicate social desirability motivated by impression management, where high self-monitor behavior demonstrates what is socially desirable and low self-monitor behavior demonstrates what is socially *un*desirable.

Social desirability is important to research in political behavior because social motivations can shape how ordinary people practice politics (Conover and Searing 2005). The social is, in essence, in the political (Sinclair 2012), with some (perhaps most) prioritizing social over political motivations (Walsh 2004; see also Carlson and Settle 2022 and Krupnikov and Ryan 2022). Thus, to present themselves well and adhere to social desirability, people will misreport political attitudes (Zaller and Feldman 1992; Kuran 1997; Daoust et al. 2021) and political behavior (Karp and Brockington 2005), suppress unpopular or contentious opinions (e.g., Berinsky 1999; Feldman and Huddy 2005; Weber et al. 2014; Carlson and Settle 2016), and change how they identify (Klar and Krupnikov 2016) or who they support (Perez-Truglia and Gruces 2017) politically.

People do not just have a general desire to impress others, however. They also have a desire to fit in with their in-group and receive its approval (Achen and Bartels 2017; Conover, Searing, and Crewe 2022). Thus, people will adopt in-group members' attitudes (Douglas and Wildavsky 1982; Kahan, Jenkins-Smith, and Braman 2011) and comply with what in-group members say and do (Turner, Brown, and Tajfel 1979).[8] These tendencies are often attributable to the fact that people become attached to their group and view it as part of their identity—something known as social identity theory (Tajfel and Turner 1979; see also Ellemers, Spears, and Doosje 2002 and Miller, Brewer, and Arbuckle 2009). Importantly, this perspective has been applied to partisanship (Green et al. 2002; see Huddy 2013 for review)—thus, just as prototypic members of social groups can shape norms and behaviors (Terry and Hogg 1996), so too can prototypic Democrats and Republicans (Hogg 2001). Indeed, research finds that partisans want to seem like typical (Toff and Suhay 2019) or good (Connors 2020) partisans, and will thus alter expressions to do so (Bakker, Lelkes, and Malka 2021; Fieldhouse, Cutts, and Bailey 2022).[9]

Together, this research demonstrates that social desirability can shape expressions and that prototypic group members can shape what is socially desirable—thus, what people believe prototypic group members do can lead in-group members to change what they (say they) do because it is viewed as socially desirable.[10] Therefore, since research demonstrates that people view the prototypic partisan as polarized (Druckman et al. 2021; Krupnikov and Ryan 2022), it is potentially the case that partisans view being polarized as socially desirable—Hypothesis 1, above—and that this social desirability can shape polarization expressions. Again, though, not

---

research, as my argument relates to outward-facing motives that can shape polarization expressions rather than the deception one can engage in to feel more psychologically content. Future research, however, could examine how self-deception motives shape polarization expressions.

[8] This can also occur because of a motivation to protect self-identity, but here I focus on the motivation to receive in-group approval.

[9] Note that a good Democrat may not be a good Republican—Connors (2020) finds that partisans adopt their own party's values because what is socially desirable differs by party. This is also an example of how perceptions of social desirability can vary by individual or group.

[10] It is not always clear if people are lying or actually changing expressions in response to social desirability, however.

everyone is equally likely to acquiesce to social desirability. Just as those higher in self-monitoring are more likely to adhere to social desirability in expressions of attitudes towards race (e.g., Feldman and Huddy 2005; Weber et al. 2014) and homosexuality (Berinsky 2004; Boysen, Vogel, and Madon 2006), partisanship (Klar and Krupnikov 2016), and values (Connors 2020), they should also be more likely to adhere to social desirability in expressions of polarization. This leads to my second hypothesis:

> **Hypothesis 2:** Social desirability can shape polarization expressions, and (when relevant) this effect is moderated by self-monitoring.

Finding support for Hypothesis 2 could help explain why telling people about high polarization increases polarization (Ahler 2014) and why children mimic their parents' polarization (Tyler and Iyengar 2022), as both could be driven by social motivations. It could also suggest why implicit measures of polarization are lower than explicit measures (Iyengar and Westwood 2015) and why polarization does not always have the outcomes we would expect (e.g. Broockman et al. 2022)—if social desirability shapes polarization expressions (and thus explicit measures of polarization), then this noise could cloud relationships between "true" polarization and certain outcomes.

Indeed, one outcome of polarization being socially desirable and this shaping expressions is that this could be reflected in survey research. Research finds that in certain contexts, people treat surveys as social interactions (Berinsky 2004), leading them to respond in socially desirable ways (Schuman, Presser, and Ludwig 1981; Sudman, Bradburn, and Schwarz 1996; Rigdon et al. 2009). This tendency can be exacerbated with reminders that researchers are "watching" (Haley and Fessler 2005), that anonymized data will be publicly available (Connors, Krupnikov, and Ryan 2019), and that they will receive survey feedback (Clifford and Jerit 2015). This research suggests that misrepresentation in survey responses is often driven by self-presentation desires. Thus, it is possible that some partisans are misrepresenting their polarization levels in surveys to adhere to social desirability. If this were the case, changing respondents' perceived survey privacy—similar to Connors et al. (2019)—should alter polarization responses. Again, though, self-monitoring should matter: the effect of privacy notifications should be moderated by self-monitoring. Indeed, this is what Connors et al. (2019) find.

However, it is also possible that privacy notifications do *not* change polarization responses. It is conceivable, for instance, that polarization responses are ingrained in partisans and thus less likely to shift based on small survey changes (see Druckman and Leeper 2012 for discussion of pre-treatment). This could especially be the case given that respondents do not know the researcher and thus may not feel the same social motivations they would in other contexts. This leads to my third, and last, set of hypotheses:

> **Hypothesis 3a:** Changing levels of perceived survey privacy can shape polarization responses, and this effect is moderated by self-monitoring.

> **Hypothesis 3b:** Changing levels of perceived survey privacy can*not* shape polarization responses.

Figure 1, below, illustrates the theory and hypotheses.

Figure 1. Illustration of Theory and Hypotheses

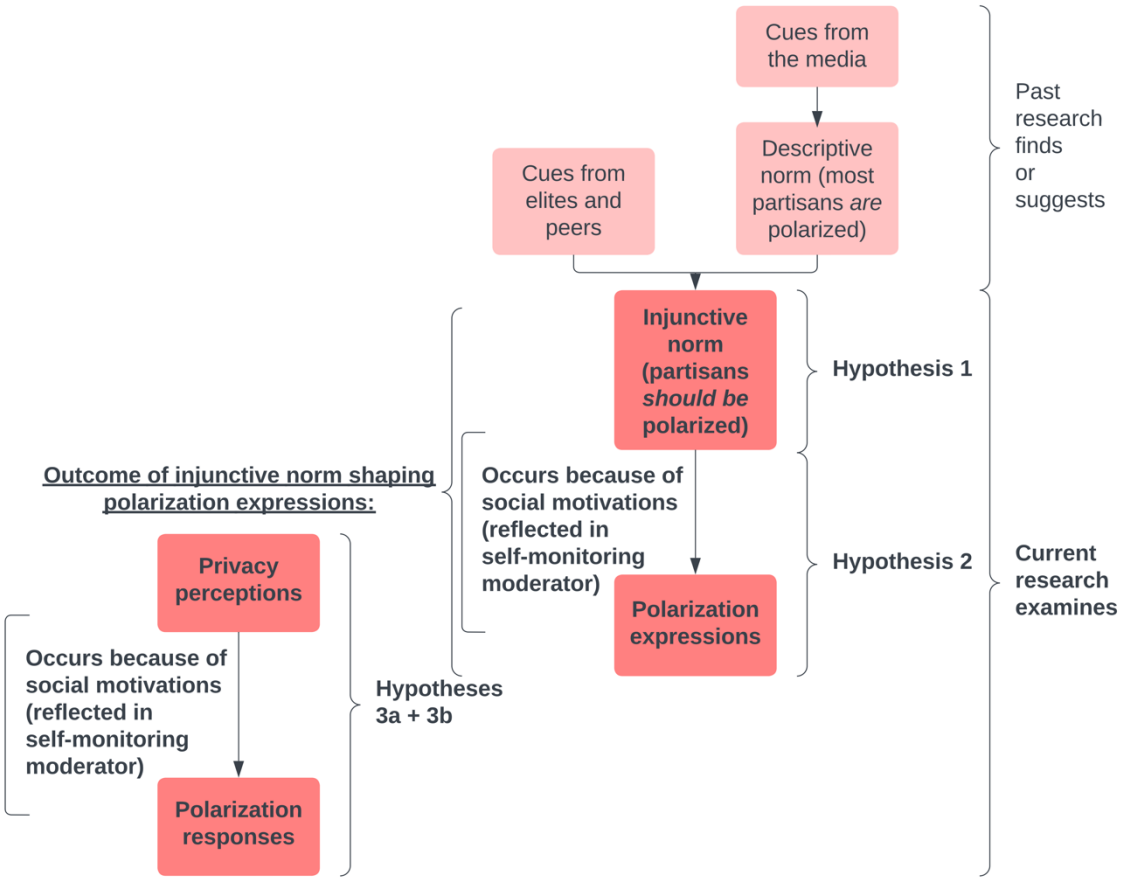


Chart illustrates the theory and hypotheses, where light pink illustrates ideas from past research and dark pink illustrates ideas examined in the current research.

Empirical Approach

Understanding the social desirability of polarization and how it shapes expressions is not a simple task. First, given nuances in how social motivations influence politics, understanding social effects with our current measures is challenging, if even possible (see Carlson and Settle 2022 for discussion). Second, many partisans—especially those "deeply involved" in politics (Krupnikov and Ryan 2022)—are likely truly and deeply polarized and thus also resistant to experimental treatments. Third, virtually everyone is already pre-treated with the information that suggests polarization is socially desirable—meaning that shifting the social desirability of polarization is difficult, if not impossible (Druckman and Leeper 2012). Thus, even those whose polarization expressions *are* susceptible to social desirability are pre-treated, making treatment effects within the context of an experiment less likely.

My empirical approach thus requires some creativity. First, to test Hypothesis 1—if polarization is socially desirable—I use a "fake good, fake bad" experiment (Holbrook, Green, and Krosnick 2003; see also Claassen and Ryan 2016 and Klar and Krupnikov 2016). This experiment randomly assigns participants to respond as someone wanting to *impress* versus *disappoint* others, where a gap between these two indicates that people perceive the *impress*

responses to be more socially desirable than the *disappoint* responses. This is a useful approach because research finds that people are often better at judging how others will behave than how they will (Epley and Dunning 2000; Vazire and Carlson 2011), implying these questions may be more insightful than self-reports are. Further, research finds that when people expect others to act a certain way, they *themselves* are more likely to act that way (Gerber and Rogers 2009)—suggesting that if people think partisans wanting to impress in-partisans will act polarized, they *too* are more likely to act polarized to impress in-partisans. Thus, in Study 1, I randomly assign participants to respond to polarization questions as a partisan wanting to impress or disappoint in-partisans. A gap between the two conditions would suggest social desirability—where higher polarization in the *impress* condition would indicate support for Hypothesis 1.

Second, to test Hypothesis 2—if social desirability can shape polarization expressions—I use a recall prompt that randomly assigns respondents to discuss a time when polarization was either socially *desirable* or *undesirable* (Study 2). I then ask respondents their reactions (including whether they pretended to agree and/or tried to fit in) and polarization questions. If social desirability can shape polarization expressions, a non-negligible proportion of respondents should say they engaged in this morphing behavior—especially because previous research shows that people will act like "chameleons" in these situations (Carlson and Settle 2016; see also Levitan and Verhulst 2016 and Fieldhouse et al. 2022). However, it is also likely that people underreport these reactions, as people may be unaware or uncomfortable admitting they changed their behavior based on social desirability (Carlson and Settle 2022).

I also conduct a more conservative test of Hypothesis 2 by interacting the treatment assignment with self-monitoring to predict respondents' polarization responses. Essentially, I treat the recall task as a treatment that could change polarization responses because respondents have just recalled a situation where it was socially desirable (undesirable) to be polarized. This could thus increase (decrease) reported polarization when moderated with self-monitoring. In particular, finding that *self-monitoring*desirable* has a positive effect on polarization reports and/or that *self-monitoring*undesirable* has a negative effect on polarization reports would demonstrate further support for Hypothesis 2. Thus, Study 2 uses recall tasks to examine the effect of social desirability both on reported behavior and survey responses.

Lastly, to test Hypotheses 3a and 3b—if changing levels of perceived survey privacy can shape polarization responses—in Studies 3 and 4 I randomly assign participants to be told that results based on their responses may be published (*public*), that responses are completely private (*private*), or nothing at all (*control*). Again, the effect of these reminders should be moderated by self-monitoring. Finding that *self-monitoring*public* has a positive effect on polarization reports and/or that *self-monitoring*private* has a negative effect on polarization reports would demonstrate support for Hypothesis 3a. This last test is the most conservative, as I use a subtle, one-line reminder to alter self-reports within the context of strong pre-treatment—this suggests that any effects found here are conservative estimates.

Data and Measurement

All of the following studies used US adult, unweighted, partisan samples (see Druckman and Levendusky 2019 for a similar partisan sampling approach), the demographics of which can be found in the Supplementary Material along with details about survey platforms and full questionnaires. In each study participants were given a consent form and no deception was used. When used, the control variables are the same across analyses: partisanship, partisan strength,

ideology, race, gender, age, and education. Lastly, the following studies rely on some of the same measures, including *self-monitoring* to measure likelihood of socially desirable responding, *feeling thermometers* and *trust* to measure polarization attitudes, and reported reactions to measure polarization behavior.

Self-monitoring asks respondents three questions about how they act in social situations—including if they are the center of attention and if they try to impress others—as well as how good of an actor they would be. Combining responses into one variable leads to a 13-point summary index. This measure originates from Snyder (1974) and Snyder and Gangestad (1986), but has been validated, shortened, and slightly edited by Gangestad and Snyder (2000), Berinsky (2004), and Berinsky and Lavine (2007, 2012). It has been used by political scientists, including the aforementioned work by Berinsky and colleagues as well as by Klar and Krupnikov (2016), Connors et al. (2019), and Connors (2020). For my particular studies, I have added Cronbach's alphas and correlates of self-monitoring to Supplementary Material F.

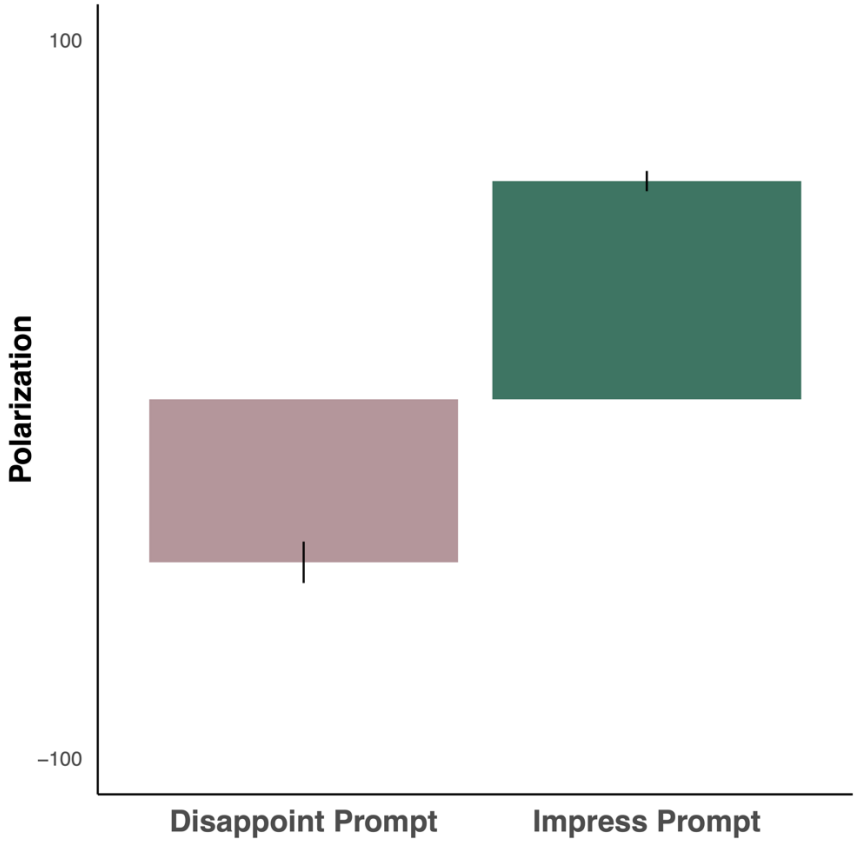
Feeling thermometers asks respondents to place each party (Democrats, Republicans) on a scale (0-100), where 50-100 indicates positive feelings and 0-50 indicates negative feelings. Polarization is then indicated by the difference between ratings of the in-party and out-party, where greater numbers indicate greater polarization. This is one of the most common measures of polarization (Iyengar et al. 2019). *Trust* is similar but asks about trust in both parties to do what is "right" for the country—polarization is then indicated by the same gap where greater numbers indicate greater polarization (Druckman and Levendusky 2019). Lastly, to measure reported behavior, I ask respondents if they would pretend to agree with people or try to fit in with others in situations when polarization was socially desirable or undesirable—this is aimed to capture morphing behavior. The goal of these three measures is to capture polarization expressions (i.e. reported attitudes and behavior).

Study 1: Is Polarization Socially Desirable? (Prolific, December 2020, N=920)

Design. Respondents from Prolific (N=920) were redirected to Qualtrics where they answered pre-treatment questions and then were randomly assigned to *impress* or *disappoint* conditions ("Please answer the following 2 questions as you think a [Republican/Democrat] wanting to [impress/disappoint] other [Republicans/Democrats] would")—the party was always in-party. Then, respondents were asked the *feeling thermometers* questions. Again, higher polarization in the *impress* condition would support Hypothesis 1—that partisans believe it is socially desirable to be polarized—while lower polarization in the *impress* condition would support the opposite—that they believe it is socially desirable to *not* be polarized.

Results. The results support Hypothesis 1 and show that partisans believe polarization is socially desirable (see Figure 2). On a 200-point scale, respondents said that partisans wanting to impress in-partisans would report high levels of affective polarization (60.76) and partisans wanting to *disappoint* in-partisans would report low levels (-45.39). This difference is 106.16 points—more than half the scale ($p<.001$).

Figure 2. Reported Polarization by *Impress* and *Disappoint* Prompts



Respondents were asked to report polarization based on wanting to impress or disappoint in-partisans. The difference between the two conditions is 106.16 points (p=.000). Lines represent 95% confidence intervals.

Further, the gap is driven by feelings toward both the in-party and out-party—the gap between the two prompts is 52.51 for in-party responses and 53.64 for out-party responses—as well as both Democrats and Republicans (although the gap for Democrats is slightly larger)—the gap is 108.55 for Democrats and 98.55 for Republicans. This demonstrates that Democrats and Republicans believe it is socially desirable for partisans to say they love the in-party but hate the out-party.

Study 2: Can Polarization Expressions be Shaped by Social Desirability? (Prolific, June 2021, N=973)

Design. Respondents from Prolific (N=973) were redirected to Qualtrics where they answered pre-treatment questions and then were randomly assigned to write about what they had for breakfast (*control*) or one of the two treatment groups that varied whether polarization was *desirable* or *undesirable* ("Think of (and write about) a specific instance where you thought you should act like [you hate [out-party members] (but like [in-party members]) / you didn't like people any more or less depending on their partisanship]"). If they had not been in this situation, they could write about an imagined time.[11] Respondents in the treatment conditions were then asked if they had been in this situation and how they reacted (or think they would have reacted). They were given a myriad

[11] 71.47% of respondents wrote about a real time, 25.67% about an imagined time, and 2.85% were unsure.

of options and could choose as many as they liked (see Supplementary Material B). Lastly, respondents were asked the *feeling thermometers* questions. Although this study involves random assignment, I present these results as largely observational, given that some respondents used recall and others used imagination—depending on their past experiences rather than random assignment.

Study 2a Results. I first examine respondents' reported reactions. Three options are pertinent to this research (see Table 1): agreeing with whoever was speaking, pretending to agree with them, and trying to fit in, where the latter two indicate changing polarization expressions (*morphing*). Honing in on the latter two, I find that 27.85% of respondents reported pretending to agree and 24.68% reported trying to fit in. Grouping these together (*morphing*: 1=pretending or trying to fit in; 0=otherwise), I find that almost half of the respondents reported engaging in this morphing behavior—41.14% said they did (or would) act differently depending on social desirability. Although this is certainly not a population estimate of morphing behavior, it is nonetheless quite surprising—especially as the outcomes are *reported* rather than observed, and thus are likely an underestimation of the sample's morphing behavior given reporting stigma and/or lack of awareness (Carlson and Settle 2022).

Table 1. Reported Reactions by Treatment

	N	% Agreed	% Pretend	% Fit In	% Morphed
Desirable Condition	319	36.05	30.41	24.76	42.63
Undesirable Condition	313	11.50	25.24	24.60	39.62
All Respondents	632	23.89	27.85	24.68	41.14

Respondents were asked to write about a social setting where affective polarization was either socially desirable (*desirable*) or where affective polarization was *not* socially desirable (*undesirable*). Respondents were asked their reactions in the social interaction, including options to say they agreed, they pretended to agree, and they tried to fit in. *Morphed* indicates respondents who said they pretended to agree and/or tried to fit in. Percentages demonstrate the percentage of respondents in each condition who selected these options (respondents could click as many options as they would like).

The goal of Study 2a was to examine if respondents reported morphing behavior in social settings. It was thus not focused on treatment effects, although respondents were randomized to examine if morphing varied by social desirability condition, which it did not (p=.441). Likewise, since there was no treatment effect examined here, I do not model self-monitoring as a moderator. I do, however, examine variables that predict reported morphing behavior to get a sense of who and in what contexts this is more likely to occur (see models in Supplementary Material B). In bivariate analysis, I find that morphing is more likely to occur among higher self-monitors (p=.028). Once including controls, however, this significance disappears, potentially suggesting multicollinearity with age (Berinsky 2004) and education (see analyses Supplementary Material F).[12] In this model, those who were younger (p<.001) and more educated (p=.001) were more likely to report morphing. This finding is suggestive of in what contexts we may see this behavior—perhaps, for example, more often in college dorms.

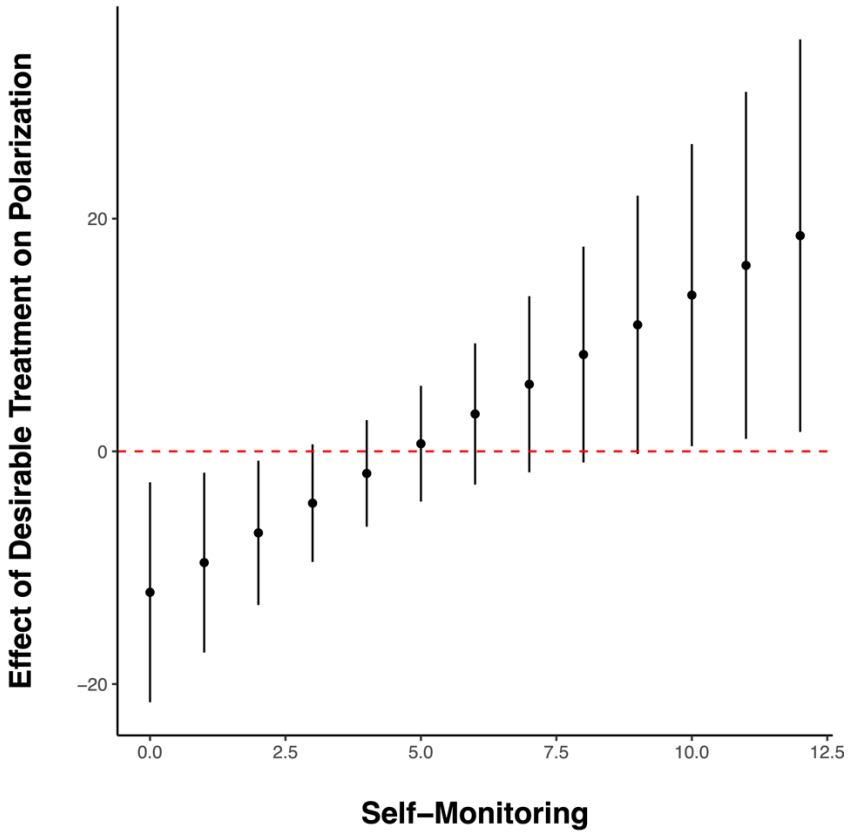
As a point of comparison, 23.89% of respondents reported agreeing with whoever was speaking—a similar, if slightly lower, percentage than those who either reported pretending (p=.099) or trying to fit in (p=.719), but *much* lower than those who said either of these responses (p<.001). Further, unlike the previous responses, these responses *did* differ by condition: more

[12] It is also possible that higher self-monitors are less honest about engaging in this morphing behavior.

participants reported agreeing when polarization was socially desirable (36.05%) than when it was socially *un*desirable (11.50%; $p<.001$). Thus, more respondents said they were polarized than were not polarized, but many respondents would pretend to be either. Lastly, note that even in the condition where the most respondents reported agreeing (36.05% in *desirable*), still (marginally) more respondents in this condition reported either pretending or trying to fit in (42.63%; $p=.079$).

Study 2b Results. Given the large proportion of respondents who reported changing their polarization behavior based on social desirability, this data demonstrates support for Hypothesis 2. This second set of analyses probes this further. I first examine if *feeling thermometers* responses differ by condition and find that this is not the case (see Supplementary Material B). Next, though, I incorporate self-monitoring by interacting it with treatment assignment to predict *feeling thermometers* responses (see Supplementary Material B). Here, I find that while there is no effect of *self-monitoring*undesirable* ($p=.215$), there is the predicted effect of *self-monitoring*desirable* ($p=.014$; see Figure 3). That is, the *desirable* treatment increased polarization reports among those who are higher in self-monitoring, suggesting partial support for Hypothesis 2.

Figure 3. Effect of Desirable Treatment on Polarization by Self-Monitoring



Respondents were given a prompt to talk about a social situation where affective polarization was socially desirable. This figure shows the marginal effect of that treatment (as compared to the control), moderated by self-monitoring (from low to high self-monitoring), on reported polarization. Full model in Supplementary Material B. Lines represent 95% confidence intervals.

Next I run robustness checks on the main finding (*self-monitoring*desirable*—all models in Supplementary Material B). Like in Study 1, I run the analyses separately for out-party and in-party feelings, finding stronger effects for the former than the latter. Although the trend is the same for both—widening the gap in feelings toward the out-party and in-party—the interaction coefficient is larger for decreasing out-party feelings (-1.94) than for increasing in-party feelings (0.66) and only the former is significant (p=.007)—the latter is not (p=.316). Similarly, I examine results separately for Democrats and Republicans and find that while effects are similar for both (interaction coefficients of 2.68 and 2.13, respectively), the sample of Republicans is so small (N=160) that the effect is no longer significant among this subgroup (p=.392) but is for Democrats (p=.020).

Next, I examine the linearity of self-monitoring and lastly, given that I am interacting a measured variable (self-monitoring) with a randomly assigned variable (see Kam and Trussler 2017), I also run the analyses with controls—*self-monitoring*desirable* remains significant (p=.007). Thus, Study 2 finds that social desirability can shape reported polarization behavior and (in some cases) polarization survey responses. Although the findings suggest support for Hypothesis 2, however, they are also not entirely straightforward. I return to this in the discussion.

Studies 3 and 4: Can Privacy Perceptions Shape Polarization Responses? (Mturk, November 2019, N=520, AmeriSpeak, March 2020, N=1,895)

"Man is least himself when he talks in his own person. Give him a mask, and he will tell you the truth." –Oscar Wilde

Design. Because Studies 3 and 4 were similar, they will be discussed together. In Study 3, respondents from Amazon's Mechanical Turk (Mturk; N=520) were redirected to Qualtrics where they answered pre-treatment questions and then were randomly assigned to be told: 1) "Just a reminder, the results from this study may be published" (*public*); 2) "Just a reminder, your responses are completely private (*private*); or 3) nothing (*control*). They then answered the *feeling thermometers* questions. In Study 4, AmeriSpeak Panel respondents (N=1,895) also answered pre-treatment questions and were randomly assigned to be told: 1) "Just a reminder, the results based on your responses may be published" (*public*); 2) "Just a reminder, your responses are completely private (*private*); or 3) nothing (*control*). They were then asked the *feeling thermometers* questions as well as the *trust* questions.[13]

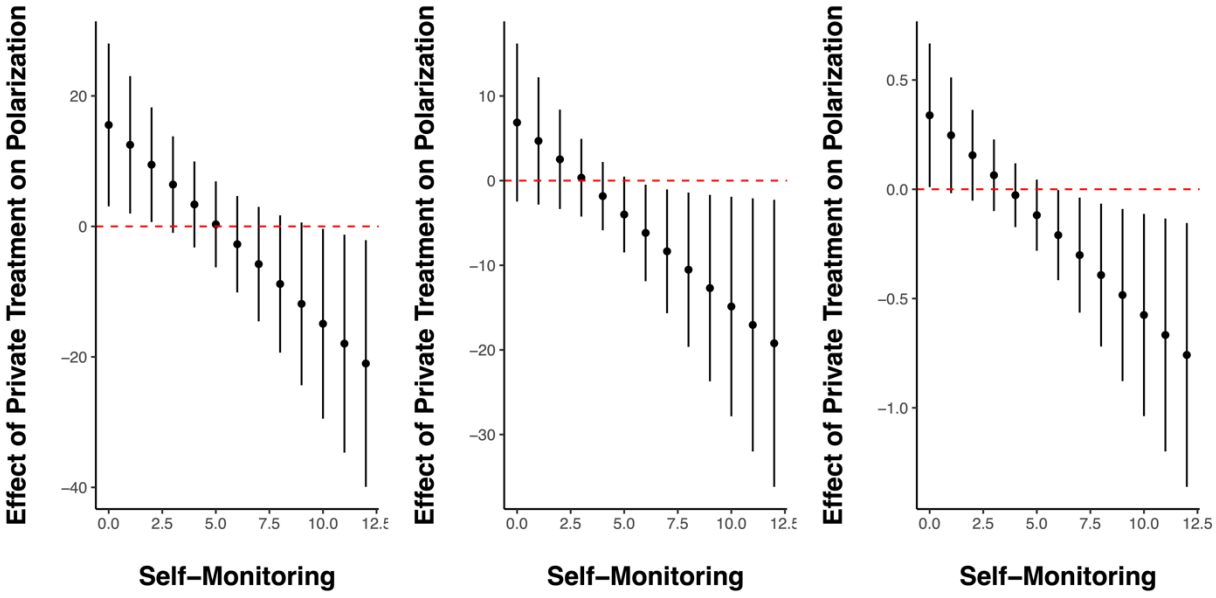
Results. I first examine if polarization responses differ by condition and find that this is not the case for all measures in either study (see Supplementary Material C and D). Next, though, I incorporate self-monitoring by interacting it with the treatments to predict *feeling thermometers* and *trust* responses (full models in Supplementary Material C and D). First, I find no effect of *self-monitoring*public* on any of the dependent measures in either study (Study 3, *feeling thermometers*: p=.545; Study 4, *feeling thermometers*: p=.216; Study 4, *trust*: p=.160). Thus, the public treatment was ineffective at changing respondents' polarization responses. This, however, makes sense when we consider that the control group as pre-treated—respondents have already reported as much polarization as they are willing to report to fit in. A public reminder will not

[13] There was one more condition that was in this design for other research.

influence their polarization responses because they are *already* driven by social desirability in the control condition.[14]

In this case, we should see the real shift in polarization when participants' perception of privacy *increases*. Indeed, here I find the predicted negative effect of *self-monitoring*private* on all three dependent measures in both studies (Study 3, *feeling thermometers*: p=.012; Study 4, *feeling thermometers*: p=.040; Study 4, *trust*: p=.015; see Figure 4). In fact, while the marginal *private* treatment effect for the lowest self-monitor on *feeling thermometers* is 15.55 and 6.86 for Studies 3 and 4, respectively, for the highest self-monitor it is -21.01 and -19.22—notable differences of 36.56 and 26.08 points. For *trust* (Study 4), this is 0.34 for the lowest self-monitor and -0.76 for the highest self-monitor. In other words, when being reminded that one's responses are completely private, higher self-monitors depress their reported polarization levels.

Figure 4. Effect of Private Treatment on Polarization by Self-Monitoring



Respondents were given a prompt that their responses were completely private. This figure shows the marginal effect of that treatment (as compared to the control), moderated by self-monitoring (from low to high self-monitoring), on reported polarization (left: *feeling thermometers* [Study 3]; middle: *feeling thermometers* [Study 4]; right: *trust* [Study 4]). Full models in Supplementary Material C and D. Lines represent 95% confidence intervals.

I next run robustness checks on this main finding (*self-monitoring*private*—all models in Supplementary Material C and D). Like in the previous studies, I run the analyses separately for out-party and in-party feelings, finding that the effect is quite similar for increasing positive feelings toward out-partisans and decreasing positive feelings toward in-partisans, although it is more robust for the latter across the two studies and different measures—the opposite asymmetry found in Study 2b. Similarly, I examine results separately for Democrats and Republicans and find that while the treatment works in the same direction for Democrats and Republicans, it is only

[14] It is also possible, given the unanticipated negative (but insignificant) effect of the public treatment, that the treatment instead reminded people to be more *accurate*, thus depressing reported polarization (similar to the effect of the private treatment)—although this explanation is post-hoc.

significant for Democrats for all measures in both studies. More research is needed to understand why this would be the case, as this is consistent with the asymmetries found in Studies 1 and 2b. Next, I examine the linearity of self-monitoring and also run the analyses with controls, replicating similar findings—although in one case (predicting *feeling thermometers* in Study 4) the interaction coefficient is no longer significant at conventional levels, it is in Study 3 and in predicting *trust* in Study 4. Lastly, I run a post-hoc manipulation check on Prolific (N=450) to examine if the treatments indeed influenced respondents' perception of privacy (see Supplementary Material E).

Thus, these findings demonstrate partial support for Hypothesis 3a but also partial support for Hypothesis 3b. The particular mix of null and significant findings, however, are quite informative. They suggest that in ordinary circumstances (e.g., on public opinion surveys or when talking with others), high self-monitors exaggerate their polarization levels to adhere to social desirability. When told they have more privacy, they are less motivated to impress others and thus temper their polarization levels. This effect is notable given the subtlety of the treatment and the existence of pre-treatment discussed earlier. However, like with Study 2, the findings also bring up questions—I discuss these in the next and final section.

Discussion & Conclusion

In this piece I found that partisans view polarization as socially desirable and that this can shape behavior and (in some cases) survey responses among partisans who want to impress others. Given obstacles to studying social dynamics especially in the context of pre-treatment, I attempted to address this topic with a variety of approaches and in doing so was also able to have diversity in both samples and measures of polarization expressions. My findings offer necessary nuance to our understanding of polarization and imply that social desirability—which can be shifted by contexts—can alter how polarized people act.

My results are a start to beginning to understand a social dimension to polarization that is currently missing from the literature—but they are certainly not the end. In fact, in some ways my findings offer more questions than answers. For example, do my findings extend to contexts outside of the US as well as other types of polarization? Although future research is needed, previous research offers some insight, suggesting that the social desirability of polarization should extend to other countries—as the US is not an outlier in polarization (Gidron, Adams, and Horne 2020)—and potentially even to other types of polarization—as research finds that social communication can influence ideological polarization (Druckman, Levendusky, and McLain 2018) and thus perhaps that social desirability could be driving some of this influence.

Additionally, why—despite in-party and out-party feelings changing differentially over time, where out-party feelings have declined substantially more than in-party feelings have increased—is there nearly perfect symmetry of these feelings in Study 1, where it is almost equally socially desirable for partisans to both love the in-party and hate the out-party? Relatedly, why do some treatments work differently with in-party and out-party attitudes as well as among Democrats and Republicans? And why does social desirability seem to shape reported behavior more than reported attitudes? This last incongruity could suggest that polarization responses are rather ingrained in partisans—that while partisans are aware of both this social desirability (Study 1) *and* its effect on their behavior (Study 2a), small changes in survey context (recall tasks and privacy prompts) do not necessarily produce large changes in polarization responses (Studies 2b, 3, and

4).[15] Perhaps stronger social pressures are needed to produce larger attitudinal effects. Relatedly, then, future research should explore how audience and group settings shape polarization expressions, as it is likely the case that social demands vary by these dynamics and that this thus shapes expressions.

Lastly, my findings cannot differentiate the "truly" polarized partisans from those whose polarization is more socially motivated—although findings from Study 2a suggest this latter group is a non-negligible proportion of partisans. Previous research offers some answers as well, though, suggesting that the truly polarized are more often strong, extreme partisans and that these partisans help to shape the polarization narrative for the broader public. For example, research shows that polarization is less contextually-dependent among strong partisans (Klar, Krupnikov, and Ryan 2018) and that strong and extreme partisans who are "deeply involved" in politics are more polarized than others (Krupnikov and Ryan 2022). The media then uses these individuals as partisan exemplars (Krupnikov and Ryan 2022), potentially helping to create the narrative that polarization is socially desirable—something that is likely reinforced in social settings, as the deeply involved are also more likely to discuss politics (Krupnikov and Ryan 2022).

While this is all suggestive of the types of partisans who are truly polarized versus those whose polarization is more socially motivated, my research unfortunately cannot definitively disentangle these two groups. Our often-used explicit measures of polarization cannot do so either, though, perhaps suggesting that we should expand the use of implicit measures to more precisely identify the truly polarized.[16] As Iyengar et al. (2019) note, "Implicit measures are known to be much harder to manipulate than explicit self-reports; they are therefore more valid and less susceptible to impression management (Boysen et al. 2006)." *However*, our measurement choice should speak to what we care about: people's "true" or expressed attitudes. And given that many political acts are expressive, perhaps an explicit measure—even if tainted by social desirability—is more informative (see Berinsky 2018).

Thinking about my findings in particular, if those who change their polarization expressions in social settings based on social desirability are *also* those who act more polarized in other settings (e.g. at political events) based on social desirability, perhaps (in some cases at least) we should care less about their true beliefs and more about their expressed beliefs. The consequences of the latter may be more important than an academic discussion of what constitutes a true belief. Future research might instead focus on which measures are more predictive of important outcomes such as political action, political trust, support for democratic norms, and political violence. In doing so, we could also better understand how polarization's social desirability shapes important outcomes. Thus, this research is a first step towards answering a multitude of important questions about polarization as a concept, polarization in social settings, and how polarization is related to certain outcomes and for whom.

[15] It is important to note, however, that responses being ingrained does not necessarily mean they are *not* shaped by social desirability—just that experimental treatments are less effective (see Druckman and Leeper 2012).

[16] Although, while this would certainly mitigate *some* socially motivated responding (indeed, Iyengar and Westwood [2015] found lower polarization levels with implicit. as compared to explicit, measures), any socially motivated responding that is internalized should still emerge with implicit measures.

References

Achen, Christopher H. and Larry M. Bartels. 2017. "Democracy for Realists." Princeton University Press.

Ahler, Douglas J. 2014. "Self-Fulfilling Misperceptions of Public Polarization." *The Journal of Politics* 76:607-20.

Ahler, Douglas J. and Gaurav Sood. 2018. "The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences." *The Journal of Politics* 80:964-81.

Arias, Eric. 2018. "How Does Media Influence Social Norms? Experimental Evidence on the Role of Common Knowledge." *Political Science Research and Methods* 7:561-78.

Bakker, Bert N., Yphtach Lelkes, and Ariel Malka. 2021. "Reconsidering the Link Between Self-Reported Personality Traits and Political Preferences." *American Political Science Review* 115:1482-98.

Banaji, Mahzarin R., and Deborah A. Prentice. 1994. "The Self in Social Contexts." *Annual Review of Psychology* 45:297-332.

Barbera, Pablo. 2014. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23:76-91.

Berinsky, Adam J. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43:1209-30.

Berinsky, Adam J. 2004. "Can We Talk? Self-Presentation and the Survey Response." *Political Psychology* 25:643-59.

Berinsky, Adam J. 2018. "Telling the Truth about Believing the Lies? Evidence for the Limited Prevalence of Expressive Survey Responding." *The Journal of Politics* 80:211-24.

Berinsky, Adam J. and Howard Lavine. 2007. "Self-Monitoring and Political Attitudes: 2006 NES Pilot Study Report." *NES Pilot Study Reports*:1–21.

Berinsky, Adam J. and Howard Lavine. 2012. "Self-Monitoring and Political Attitudes." *Improving Public Opinion Surveys: Interdisciplinary Innovation and the American National Election Studies*:27-45.

Bishop, Bill. 2008. *The Big Sort: Why the Clustering of Like-Minded America is Tearing Us Apart.* Wilmington, Delaware: Mariner Books.

Boysen, Guy A., David L. Vogel, and Stephanie Madon. 2006. "A Public Versus Private Administration of the Implicit Associations Test." *European Journal of Social Psychology* 36:845-56.

Broockman, David E., Joshua L. Kalla, and Sean J. Westwood. 2022. "Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not." *American Journal of Political Science.*

Butters, Ross, and Christopher Hare. 2022. "Polarized Networks? New Evidence on American Voters' Political Discussion Networks." *Political Behavior* 44:1079-103.

Carlson, Taylor N., and Jaime E. Settle. 2016. "Political Chameleons: An Exploration of Conformity in Political Discussions." *Political Behavior* 38:817-59.

Carlson, Taylor N. and Jaime Settle. 2022. *What Goes Without Saying: Navigating Political Discussion in America*. Cambridge University Press: New York.

Cassese, Erin. 2021. "Partisan Dehumanization in American Politics." *Political Behavior* 43:29-50.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren. 1990. "A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public

Places." *Journal of Personality and Social Psychology* 58:1015.
Claassen, Ryan L. and John Barry Ryan. 2016. "Social Desirability, Hidden Biases, and Support for Hillary Clinton." *PS* 49:730-35.
Clifford, Scott and Jennifer Jerit. 2015. "Do Attempts to Improve Respondents Attention Increase Social Desirability Bias?" *Public Opinion Quarterly* 79:790–802.
Connors, Elizabeth C. 2020. "The Social Dimension of Political Values." *Political Behavior* 42:961-82.
Connors, Elizabeth C., Yanna Krupnikov, John Barry Ryan. 2019. "How Transparency Affects Survey Responses." *Public Opinion Quarterly* 83:185-209.
Conover, Pamela Johnston, and Donald D. Searing. 2005. "Studying 'Everyday Political Talk' in the Deliberative System." *Acta Politica* 40:269-83.
Conover, Pamela Johnston, Donald D. Searing, and Ivor M. Crewe. 2002. "The Deliberative Potential of Political Discussion." *British Journal of Political Science* 32:21-62.
Cosmides, Leda, and John Tooby. 1992. "Cognitive Adaptations for Social Exchange." *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*:163-228.
Daoust, Jean-François, Richard Nadeau, Ruth Dassonneville, Erick Lachapelle, Éric Bélanger, Justin Savoie, and Clifton van der Linden. 2021. "How to Survey Citizens' Compliance with COVID-19 Public Health Measures: Evidence from Three Survey Experiments." *Journal of Experimental Political Science* 8:310-17.
Dias, Nicholas, and Yphtach Lelkes. 2022. "The Nature of Affective Polarization: Disentangling Policy Disagreement from Partisan Identity." *American Journal of Political Science* 66:775-90.
Douglas, Mary, and Aaron Wildavsky. 1982. *Risk and Culture: An Essay on the Selection of Technical and Environmental Dangers*. Berkeley: University of California Press.
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* 8:223-34.
Druckman, James N., Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2022. "(Mis) estimating Affective Polarization." *The Journal of Politics* 84:1106-17.
Druckman, James N., and Thomas J. Leeper. 2012. "Learning More from Political Communication Experiments: Pre-Treatment and its Effects." *American Journal of Political Science* 56:875-96.
Druckman, James N. and Matthew S. Levendusky. 2019. "What Do We Measure When We Measure Affective Polarization?" *Public Opinion Quarterly*, 83:114-22.
Druckman, James N., Matthew S. Levendusky, and Audrey McLain. 2018. "No Need to Watch: How the Effects of Partisan Media Can Spread Via Interpersonal Discussions." *American Journal of Political Science* 62:99-112.
Ellemers, Naomi, Russell Spears, and Bertjan Doosje. 2002. "Self and Social Identity." *Annual Review of Psychology* 53:161-86.
Epley, Nicholas and David Dunning. 2000. "Feeling 'Holier Than Thou': Are Self-Serving Assessments Produced by Errors in Self or Social Prediction." *Journal of Personality and Social Psychology* 79:861.
Feldman, Stanley and Leonie Huddy. 2005. "Racial Resentment and White Opposition to Race-Conscious Programs: Principles or Prejudice?" *American Journal of Political Science*

49:168–83.
Fieldhouse, Edward, David Cutts, and Jack Bailey. 2022. "Who Cares If You Vote? Partisan Pressure and Social Norms of Voting." *Political Behavior* 44: 1297-316.
Gangestad, Steven W., and Mark Snyder. 2000. "Self-Monitoring: Appraisal and Reappraisal." *Psychological Bulletin* 126:530-55.
Gerber, Alan S. and Todd Rogers. 2009. "Descriptive Social Norms and Motivation to Vote: Everybody's Voting and So Should You." *The Journal of Politics* 71:178-91.
Gidron, Noam, James Adams, and Will Horne. 2020. *American Affective Polarization in Comparative Perspective*. Cambridge University Press.
Goffman, Erving. 1955. "On Face Work: An Analysis of Ritual Elements in Social Interaction." *Psychiatry* 18:213-21.
Goffman, Erving. 1967. "On Face-Work." *Interaction Ritual*:5-45.
Green, Donald, Bradley Palmquist, and Eric Schickler. 2002. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters.* New Haven, CT: Yale University Press.
Haley, Kevin J., and Daniel MT Fessler. 2005. "Nobody's Watching?: Subtle Cues Affect Generosity in an Anonymous Economic Game." *Evolution and Human Behavior* 26:245-56.
Hetherington, Marc and Thomas Rudolph. 2015. *Why Washington Won't Work: Polarization, Political Trust, and the Governing Crisis.* Chicago: University of Chicago Press.
Hogg, Michael A. 2001. "A Social Identity Theory of Leadership." *Personality and Social Psychology Review* 5:184-200.
Holbrook, Allyson K., Melanie C. Green, and Jon A. Krosnick. 2003. "Telephone Versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparison of Respondent Satisficing and Social Desirability Response Bias." *Public Opinion Quarterly* 61:79-125.
Holtgraves, Thomas. 1992. "The Linguistic Realization of Face Management: Implications for Language Production and Comprehension, Person Perception, and Cross-Cultural Communication." *Social Psychology Quarterly*:141-59.
Holtgraves, Thomas. 2004. "Social Desirability and Self Reports: Testing Models of Socially Desirable Responding." *Personality and Social Psychology Bulletin* 30:161-72.
Huddy, Leonie. 2013. "From Group Identity to Political Cohesion and Commitment." In Leonie Huddy, David O. Sears, and Jack S. Levy, eds., *The Oxford Handbook of Political Psychology*. Oxford University Press.
Huddy, Leonie, and Omer Yair. 2021. "Reducing Affective Polarization: Warm Group Relations or Policy Compromise?" *Political Psychology* 42:291-309.
Iyengar, Shanto, Tobias Konitzer, and Kent Tedin. 2018. "The Home as a Political Fortress: Family Agreement in an Era of Polarization." *Journal of Politics* 80:1326-38.
Iyengar, Shanto, and Masha Krupenkin. 2018. "The Strengthening of Partisan Affect." *Political Psychology* 39:201-18.
Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. "The Origins and Consequences of Affective Polarization in the United States." *Annual Review of Political Science* 22:129-46.
Iyengar, Shanto, and Sean J. Westwood. 2015. "Fear and Loathing across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* 59:690-707.
Kahan, Dan M., and Hank Jenkins-Smith, and Donald Braman. 2011. "Cultural Cognition of Scientific Consensus." *Journal of Risk Research* 14:147-74.

Kam, Cindy D., and Marc J. Trussler. 2017. "At the Nexus of Experimental and Observational Research: Theory, Specification, and Analysis of Experiments with Heterogeneous Treatment Effects." *Political Behavior* 39:789-815.

Karp, Jeffrey A. and David Brockington. 2005. "Social Desirability and Response Validity: A Comparative Analysis of Overreporting Voter Turnout in Five Countries." *The Journal of Politics* 67:825-40.

Kingzette, Jon, James N. Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2021. "How Affective Polarization Undermines Support for Democratic Norms." *Public Opinion Quarterly* 85:663-77.

Klar, Samara, and Yanna Krupnikov. 2016. *Independent Politics: How American Disdain for Parties Leads to Political Inaction*. Cambridge University Press.

Klar, Samara, Yanna Krupnikov, and John Ryan. 2018. "Affective Polarization or Partisan Disdain?: Untangling a Dislike for the Opposing Party from a Dislike of Partisanship." *Public Opinion Quarterly* 82:379-90.

Krupnikov, Yanna, and John Barry Ryan. 2022. *The Other Divide: Polarization and Disengagement in American Politics*. Cambridge University Press.

Kuran, Timur. 1997. *Private Truths, Public Lies: The Social Consequences of Preference Falsification.* Harvard University Press.

Lee, Amber Hye-Yon. 2022. "Social Trust in Polarized Times: How Perceptions of Political Polarization Affect Americans' Trust in Each Other." *Political Behavior* 44:1533-54.

Levendusky, Matthew. 2009. *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans.* University of Chicago Press.

Levendusky, Matthew and Neil Malhotra. 2016a. "(Mis)perceptions of Partisan Polarization in the American Public." *Public Opinion Quarterly* 80:378-91.

Levendusky, Matthew and Neil Malhotra. 2016b. "Does Media Coverage of Partisan Polarization Affect Political Attitudes?" *Political Communication* 33:283-301.

Levitan, Lindsey C., and Brad Verhulst. 2016. "Conformity in Groups: The Effects of Others' Views on Expressed Attitudes and Attitude Change." *Political Behavior* 38:277-315.

Martherus, James L., Andres G. Martinez, Paul K. Piff, and Alexander G. Theodoridis. 2021. "Party Animals? Extreme Partisan Polarization and Dehumanization." *Political Behavior* 43:517-40.

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

Miller, Kevin P., Marilynn B. Brewer, and Nathan L. Arbuckle. 2009. "Social Identity Complexity: Its Correlates and Antecedents." *Group Processes & Intergroup Relations* 12:79-94.

Mutz, Diana C. 2002. "Cross-Cutting Social Networks: Testing Democratic Theory in Practice." *American Political Science Review* 96:111-26.

Orr, Lilla V., and Gregory A. Huber. 2020. "The Policy Basis of Measured Partisan Animosity in the United States." *American Journal of Political Science* 64:569-86.

Paulhus, Delroy L. 1991. "Measurement and Control of Response Bias." In *Measures of Personality and Social Psychology*, edited by John P. Robinson, Philip R. Shaver, and Lawrence S. Wrightsman, 17-59. New York: Academic Press.

Paulhus, Delroy L., Peter D. Harms, M. Nadine Bruce, and Daria C. Lysy. 2003. "The Over-Claiming Technique: Measuring Self-Enhancement Independent of Ability." *Journal of Personality and Social Psychology* 84:890-904.

Perez-Truglia, Ricardo, and Guillermo Cruces. 2017. "Partisan Interactions: Evidence from a Field

Experiment in the United States." *Journal of Political Economy* 125:1208-43.
Premeaux, Sonya Fontenot, and Arthur G. Bedeian. 2003. "Breaking the Silence: The Moderating Effects of Self-Monitoring in Predicting Speaking Up in the Workplace." *Journal of Management Studies* 40:1537-62.
Rigdon, Mary, Keiko Ishii, Motoki Watabe and Shinobu Kitayama. 2009. "Minimal Social Cues in the Dictator Game." *Journal of Economic Psychology* 30:358-67.
Rimal, Rajiv M., and Kevin Real. 2003. "Understanding the Influence of Perceived Norms on Behaviors." *Communication Theory* 13:184-203.
Rimal, Rajiv M., and Kevin Real. 2005. "How Behaviors are Influenced by Perceived Norms: A Test of the Theory of Normative Social Behavior." *Communication Research* 32:389-414.
Robison, Joshua, and Kevin J. Mullinix. 2016. "Elite Polarization and Public Opinion: How Polarization is Communicated and its Effects." *Political Communication* 33:261-282.
Ross, Andrew RN, and Delia Dumetrescu. 2019. "'Vox Twitterati': Investigating the Effects of Social Media Exemplars in Online News Articles." *New Media & Society* 21:962-83.
Schuman, Howard, Stanley Presser, and Jacob Ludwig. 1981. "Context Effects on Survey Responses to Questions about Abortion." *Public Opinion Quarterly* 45:216-23.
Sinclair, Betsy. 2012. *The Social Citizen: Peer Networks and Political Behavior.* University of Chicago Press.
Snyder. 1974. "Self-Monitoring of Expressive Behavior." *Journal of Personality and Social Psychology* 30:526.
Snyder, Mark. 1979. "Cognitive, Behavioral, and Interpersonal Consequences of Self-Monitoring." *Perception of Emotion in Self and Others*. Boston, MA: Springer.
Snyder, Mark, and Steven W. Gangestad. 1986. "On the Nature of Self-Monitoring: Matters of Assessment, Matters of Validity." *Journal of Personality and Social Psychology* 51:125-39.
Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.
Tajfel, Henri, and John C. Turner. 1979. "An Integrative Theory of Intergroup Conflict." In William G. Austin and Stephen Worchel, eds., *The Social Psychology of Inter-Group Relations*. Monterey, CA: Brooks Cole, 33-47.
Terry, Deborah J., and Michael A. Hogg. 1996. "Group Norms and the Attitude-Behavior Relationship: A Role for Group Identification." *Personality and Social Psychology Bulletin* 22:776-93.
Toff, Benjamin, and Elizabeth Suhay. 2019. "Partisan Conformity, Social Identity, and the Formation of Policy Preferences." *International Journal of Public Opinion Research* 31:349-67.
Turner, John C., Rupert J. Brown, and Henri Tajfel. 1979. "Social Comparison and Group Interest in Ingroup Favouritism." *European Journal of Social Psychology* 9:187-204.
Tyler, Matthew and Shanto Iyengar. 2022. "Learning to Dislike Your Opponents: Political Socialization in the Era of Polarization." *American Political Science Review*:1-8.
Vazire, Simine, and Erika N. Carlson. 2011. "Others Sometimes Know Us Better Than We Know Ourselves." *Current Directions in Psychological Science* 20:104-8.
Voelkel, Jan G., James Chu, Michael N. Stagnaro, Joseph S. Mernyk, Chrystal Redekopp, Sophia L. Pink, James N. Druckman, David G. Rand, and Robb Willer. 2022. "Interventions

Reducing Affective Polarization Do Not Improve Anti-Democratic Attitudes." *Working Paper.*
Walsh, Katherine. 2004. *Talking about Politics: Informal Groups and Social Identity in American Life*. Chicago, IL: University of Chicago Press.
Weber, Christopher, Howard Lavine, Leonie Huddy and Christopher Federico. 2014. "Placing Racial Stereotypes in Context: Social Desirability and the Politics of Racial Hostility." *American Journal of Political Science* 58:63-78.
Webster, Steven W., and Alan I. Abramowitz. 2017. "The Ideological Foundations of Affective Polarization in the U.S., Electorate." *American Politics Research* 45:621-47.
West, Emily A., and Shanto Iyengar. 2022. "Partisanship as a Social Identity: Implications for Polarization." *Political Behavior* 44:807-38.
Zaller, John. 1992. *The Nature and Origins of Mass Opinion.* New York: Cambridge University Press.

Supplementary Material A: Study 1

Sample Information:

Prolific is an online platform where respondents opt-in to take surveys and get paid for their participation (for more information see https://www.prolific.co). Although Prolific produces a quality sample, I also prevented multiple submissions to avoid "ballot stuffing."

For this particular survey, I recruited 943 US adult partisans, 920 of which completed the survey. The sample was 79.78% Democrat and 20.22% Republican (with 60.11% of these strong partisans and 39.89% of these weak partisans); with a mean of 2.78 and standard deviation of 1.79 from 1 (extremely liberal) to 7 (extremely conservative); 55.66% women, 43.14% men, and 1.20% other; 69.35% white and 30.65% either mixed or full minority; a mean age of 33.30 and standard deviation of 12.23; and 56.09% received at least a bachelors college.

As a comparison, American National Election Studies (ANES) 2020 data has the following breakdown. 23.78% were strong Democrats, 10.92% were weak Democrats, 11.83% were leaning Democrats, 10.66% were leaning Republicans, 10.09% were weak Republicans, 20.98% were strong Republicans, and 11.74% were pure independents. The ANES sample had a mean ideology of 4.09 and standard deviation of 1.67 on a scale from 1 (extremely liberal) to 7 (extremely conservative). It was 45.45% male, 53.74% female, and 0.81% NA; 72.92% white; with a mean age of 51.59 and standard deviation of 17.21. Lastly, 44.75% had a bachelor's degree or more.

Survey:

1. [PID] Generally speaking, do you think of yourself as a Republican, a Democrat, an Independent, or what? [Republican / Democrat / independent / something else [____]]
    a. Would you call yourself a strong [Democrat/Republican] or a not very strong [Democrat/Republican]? [strong [Democrat/Republican] / not very strong [Democrat/Republican]]
2. [ideology] We hear a lot of talk these days about liberals and conservatives. Here is a 7-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale, or haven't you thought much about this? [extremely liberal / liberal / slightly liberal / moderate / slightly conservative / conservative / extremely conservative / don't know]
3. [gender] What is your gender? [man / woman / other]
4. [age] What is your age? [ ]
5. [race] What racial or ethnic group or groups best describes you? [white / black / Hispanic / Asian / Native American / other]
6. [education] What is the highest level of education that you have completed? [did not complete a high school degree / high school degree / some college / Associate's degree / Bachelor's degree / graduate or professional degree]
7. [media use] During a typical week, how many days do you watch, read, or listen to news on the following medium: the Internet (including online newspapers) / the TV / print newspapers / the radio [0 days → 7 days]
8. [discuss] During a typical week, how many days do you discuss politics with your family and/or friends? [0 days → 7 days]
9. [self-monitoring 1] When you are with other people, how often do you put on a show to impress or entertain them? [always / most of the time / some of the time / once in a while / never]
10. [self-monitoring 2] When you are in a group of people, how often are you the center of

attention? [always / most of the time / some of the time / once in a while / never]

11. [self-monitoring 3] How good or poor of an actor would you be? [excellent / good / fair / poor / very poor]
12. [randomize to a or b]
    a. [fake good] **Please read the following 2 questions and answer them in a way that you think a [Democrat / Republican] would in order to impress other [Democrats / Republicans] (even if this is not your actual opinion):**
    We'd like to get your feelings toward the two national parties. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the party. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the party and that you don't care too much for that party. You would rate the party at the 50-degree mark if you don't feel particularly warm or cold toward the party. [randomize order of i and ii]
        i. How would you rate Democrats (again, as a [Democrat / Republican] wanting to impress other [Democrats / Republicans])? [0 to 100 degrees]
        ii. How would you rate Republicans (again, as a [Democrat / Republican] wanting to impress other [Democrats / Republicans])? [0 to 100 degrees]
    b. [fake bad] **Please read the following 2 questions and answer them in a way that you think a [Democrat / Republican] would in order to disappoint other [Democrats / Republicans] (even if this is not your actual opinion):**
    We'd like to get your feelings toward the two national parties. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the party. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the party and that you don't care too much for that party. You would rate the party at the 50-degree mark if you don't feel particularly warm or cold toward the party. [randomize order of i and ii]
        i. How would you rate Democrats (again, as a [Democrat / Republican] wanting to disappoint other [Democrats / Republicans])? [0 to 100 degrees]
        ii. How would you rate Republicans (again, as a [Democrat / Republican] wanting to disappoint other [Democrats / Republicans])? [0 to 100 degrees]
13. If you would like to add comments or feedback: [_____]

Supplementary Material B: Study 2

Sample Information:
Prolific is an online platform where respondents opt-in to take surveys and get paid for their participation (for more information see https://www.prolific.co). Although Prolific produces a quality, attentive sample, I also included a Captcha verification at the beginning of the survey to prohibit bots and prevented multiple submissions to avoid "ballot stuffing."

For this particular survey, I recruited 1,003 US adults, 30 of which were not included in this research as they were not partisans. The final sample of 973 was 74.72% Democrat and 25.28% Republican (with 57.97% of these strong partisans and 42.03% of these weak partisans); with a mean of 3.08 and standard deviation of 1.92 from 1 (extremely liberal) to 7 (extremely conservative); 53.09% women, 45.99% men, and 0.93% other; 65.36% white and 34.64% either mixed or full minority; with a mean age of 36.25 and standard deviation of 13.14; and 64.09% received at least a bachelors degree.

As a comparison, American National Election Studies (ANES) 2020 data has the following breakdown. 23.78% were strong Democrats, 10.92% were weak Democrats, 11.83% were leaning Democrats, 10.66% were leaning Republicans, 10.09% were weak Republicans, 20.98% were strong Republicans, and 11.74% were pure independents. The ANES sample had a mean ideology of 4.09 and standard deviation of 1.67 on a scale from 1 (extremely liberal) to 7 (extremely conservative). It was 45.45% male, 53.74% female, and 0.81% NA; 72.92% white; with a mean age of 51.59 and standard deviation of 17.21. Lastly, 44.75% had a bachelor's degree or more.

Survey:
First, we'd like to know a bit about you.

1. [PID] Generally speaking, do you think of yourself as a Republican, a Democrat, an Independent, or what? [Republican / Democrat / independent / something else [_____]]
    a. [if Democrat or Republican] Would you call yourself a strong [Democrat/Republican] or a not very strong [Democrat/Republican]? [strong [Democrat/Republican] / not very strong [Democrat/Republican]]
    b. [if independent or something else] Do you think of yourself as closer to the Republican Party or the Democratic Party? [closer to the Republican Party / closer to the Democratic Party / neither]
2. [identity] How important is being a [Democrat / Republican] to your identity? [not at all important / a little important / moderately important / very important / extremely important]
3. [ideology] We hear a lot of talk these days about liberals and conservatives. Here is a 7-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale, or haven't you thought much about this? [extremely liberal / liberal / slightly liberal / moderate / slightly conservative / conservative / extremely conservative / don't know]
4. [interest] Some people don't pay much attention to political news. How about you? Would you say that you are very much interested, somewhat interested, or not interested at all interested in political news? [not at all interested / somewhat interested / very much interested]
5. [gender] What is your gender? [man / woman / other]
6. [age] What is your age? [ ]
7. [race] What racial or ethnic group or groups best describes you? [white / black / Hispanic / Asian / Native American / other (please specify):___]

8. [education] What is the highest level of education that you have completed? [did not complete a high school degree / high school degree / some college / Associate's degree / Bachelor's degree / graduate or professional degree]
9. [media] During a typical week, how many days do you watch, read, or listen to news on the following medium: (the internet (including online newspapers), the TV, print newspapers, the radio) [0 days / 1 day / 2 days / 3 days / 4 days / 5 days / 6 days / 7 days]
10. [discuss] During a typical week, how many days do you discuss politics with your family and/or friends? [0 days / 1 day / 2 days / 3 days / 4 days / 5 days / 6 days / 7 days]
11. [self-monitoring 1] When you are with other people, how often do you put on a show to impress or entertain them? [always / most of the time / some of the time / once in a while / never]
12. [self-monitoring 2] When you are in a group of people, how often are you the center of attention? [always / most of the time / some of the time / once in a while / never]
13. [self-monitoring 3] How good or poor of an actor would you be? [excellent / good / fair / poor / very poor]
14. [randomize to treatment a, b, or c]
    a. [control] Next, we'd like you to write about what you had for breakfast this morning. Be as specific as possible.
    b. [desirable] Next, we'd like to ask you about situations where social norms may have shaped your behavior. Think of (and write about) a specific instance where you thought you should act like you hate [out-party members] (but like [in-party members]). Be as specific as possible. If you have not been in this situation, write about an imagined time.
    c. [undesirable] Next, we'd like to ask you about situations where social norms may have shaped your behavior. Think of (and write about) a specific instance where you thought you should act like you didn't like people any more or less depending on their partisanship. Be as specific as possible. If you have not been in this situation, write about an imagined time.
15. [if treatment b or c] Have you been in situations like the one from before or did you write about an imagined time? [yes, I have been in situations like this / no I haven't been in situations like this / I can't remember]
    a. [if yes] How did you react in the situation? (check **all** that apply) [agreed with whoever was speaking / pretended I agreed with them / tried to fit in / tried to leave the situation / tried to change the topic / tried to be as quiet as possible / disagreed with whoever was speaking / tried to justify my opinion / tried to convince whoever was speaking that they were wrong / other (please explain):___]
    b. [if no or can't remember] How do you think you would react in the situation? (check **all** that apply) [agree with whoever was speaking / pretend I agreed with them / try to fit in / try to leave the situation / try to change the topic / try to be as quiet as possible / disagree with whoever was speaking / try to justify my opinion / try to convince whoever was speaking that they were wrong / other (please explain):___]
16. [affective polarization – if treatment a] Now, what are your feelings toward the two national parties? Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the party. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the party and that you don't care too much for that party. You would rate the party at the 50-degree mark if you don't feel particularly warm or cold toward the party. [randomize order of a and b]

a. How would you rate Democrats? [0 to 100 degrees]
b. How would you rate Republicans? [0 to 100 degrees]

17. [affective polarization – if treatment b or c] Still thinking about that situation, how would you place yourself on the following scale during or after that conversation? What are your feelings toward the two national parties? Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the party. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the party and that you don't care too much for that party. You would rate the party at the 50-degree mark if you don't feel particularly warm or cold toward the party. [randomize order of a and b]
    a. How would you rate Democrats? [0 to 100 degrees]
    b. How would you rate Republicans? [0 to 100 degrees]
18. Thank you for your participation! If you have any comments or feedback, add them below (if not, leave blank): [_____]

Study 2a Models:

Table B1. Predicting Morphing

Variable	Model 1	Model 2
Self-Monitoring	0.08	0.02
	(.03)	(.04)
Democrat	-	0.02
	-	(.32)
Strength	-	-0.13
	-	(.21)
Ideology 2	-	0.57
	-	(.24)
Ideology 3	-	0.28
	-	(.34)
Ideology 4	-	0.52
	-	(.43)
Ideology 5	-	0.19
	-	(.50)
Ideology 6	-	0.51
	-	(.40)
Ideology 7	-	0.69
	-	(.51)
White	-	-0.20
	-	(.18)
Woman	-	-0.18
	-	(.18)
Age	-	-0.03
	-	(.01)
Education	-	0.23
	-	(.07)
Constant	-0.66	-0.35
	(.16)	(.62)

Table shows coefficients with standard errors in parentheses. Model 1 (logit) = predicting *morphing* by *self-monitoring* (N=632—1 respondent did not answer at least 1 of the self-monitoring questions and thus was dropped from this analysis). Model 2 (logit) = predicting *morphing* by *self-monitoring*, *partisanship*, *partisan strength*, *ideology*, *race*, *gender*, *age*, and *education* (N=629—1 respondent did not answer at least 1 of the self-monitoring questions, 2 respondents did not answer the gender question, and 1 respondent did not answer the education question [these 3 respondents were dropped from the analysis]). *Ideology* is in comparison to extremely liberal and values range from 1 to 7, extremely liberal to extremely conservative.

Study 2b Models:

Table B2. Polarization Responses by Treatment

Condition	N	Mean	SE
Desirable	317	51.81	1.70
Control	337	53.47	1.61
Undesirable	306	52.03	1.84

Table shows respondents' polarization responses by treatment. None of these are statistically different from one another at $p<.05$.

Table B3. Polarization Responses by Treatment and Self-Monitoring

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Desirable	-	-12.27	8.31	-4.08	-11.83	-15.77	-12.09
	-	(4.83)	(3.17)	(3.26)	(5.03)	(12.47)	(4.32)
Undesirable	-6.93	-	-	-	-	-	-
	(5.07)	-	-	-	-	-	-
Self-Monitoring	-1.43	-1.43	2.51	1.04	-1.73	-0.36	-1.41
	(.71)	(.71)	(.52)	(.46)	(.80)	(1.60)	(.72)
Desirable*Self-	-	2.58	-1.94	0.66	2.68	2.13	2.58
Monitoring	-	(1.04)	(.72)	(.66)	(1.13)	(2.48)	(.96)
Undesirable*	1.31	-	-	-	-	-	-
Self-Monitoring	(1.05)	-	-	-	-	-	-
Democrat	-	-	-	-	-	-	-2.47
	-	-	-	-	-	-	(4.75)
Strength	-	-	-	-	-	-	15.11
	-	-	-	-	-	-	(2.62)
Ideology 2	-	-	-	-	-	-	-6.88
	-	-	-	-	-	-	(2.59)
Ideology 3	-	-	-	-	-	-	-18.42
	-	-	-	-	-	-	(3.72)
Ideology 4	-	-	-	-	-	-	-25.56
	-	-	-	-	-	-	(5.92)
Ideology 5	-	-	-	-	-	-	-44.53
	-	-	-	-	-	-	(7.00)
Ideology 6	-	-	-	-	-	-	-16.60
	-	-	-	-	-	-	(5.31)
Ideology 7	-	-	-	-	-	-	-2.26
	-	-	-	-	-	-	(6.75)
White	-	-	-	-	-	-	-0.45

	-	-	-	-	-	-	(2.23)
Woman	-	-	-	-	-	-	2.57
	-	-	-	-	-	-	(2.06)
Other Gender	-	-	-	-	-	-	11.15
	-	-	-	-	-	-	(9.42)
Age	-	-	-	-	-	-	0.35
	-	-	-	-	-	-	(.09)
Education	-	-	-	-	-	-	-1.27
	-	-	-	-	-	-	(.81)
Constant	59.43	59.43	11.09	70.76	62.32	50.66	39.91
	(3.33)	(3.32)	(2.24)	(2.17)	(3.52)	(8.05)	(7.98)

Table shows coefficients predicting reported polarization with robust standard errors in parentheses. 1 respondent did not answer at least 1 of the self-monitoring questions and thus was dropped from all the following analyses. Model 1 (OLS) = *self-monitoring*undesirable* (N=642—11 respondents did not answer at least one polarization question in these two conditions and were dropped from this model). Model 2 (OLS) = *self-monitoring*desirable* (N=653—6 respondents did not answer at least one polarization question in these two conditions and were dropped from this model). Model 3 (OLS) = *self-monitoring*desirable*, just out-party feelings (N=654). Model 4 (OLS) = *self-monitoring* desirable*, just in-party feelings (N=657). Model 5 (OLS) = *self-monitoring*desirable*, just Democrats (N=493). Model 6 (OLS) = *self-monitoring*desirable*, just Republicans (N=160). Model 7 (OLS) = *self-monitoring*desirable* with controls (N=648—3 respondents did not answer the ideology question, 1 respondent did not answer the education question, and 1 respondent did not answer the gender question [these 5 respondents were dropped from the analysis]). *Ideology* is in comparison to extremely liberal and values range from 1 to 7, extremely liberal to extremely conservative. *Woman* and *other gender* are in comparison to men.

Quartile Analysis. Table B4 examines the potential non-linear moderating effect of self-monitoring with the desirable treatment (replicating the model used in Figure 3, but interacting the treatment with quartiles of self-monitoring rather than a continuous self-monitoring variable). The table shows that there are no non-linear interaction effects.

Table B4. Modeling Self-Monitoring as Non-Linear

Variable	**Model 1**
Desirable	-3.05
	(4.73)
Self-Monitoring Quartile 2	2.59
	(4.14)
Self-Monitoring Quartile 3	5.12
	(4.99)
Self-Monitoring Quartile 4	-5.30
	(4.64)
Desirable*Self-Monitoring Quartile 2	-0.66
	(5.99)
Desirable* Self-Monitoring Quartile 3	-7.11
	(7.33)
Desirable*Self-Monitoring Quartile 4	9.07
	(6.96)
Constant	53.43
	(3.21)

Just like in the table above, this table shows coefficients predicting reported polarization with robust standard errors in parentheses (OLS). Model 1 models self-monitoring as non-linear, using four different dummy variables (1=below or equal to 25%, 2=between 25% and 50% [inclusive], 3=between 50% and 75% [inclusive], 4=above 75%) each interacted with the treatment. N=653—1 respondent did not answer at least 1 of the self-monitoring questions and thus was dropped from the analysis and 6 respondents did not answer at least one polarization question in these two conditions and were also dropped from this analysis.

Supplementary Material C: Study 3

Sample Information:
Amazon's Mechanical Turk (Mturk) is an online platform where people sign up to get paid to take surveys (see Berinsky, Huber, and Lenz 2012). For this study, I prevented multiple submissions to avoid "ballot stuffing."

For the study, I recruited 742 US adults, 222 of which were not included in this research as they were not partisans. The final sample of 520 was 63.65% Democrat and 36.35% Republican (with 59.81% of these strong partisans and 40.19% of these weak partisans); with a mean of 3.45 and standard deviation of 1.98 from 1 (extremely liberal) to 7 (extremely conservative); 40.73% women and 59.27% men; 70.77% white and 29.23% either mixed or full minority; with a mean age of 35.91 and standard deviation of 10.50; and 64.24% received at least a bachelors degree.

As a comparison, American National Election Studies (ANES) 2020 data has the following breakdown. 23.78% were strong Democrats, 10.92% were weak Democrats, 11.83% were leaning Democrats, 10.66% were leaning Republicans, 10.09% were weak Republicans, 20.98% were strong Republicans, and 11.74% were pure independents. The ANES sample had a mean ideology of 4.09 and standard deviation of 1.67 on a scale from 1 (extremely liberal) to 7 (extremely conservative). It was 45.45% male, 53.74% female, and 0.81% NA; 72.92% white; with a mean age of 51.59 and standard deviation of 17.21. Lastly, 44.75% had a bachelor's degree or more.

Survey:
1. [pid] Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what? [Republican, Democrat, Independent, other, don't know]
   - 1a. If 1 or 2: Would you call yourself a strong [Republican / Democrat] or a not very strong [Republican / Democrat]? [strong, not strong]
   - 1b. If 3 or 4 or 5: Do you think of yourself as closer to the Republican or Democratic party? [Republican, Democrat, neither]
2. [ideology] We hear a lot of talk these days about liberals and conservatives. Here is a 7-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale, or haven't you thought much about this? [extremely liberal, liberal, slightly liberal, moderate, slightly conservative, conservative, extremely conservative, don't know]
3. [gender] What is your gender? [male, female, other]
4. [age] What is your age? [ ]
5. [race] What racial or ethnic group or groups best describes you? [white, black, Hispanic, Asian, Native American, other]
6. [education] What is the highest level of education that you have completed? [did not complete a high school degree, high school degree, some college, Associate's degree, Bachelor's degree, graduate or professional degree]
7. [self-monitoring 1] When you are with other people, how often do you put on a show to impress or entertain them? [always, most of the time, some of the time, once in a while, never]
8. [self-monitoring 2] When you are in a group of people, how often are you the center of attention? [always, most of the time, some of the time, once in a while, never]
9. [self-monitoring 3] How good or poor of an actor would you be? [excellent, good, fair, poor, very poor]

[Random assignment to one of three conditions]:

Condition 1 [control]: [nothing]
Condition 2 [public treatment]: **Just a reminder, the results from this study may be published.**
Condition 3 [private treatment]: **Just a reminder, your responses are completely private.**
Post-Treatment Questions:
[feeling thermometers]
I'd like to get your feelings toward the two national parties.
Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the party. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the party and that you don't care too much for that party. You would rate the party at the 50-degree mark if you don't feel particularly warm or cold toward the party.
[randomize party order]
1. How would you rate Democrats? [0 to 100 degrees]
2. How would you rate Republicans? [0 to 100 degrees]

Table C1. Polarization Responses by Treatment

Condition	N	Mean	SE
Public	171	46.97	2.54
Control	179	49.10	2.38
Private	170	47.67	2.83

Table shows respondents' polarization responses by treatment. None of these are statistically different from one another at $p<.05$.

Interactive Treatment Effects and Robustness Checks:
Note: 5 respondents did not answer at least 1 of the self-monitoring questions and thus were dropped from the following analyses.

Table C2. Polarization Responses by Treatment and Self-Monitoring

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Private	-	15.55	-6.31	9.24	19.84	2.94	10.70
	-	(6.35)	(4.45)	(4.09)	(7.99)	(11.65)	(6.00)
Public	0.14	-	-	-	-	-	-
	(5.98)	-	-	-	-	-	-
Self-Monitoring	-2.78	-2.78	4.17	1.39	-1.07	-4.80	-3.22
	(.75)	(.75)	(.58)	(.50)	(1.04)	(1.00)	(.80)
Private*Self-	-	-3.05	1.51	-1.53	-3.66	-1.28	-2.37
Monitoring	-	(1.20)	(.93)	(.71)	(1.71)	(1.75)	(1.19)
Public*Self-	-0.68	-	-	-	-	-	-
Monitoring	(1.12)	-	-	-	-	-	-
Democrat	-	-	-	-	-	-	12.19
	-	-	-	-	-	-	(5.21)
Strength	-	-	-	-	-	-	17.69
	-	-	-	-	-	-	(3.45)
Ideology 2	-	-	-	-	-	-	3.64
	-	-	-	-	-	-	(4.91)
Ideology 3	-	-	-	-	-	-	-1.87
	-	-	-	-	-	-	(6.18)
Ideology 4	-	-	-	-	-	-	-18.63

	-	-	-	-	-	-	(8.24)
Ideology 5	-	-	-	-	-	-	-6.79
	-	-	-	-	-	-	(6.23)
Ideology 6	-	-	-	-	-	-	10.24
	-	-	-	-	-	-	(6.97)
Ideology 7	-	-	-	-	-	-	18.94
	-	-	-	-	-	-	(7.87)
White	-	-	-	-	-	-	2.69
	-	-	-	-	-	-	(3.57)
Woman	-	-	-	-	-	-	-3.46
	-	-	-	-	-	-	(3.32)
Age	-	-	-	-	-	-	0.18
	-	-	-	-	-	-	(.16)
Education	-	-	-	-	-	-	-2.30
	-	-	-	-	-	-	(1.30)
Constant	63.33	63.33	5.63	68.96	57.01	72.48	12.95
	(4.29)	(4.29)	(2.80)	(2.98)	(5.42)	(7.10)	(13.32)

Table shows coefficients predicting reported polarization with robust standard errors in parentheses. Model 1 (OLS) = *self-monitoring*public* (N=348). Model 2 (OLS) = *self-monitoring*private* (N=344). Model 3 (OLS) = *self-monitoring*private*, just out-party feelings (N=344). Model 4 (OLS) = *self-monitoring*private*, just in-party feelings (N=344). Model 5 (OLS) = *self-monitoring*private*, just Democrats (N=218). Model 6 (OLS) = *self-monitoring*private*, just Republicans (N=126). Model 7 (OLS) = *self-monitoring*private* with controls (N=343—1 respondent did not answer the gender question and was dropped from this model). *Ideology* is in comparison to extremely liberal and values range from 1 to 7, extremely liberal to extremely conservative.

Quartile Analysis. Table C3 examines the potential non-linear moderating effect of self-monitoring with the desirable treatment (replicating the model used in Figure 4 [left], but interacting the treatment with quartiles of self-monitoring rather than a continuous self-monitoring variable). The table shows that there are no non-linear interaction effects except with the interaction between self-monitoring quartile 4 and the treatment.

Table C3. Modeling Self-Monitoring as Non-Linear

Variable	**Model 1**
Private	10.09
	(5.39)
Self-Monitoring Quartile 2	-8.88
	(6.51)
Self-Monitoring Quartile 3	-8.13
	(6.42)
Self-Monitoring Quartile 4	-19.43
	(6.07)
Private*Self-Monitoring Quartile 2	-7.28
	(8.86)
Private* Self-Monitoring Quartile 3	-10.42
	(8.81)
Private*Self-Monitoring Quartile 4	-28.76
	(9.87)

Constant	57.52
	(3.78)

Just like in the table above, this table shows coefficients predicting reported polarization with robust standard errors in parentheses (OLS). Model 1 models self-monitoring as non-linear, using four different dummy variables (1=below or equal to 25%, 2=between 25% and 50% [inclusive], 3=between 50% and 75% [inclusive], 4=above 75%) each interacted with the treatment. N=344.

Supplementary Material D: Study 4

Sample Information:

This study was funded by Time-Sharing Experiments for the Social Sciences (TESS; for more information, see https://www.tessexperiments.org/). It was thus fielded by National Opinion Research Center (NORC) at the University of Chicago using their AmeriSpeak Panel (for more information see https://amerispeak.norc.org/). NORC says the following about their samples: "The sample for a specific study is selected from the AmeriSpeak Panel using sampling strata based on age, race/ethnicity, education, and gender (48 strata in total). The size of the selected sample per sampling stratum is determined by the population distribution for each stratum. In addition, sample selection takes into account expected differential survey completion rates by demographic groups so that the set of panel members with a completed interview for a study is a representative sample of the target population. If panel household has one more than one active adult panel member, only one adult in the household is eligible for selection (random within-household sampling). Panelists selected for an AmeriSpeak study earlier in the business week are not eligible for sample selection until the following business week."

For this particular survey, 3,333 participants were recruited, 2,459 of which were randomly assigned to one of the three conditions used here. 1,895 of these were partisans. The final sample was 48.02% Democrat and 51.98% Republican (with 42.01% of these strong partisans and 57.99% of these weak partisans); with a mean of 3.91 and standard deviation of 1.80 from 1 (extremely liberal) to 7 (extremely conservative); 53.98% women and 46.02% men; 68.97% white and 31.03% either mixed or full minority; with 10.45% ages 18-29, 29.18% ages 30-44, 27.28% ages 45-59, and 33.09% ages 60 or more; and 35.77% received a at least a bachelors degree.

As a comparison, American National Election Studies (ANES) 2020 data has the following breakdown. 23.78% were strong Democrats, 10.92% were weak Democrats, 11.83% were leaning Democrats, 10.66% were leaning Republicans, 10.09% were weak Republicans, 20.98% were strong Republicans, and 11.74% were pure independents. The ANES sample had a mean ideology of 4.09 and standard deviation of 1.67 on a scale from 1 (extremely liberal) to 7 (extremely conservative). It was 45.45% male, 53.74% female, and 0.81% NA; 72.92% white; with a mean age of 51.59 and standard deviation of 17.21. Lastly, 44.75% had a bachelor's degree or more.

101 respondents said they "haven't thought much about it" to ideology and 2 skipped this question. In models with this variable, these 103 respondents were coded as moderate. 122 respondents skipped the *feeling thermometers* questions and were dropped from relevant analyses (this was not predicted by treatment assignment, p=.961). 9 respondents skipped the *trust* question and were dropped from relevant analyses (this was not predicted by treatment assignment, p=.297). Lastly, 25 respondents did not answer at least one of the three self-monitoring questions and were thus dropped from the interactive models. NORC also conducts quality control of their studies, removing speeders (who complete the survey in less than 1/3 the median duration) and those with high refusal rates (who skip or refuse more than 50% of the questions). 35 total cases were removed from the data for one or both of these reasons before being delivered to me.

Survey:

1. [self-monitoring 1] When you are with other people, how often do you put on a show to impress or entertain them? [always, most of the time, some of the time, once in a while, never]
2. [self-monitoring 2] When you are in a group of people, how often are you the center of attention? [always, most of the time, some of the time, once in a while, never]

3. [self-monitoring 3] How good or poor of an actor would you be? [excellent, good, fair, poor, very poor]

[Random assignment to one of four conditions – fourth condition for separate project and thus not included here]:

Condition 1 [control]: [nothing]

Condition 2 [public treatment]: Just a reminder, the results based on your responses may be published.

Condition 3 [private treatment]: Just a reminder, your responses are completely private.

Post-Treatment Questions:

[feeling thermometers]

I'd like to get your feelings toward the two national parties.

Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the party. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the party and that you don't care too much for that party. You would rate the party at the 50-degree mark if you don't feel particularly warm or cold toward the party.

[randomize party order]

1. How would you rate Democrats? [0 to 100 degrees]
2. How would you rate Republicans? [0 to 100 degrees]

[trust]

[randomize party order]

How much of the time do you think you can trust Democrats to do what is right for the country? [almost never / once in a while / about half the time / most of the time / almost always]

How much of the time do you think you can trust Republicans to do what is right for the country? [almost never / once in a while / about half the time / most of the time / almost always]

Table D1. Polarization Responses by Treatment

	Feeling Thermometers			***Trust***		
Condition	**N**	**Mean**	**SE**	**N**	**Mean**	**SE**
Public	606	49.74	1.42	644	1.81	0.05
Control	566	52.46	1.41	604	1.84	0.05
Private	601	50.39	1.47	638	1.82	0.05

Table shows Study 4 respondents' polarization responses by treatment. None of these are statistically different from one another at $p<.05$.

Interactive Treatment Effects and Robustness Checks:

Table D2. Polarization Responses by Treatment and Self-Monitoring (*Feeling Thermometers*)

Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Private	-	6.86	-3.26	3.87	11.99	1.59	4.91
	-	(4.76)	(2.94)	(2.73)	(6.85)	(6.58)	(4.58)
Public	1.98	-	-	-	-	-	-
	(4.36)	-	-	-	-	-	-
Self-Monitoring	0.42	0.42	0.25	0.69	0.73	0.11	0.64
	(.68)	(.68)	(.42)	(.41)	(.90)	(1.05)	(.62)
Private*Self-Monitoring	-	-2.17	1.12	-1.23	-3.74	-0.48	-1.66
	-	(1.06)	(.66)	(.60)	(1.45)	(1.54)	(1.04)

Public*Self-	-1.20	-	-	-	-	-	-
Monitoring	(.97)	-	-	-	-	-	-
Democrat	-	-	-	-	-	-	0.22
	-	-	-	-	-	-	(3.02)
Strength	-	-	-	-	-	-	18.34
	-	-	-	-	-	-	(2.04)
Ideology 2	-	-	-	-	-	-	-0.94
	-	-	-	-	-	-	(3.65)
Ideology 3	-	-	-	-	-	-	-2.76
	-	-	-	-	-	-	(4.18)
Ideology 4	-	-	-	-	-	-	-12.67
	-	-	-	-	-	-	(4.23)
Ideology 5	-	-	-	-	-	-	-10.62
	-	-	-	-	-	-	(5.21)
Ideology 6	-	-	-	-	-	-	0.12
	-	-	-	-	-	-	(4.41)
Ideology 7	-	-	-	-	-	-	11.91
	-	-	-	-	-	-	(5.18)
White	-	-	-	-	-	-	4.88
	-	-	-	-	-	-	(2.51)
Woman	-	-	-	-	-	-	0.25
	-	-	-	-	-	-	(1.96)
Age	-	-	-	-	-	-	1.66
	-	-	-	-	-	-	(.58)
Education	-	-	-	-	-	-	-0.51
	-	-	-	-	-	-	(.76)
Constant	50.93	50.93	20.70	71.51	48.96	52.74	3.82
	(3.09)	(3.09)	(1.88)	(1.90)	(4.46)	(4.35)	(11.66)

Table shows coefficients predicting reported polarization with robust standard errors in parentheses. Model 1 (OLS) = *self-monitoring*public* (N=1,160). Model 2 (OLS) = *self-monitoring*private* (N=1,152). Model 3 (OLS) = *self-monitoring*private*, just out-party feelings (N=1,183). Model 4 (OLS) = *self-monitoring*private*, just in-party feelings (N=1,179). Model 5 (OLS) = *self-monitoring*private*, just Democrats (N=553). Model 6 (OLS) = *self-monitoring*private*, just Republicans (N=599). Model 7 (OLS) = *self-monitoring*private* with controls (N=1,152). *Ideology* is in comparison to extremely liberal and values range from 1 to 7, extremely liberal to extremely conservative.

Table D3. Polarization Responses by Treatment and Self-Monitoring (*Trust*)

Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Private	-	0.34	-0.10	0.24	0.57	0.12	0.30
	-	(.17)	(.10)	(.11)	(.24)	(.23)	(.16)
Public	0.16	-	-	-	-	-	-
	(.16)	-	-	-	-	-	-
Self-Monitoring	0.02	0.02	0.02	0.04	0.03	0.00	0.03
	(.02)	(.02)	(.02)	(.02)	(.03)	(.04)	(.02)
Private*Self-	-	-0.09	0.03	-0.07	-0.15	-0.03	-0.08
Monitoring	-	(.04)	(.02)	(.03)	(.05)	(.05)	(.04)
Public*Self-	-0.05	-	-	-	-	-	-

Monitoring	(.04)	-	-	-	-	-	-
Democrat	-	-	-	-	-	-	0.14
	-	-	-	-	-	-	(.11)
Strength	-	-	-	-	-	-	0.77
	-	-	-	-	-	-	(.08)
Ideology 2	-	-	-	-	-	-	-0.13
	-	-	-	-	-	-	(.14)
Ideology 3	-	-	-	-	-	-	0.04
	-	-	-	-	-	-	(.15)
Ideology 4	-	-	-	-	-	-	-0.39
	-	-	-	-	-	-	(.16)
Ideology 5	-	-	-	-	-	-	-0.42
	-	-	-	-	-	-	(.20)
Ideology 6	-	-	-	-	-	-	0.08
	-	-	-	-	-	-	(.16)
Ideology 7	-	-	-	-	-	-	0.41
	-	-	-	-	-	-	(.18)
White	-	-	-	-	-	-	0.15
	-	-	-	-	-	-	(.09)
Woman	-	-	-	-	-	-	0.13
	-	-	-	-	-	-	(.07)
Age	-	-	-	-	-	-	0.09
	-	-	-	-	-	-	(.02)
Education	-	-	-	-	-	-	-0.03
	-	-	-	-	-	-	(.03)
Constant	1.78	1.78	1.63	3.41	1.76	1.81	-0.29
	(.11)	(.11)	(.07)	(.08)	(.16)	(.16)	(.45)

Table shows coefficients predicting reported polarization with robust standard errors in parentheses. Model 1 (OLS) = *self-monitoring*public* (N=1,235). Model 2 (OLS) = *self-monitoring*private* (N=1,225). Model 3 (OLS) = *self-monitoring*private*, just out-party feelings (N=1,226). Model 4 (OLS) = *self-monitoring*private*, just in-party feelings (N=1,228). Model 5 (OLS) = *self-monitoring*private*, just Democrats (N=591). Model 6 (OLS) = *self-monitoring*private*, just Republicans (N=634). Model 7 (OLS) = *self-monitoring*private* with controls (N=1,225). *Ideology* is in comparison to extremely liberal and values range from 1 to 7, extremely liberal to extremely conservative.

Quartile Analysis. Table D4 examines the potential non-linear moderating effect of self-monitoring with the desirable treatment (replicating the models used in Figure 4 [middle and right], but interacting the treatment with quartiles of self-monitoring rather than a continuous self-monitoring variable). The table shows that there are no non-linear interaction effects with feeling thermometer affective polarization as the dependent variable (Model 1), and the only non-linear interaction effect with trust affective polarization as the dependent variable (Model 2) is with the interaction between self-monitoring quartile 4 and the treatment.

Table D4. Modeling Self-Monitoring as Non-Linear

Variable	**Model 1**	**Model 2**
Private	0.60	0.18
	(4.56)	(.16)

Self-Monitoring Quartile 2	3.40	0.24
	(3.72)	(.13)
Self-Monitoring Quartile 3	4.44	0.05
	(4.40)	(.16)
Self-Monitoring Quartile 4	1.88	0.16
	(4.69)	(.18)
Private*Self-Monitoring Quartile 2	0.43	-0.25
	(5.48)	(.19)
Private* Self-Monitoring Quartile 3	-5.12	-0.15
	(6.15)	(.22)
Private*Self-Monitoring Quartile 4	-10.95	-0.61
	(7.57)	(.27)
Constant	50.06	1.72
	(3.03)	(.11)

Just like in the table above, this table shows coefficients predicting reported polarization with robust standard errors in parentheses (OLS). Models 1 and 2 model self-monitoring as non-linear, using four different dummy variables (1=below or equal to 25%, 2=between 25% and 50% [inclusive], 3=between 50% and 75% [inclusive], 4=above 75%) each interacted with the treatment. Model 1 predicts affective polarization as measured by feeling thermometers (N=1,152) and Model 2 predicts affective polarization as measured by trust (N=1,225).

Supplementary Material E: Post-Hoc Manipulation Check (January 2023)

Study Goals and Design:

Studies 3 and 4 utilized two treatments aiming to change respondents' perception of privacy. Respondents were assigned to either no prompt (control), to be reminded that "Just a reminder, the results from this study may be published" [Study 3] or "Just a reminder, the results based on your responses may be published" [Study 4] (public), or to be reminded that "Just a reminder, your responses are completely private" (private). They were then asked questions to measure affective polarization.

One drawback to this design is that it is unclear if the treatments affected respondents' perceptions of survey privacy or if they affected respondents' perceptions of survey accountability. Changes in survey accountability would mean that respondents believed they might have to justify their choices or responses to others (see Krosnick 1991; Tetlock 1983; and Tetlock and Kim 1987), and this anticipation of having to justify or explain their responses could shift responses to questions about affective polarization. To examine which (if either) of these were shifted by the privacy treatments, I ran a post-hoc manipulation check with a sample from Prolific (N=450, see below), where I randomly assigned respondents to the same control, public, and private conditions from Study 3 and Study 4 (using the public treatment wording from Study 4).

Respondents were then asked 2 questions to measure privacy perceptions, 1 question to measure privacy concerns, and 2 questions to measure perceived response accountability. They were randomly assigned to either receive the privacy or the accountability questions first, in order to avoid question ordering effects. The first privacy question was taken from Mueller et al. (2014), who asked respondents to answer on a 5-point Likert scale: "How do you rate the anonymity of this survey?" Respondents were given the options of not at all anonymous, not very anonymous, somewhat anonymous, anonymous, and completely anonymous. The second question focused more specifically on privacy, asking respondents "On the following scale, how private do you believe your responses are?" with a sliding scale from 0 (completely public) to 100 (completely private). The third question asked about concerns with privacy and was adapted from Jenson, Potts, and Jenson (2005). It said: "I am concerned about my privacy in this survey" (note: Jenson et al. 2005 said "online" instead of "in this survey"), with 5 response options from strongly agree to strongly disagree.

The two questions to measure accountability were written in a way to encompass research from Krosnick (1991), Tetlock (1983), and Tetlock and Kim (1987). Krosnick (1991), for example, argued that accountability could reduce satisficing in surveys because respondents would believe they have to justify their responses to others. Relatedly, Tetlock and Kim (1987) manipulated perceived accountability by telling respondents that researchers wanted to interview them to explore "the types of information that people use to form impressions of others" and then asked them to allow taping of the interview "for future data analysis purposes." Since I was measuring rather than manipulating perceived accountability, I asked two questions aimed to measure respondents' belief that they would have to justify and/or explain their survey responses to others. The first question asked: "To what extent do you believe you will have to justify your survey responses to others" and the second asked: "To what extent do you believe you will have to explain your survey responses to others." Respondents were given response options of: not at all, somewhat, a fair amount, and completely.

Sample Information:
Prolific is an online platform where respondents opt-in to take surveys and get paid for their participation (for more information see https://www.prolific.co). Although Prolific produces a quality, attentive sample, I also included a Captcha verification at the beginning of the survey to prohibit bots and prevented multiple submissions to avoid "ballot stuffing."

For this particular survey, I recruited 450 US adults. The sample was 73.54% Democrat and 26.46% Republican (with 34.67% of these strong partisans, 30.22% of these weak partisans, 19.11% leaning partisans, and 16% pure independents); with a mean of 3.15 and standard deviation of 1.70 from 1 (extremely liberal) to 7 (extremely conservative); 50% women, 48.89% men, and 1.11% other; 70.89% white and 29.11% either mixed or full minority; a mean age of 38.10 and standard deviation of 13.76; and 51.34% received at least a bachelors college.

As a comparison, American National Election Studies (ANES) 2020 data has the following breakdown. 23.78% were strong Democrats, 10.92% were weak Democrats, 11.83% were leaning Democrats, 10.66% were leaning Republicans, 10.09% were weak Republicans, 20.98% were strong Republicans, and 11.74% were pure independents. The ANES sample had a mean ideology of 4.09 and standard deviation of 1.67 on a scale from 1 (extremely liberal) to 7 (extremely conservative). It was 45.45% male, 53.74% female, and 0.81% NA; 72.92% white; with a mean age of 51.59 and standard deviation of 17.21. Lastly, 44.75% had a bachelor's degree or more.

Survey:
[pre-treatment questions]

1. [gender] What is your gender? [man / woman / other]
2. [age] What is your age? [ ]
3. [race] What racial or ethnic group or groups best describes you? [white / black / Hispanic / Asian / Native American / other (please specify):___]
4. [education] What is the highest level of education that you have completed? [did not complete a high school degree / high school degree / some college / Associate's degree / Bachelor's degree / graduate or professional degree]
5. [self-monitoring 1] When you're with other people, how often do you put on a show to impress or entertain them? [always, most of the time, some of the time, once in a while, never]
6. [self-monitoring 2] When you're in a group of people, how often are you the center of attention? [always, most of the time, some of the time, once in a while, never]
7. [self-monitoring 3] How good or poor of an actor would you be? [excellent, good, fair, poor, very poor]
8. [PID] Generally speaking, do you think of yourself as a Republican, a Democrat, an Independent, or what? [Republican / Democrat / independent / something else [_____]]
    a. [if Democrat or Republican] Would you call yourself a strong [Democrat/Republican] or a not very strong [Democrat/Republican]? [strong [Democrat/Republican] / not very strong [Democrat/Republican]]
    b. [if independent or something else] Do you think of yourself as closer to the Republican Party or the Democratic Party? [closer to the Republican Party / closer to the Democratic Party / neither]
9. [ideology] We hear a lot of talk these days about liberals and conservatives. Here is a 7-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale, or haven't you thought much about this? [extremely liberal / liberal / slightly liberal / moderate / slightly

conservative / conservative / extremely conservative / don't know]

10. [interest] Some people don't pay much attention to political news. How about you? Would you say that you are very much interested, somewhat interested, or not interested at all interested in political news? [not at all interested / somewhat interested / very much interested]
11. [media] During a typical week, how many days do you watch, read, or listen to political news on the following medium: (the internet (including online newspapers), the TV, print newspapers, the radio) [0 days / 1 day / 2 days / 3 days / 4 days / 5 days / 6 days / 7 days]
12. [discuss] During a typical week, how many days do you discuss politics with your family and/or friends? [0 days / 1 day / 2 days / 3 days / 4 days / 5 days / 6 days / 7 days]

[random assignment to one of three conditions]:

a. [control]: [nothing]
b. [public treatment]: Just a reminder, the results based on your responses may be published.
c. [private treatment]: Just a reminder, your responses are completely private.

[post-hoc manipulation checks: randomly assign to privacy or accountability questions first]

13. [privacy 1] **How would you rate the anonymity of this survey? [not at all anonymous / not very anonymous / somewhat anonymous / anonymous / completely anonymous]**
14. [privacy 2] On the following scale, how private do you believe your responses are? [0=completely public → 50=neither public nor private → 100=completely private]
15. [privacy concern] **I am concerned about my privacy in this survey [strongly agree / agree / neither agree nor disagree / disagree / strongly disagree]**
16. [accountability 1] **To what extent do you believe you will have to justify your survey responses to others? [not at all / somewhat / a fair amount / completely]**
17. **[accountability 2] To what extent do you believe you will have to explain your survey responses to others? [not at all / somewhat / a fair amount / completely]**

Survey Findings:

I then ran 10 OLS models with robust standard errors, predicting responses to each of the 5 questions by a private treatment dummy (1=private treatment, 0=control) and then predicting responses to each of the 5 questions by a *public* treatment dummy (1=public treatment, 0=control).

As shown in Table E1 and Table E2, I find that in comparison to those in the control group, those in the privacy group reported believing the survey was more anonymous (p=.023) and their responses more private (p=.093). Indeed, standardizing each question (0-1) and merging them together (0-2), I find that the private treatment increased perceptions of survey privacy (coefficient: 0.09, p=.030). Those in the privacy group were not, however, less *concerned* about the privacy of the survey (p=.682)—perhaps because people were largely unconcerned anyways (mean=4.12, SD=1.01 on scale from 1 [strongly agree that they are concerned] to 5 [strongly disagree that they are concerned])—nor did they have any different responses about the *accountability* of their responses (p=.523, p=.914). Further, while (again compared to the control) the public treatment did not influence responses about perceptions of anonymity (p=.100), privacy concerns (p=.954) or accountability (p=.613, p=.921), it *did* make respondents believe the survey was less private (p=.021). These findings suggest that indeed, the privacy treatment increased respondents' perception of privacy, and that is why in Study 3 and Study 4 it also shaped responses about affective polarization.

Table E1. Privacy Responses by Treatment

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Private	0.21	4.37	0.05	-	-	-
	(.09)	(2.59)	(.11)	-	-	-
Public	-	-	-	-0.00	-7.00	-0.01
	-	-	-	(.10)	(3.02)	(.12)
Constant	4.01	75.33	4.11	4.01	75.33	4.11
	(.07)	(1.85)	(.08)	(.07)	(1.85)	(.08)

Table shows coefficients predicting privacy responses with robust standard errors in parentheses. Model 1 (OLS): *privacy 1 = private treatment* (N=300). Model 2 (OLS): *privacy 2 = private treatment* (N=297; 1 respondent in the private treatment and 2 in the control condition did not answer the second privacy question). Model 3 (OLS): *privacy concern = private treatment* (N=300). Model 4 (OLS): *privacy 1 = public treatment* (N=300). Model 5 (OLS): *privacy 2 = public treatment* (N=294; 4 respondents in the public condition and 2 in the control condition did not answer the second privacy question). Model 6 (OLS): *privacy concern = public treatment* (N=300).

Table E2. Accountability Responses by Treatment

Variable	Model 1	Model 2	Model 3	Model 4
Private	0.05	-0.01	-	-
	(.07)	(.06)	-	-
Public	-	-	0.03	0.01
	-	-	(.07)	(.07)
Constant	1.21	1.21	1.21	1.21
	(.05)	(.05)	(.05)	(.05)

Table shows coefficients predicting accountability responses with robust standard errors in parentheses. Model 1 (OLS): accountability 1 = private treatment (N=300). Model 2 (OLS): accountability 2 = private treatment (N=300). Model 3 (OLS): accountability 1 = public treatment (N=299; 1 respondent in the public condition did not answer the first accountability question). Model 4 (OLS): accountability 2 = public treatment (N=300).

Supplementary Material F: Self-Monitoring

"I'm not good at breaking ties. I'm good at being in the majority after the majority has already voted so that I can see which way things are going. That's my thing." –Truly, an (assumed) high-self-monitor from the show Bunheads

Self-Monitoring Cronbach's Alphas:
Study 1: .61; *Study 2:* .70; *Study 3:* .83, and *Study 4:* .66.

Predicting Self-Monitoring:
To examine the correlates of self-monitoring, I run OLS regressions in each study predicting self-monitoring with the controls from the previous analyses (partisanship, partisan strength, ideology, race, gender, age, and education) and robust standard errors. I find that in Studies 1, 2, 3, and 4, age is negatively correlated with self-monitoring (-0.04, $p<.001$; -0.03, $p<.001$; -0.06, $p<.001$; and -0.27, $p<.001$; respectively), whereby younger people are on average higher self-monitors. This aligns with Berinsky (2004) and Berinsky and Lavine (2012). Further, in Studies 1, 2, 3, and 4, gender is correlated with self-monitoring (0.33, $p=.024$; 0.90, $p<.001$; 0.85, $p=.001$; and 0.56, $p<.001$; respectively), whereby men are on average higher self-monitors than women. This also aligns with Berinsky (2004). Lastly, in Studies 1, 2, and 3, education is positively correlated with self-monitoring (0.15, $p=.005$; 0.25, $p<.001$; and 0.42, $p<.001$; respectively), whereby those who are more highly educated are higher self-monitors on average—this was not replicated in Study 4.

In Study 3, Republicans (-0.93, $p=.015$) and strong partisans (1.61, $p<.001$) were higher self-monitors, but this was not replicated in the other studies. In Study 4, being white was also negatively correlate with self-monitoring (-0.53, $p<.001$), such that mixed or full minorities were higher self-monitors. This is in line with Berinsky (2004) but was only marginally significant in Study 3 (-0.48, $p=.073$) and did not replicate in the other two studies.

Supplementary Material References

Berinsky, Adam J. 2004. "Can We Talk? Self-Presentation and the Survey Response." *Political Psychology* 25(4): 643-659.

Berinsky, Adam J. and Howard Lavine. 2012. "Self-Monitoring and Political Attitudes." *Improving Public Opinion Surveys: Interdisciplinary Innovation and the American National Election Studies*: 27-45.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351-368.

Jensen, Carlos, Colin Potts, and Christian Jensen. 2005. "Privacy Practices of Internet Users: Self-Reports versus Observed Behavior." *International Journal of Human-Computer Studies* 63:203–27.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213–36.

Mueller, Karsten, Tammo Straatman, Kate Hattrup, and Marco Jochum. 2014. "Effects of Personalized versus Generic Implementation of an Intra-Organizational Online Survey on Psychological Anonymity and Response Behavior: A Field Experiment." *Journal of Business and Psychology* 29:169–81.

Tetlock, Philip E. 1983. "Accountability and Complexity of Thought." *Journal of Personality and Social Psychology* 45:74-83.

Tetlock, Philip E. and Jae Il Kim. 1987. "Accountability and Judgment Processes in a Personality Prediction Task." *Journal of Personality and Social Psychology* 52(4):700-709.