Partisan Social Pressure and Affective Polarization

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

***Abstract*.** American politics today is affectively polarized—partisans report disliking and distrusting out-partisans while liking and trusting in-partisans. I examine if this climate could encourage partisans to report higher levels of affective polarization. I test this with two survey experiments (one with a convenience sample and one with a nationally representative sample from NORC's AmeriSpeak panel) and a set of analyses of 2008 American National Election Studies (ANES) data. My findings demonstrate that there is partisan social pressure to report affective polarization and that this could inflate survey responses, suggesting implications for the measurement of polarization, greater nuance in our understanding of polarization, and the potential for a "snowball effect" of political climates—where a climate of polarization could beget more polarization. Future work is needed for a broader understanding of how social context shapes expressions of partisanship.

Word Count: 9,892

Key Words: social influence; affective polarization; partisanship

Introduction

American politics today is affectively polarized—partisans dislike and distrust out-partisans while liking and trusting in-partisans (e.g., Iyengar and Westwood 2015; Mason 2015, 2018). I examine if this climate could encourage partisans to report higher levels of animosity towards out-partisans and favorability toward in-partisans, arguing that in today's political climate there is partisan social pressure to do so and that this can motivate partisans to inflate their reports of in-party like and out-party dislike (i.e., the measures that create the affective polarization construct—heretofore just referred to as affective polarization). I argue that this partisan social pressure to report affective polarization should influence self-reports of these measures because people have a desire to impress others (Goffman 1967; Holtgraves 1992)—thus, this pressure should mean that at least some partisans acquiesce and exaggerate their reports of affective polarization.

Indeed, social contexts can influence individuals' reported political attitudes, values, behaviors, and expressions of partisanship (Carlson and Settle 2016; Connors 2019; Daoust et al. 2020; Klar 2014; Klar and Krupnikov 2016). The power of social contexts often lies in people's desire to impress others and thus adjust themselves to do so (Goffman 1967; Holtgraves 1992). This type of morphing behavior, driven by self-presentation desires, varies by individual and can be tracked with the self-monitoring trait (Berinsky 2004; Berinsky and Lavine 2012; Gangestad and Snyder 2000; Snyder 1974)—an independent variable of interest in this piece. Of course, it may also be the case that partisan social pressure's influence on reports of affective polarization is a story about partisan social identity (see Green et al. 2002; Tajfel and Turner 1979) rather than self-presentation desires. In my analyses I acknowledge this possibility but find no support for this alternative mechanism.

I test my argument with three sets of analyses: 1) a survey experiment to examine if there is indeed partisan social pressure to report affective polarization; 2) a survey experiment conducted with a nationally representative sample that attempts to alter reports of affective polarization by varying the privacy respondents perceive they have—and to see if this effect is moderated by self-monitoring; and 3) observational analysis of the 2008 American National Election Studies (ANES) data to see if we see the relationships we would expect to see reflected in the real world. Together, the analyses demonstrate: 1) that there is strong partisan social pressure to report affective polarization; 2) that this is likely influencing some self-reports of affective polarization; and 3) that this is often moderated by self-monitoring, further demonstrating that partisan social pressure can influence these self-reports because of the desire to impress others (Goffman 1967; Holtgraves 1992).

My findings add important nuance to earlier findings on partisanship and self-presentation desires (Klar and Krupnikov 2016). While previous research finds that in today's polarized climate, self-presentation desires can lead some to retreat into political independence and to avoid political activity (see Klar and Krupnikov 2016), I argue that among those who are *willing* to identify as a partisan, today's polarized climate and self-presentation desires instead can lead to greater reports of affective polarization. Although examining which variables lead people to these two groups is beyond the scope of this manuscript, it is likely the case that those who are willing to identify as partisans receive different social demands than those who refuse to do so, potentially leading to these divergent trends. My argument, then, adds a broader perspective to how self-presentation desires can influence expressed partisanship (or lack thereof)—bringing attention to those who respond to a culture of polarization by exaggerating their own.

The implications of my results are two-fold. First, the findings speak to the measurement

and understanding of an important political construct—affective polarization—as well as suggests multiple ways to decrease self-reports of affective polarization to their perhaps more truthful levels (although it remains to be seen what a "true" report of affective polarization would entail). Second, the findings suggest that political climates can have a snowball effect—that (for example) a climate of political polarization can create partisan social pressure to report polarization, thus leading to both overreports of polarization as well as reactions *of* increased polarization *to* those overreports (see Levendusky and Malhotra, 2015; 2016 who find that when partisans hear that out-partisans dislike them they react by reciprocating that dislike). On the other hand, this suggests that a climate of partisan compromise and congeniality could lead to even *more* compromise and congeniality, if that is what partisans perceive to be socially desirable.

Polarization & Social Influence

Affective Polarization. We have many different conceptualizations of polarization, but here I focus on affective polarization—partisans liking in-partisans and disliking out-partisans (see Iyengar et al. 2019 for review). Note that affective polarization need not be rooted in ideology: people do not necessarily dislike opposing partisans while liking their own because they disagree on issues—they abide by this polarization *even when they agree with each other* (Iyengar and Westwood 2015; Iyengar et al. 2012; Mason 2015).[1] As Mason (2015) explains, America "agrees on many things but is bitterly divided nonetheless" (pg. 128). In sum, partisanship is a social identification and by identifying with one's own party, the other party becomes a disliked out-group that is perceived to be highly dissimilar to oneself (Green et al. 2002; Huddy et al. 2015; Mason 2015).

Affective polarization has increased consistently and tremendously over the past 40 years (Iyengar, Sood, and Lelkes 2012; Iyengar and Westwood 2015; Mason 2015, 2018; Westwood, Peterson, and Lelkes 2018; Iyengar et al. 2019). These high levels of polarization can 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). As Iyengar and Westwood (2015) explain, "Partisans discriminate against opposing partisans, doing so to a degree that exceeds discrimination based on race" (pg. 690).

This era's growing and intense affective polarization has received large amounts of media attention. Levendusky (2009), for example, finds that media coverage of polarization has increased dramatically since the 1990s—a trend that has been replicated more recently (Klar and Krupnikov 2016; see also Robison and Mullinix 2016). Further, McGregor (2019) has demonstrated journalists' use of tweets as judgments of public opinion—something that inevitably gives extremists' (including the polarized public) a louder voice in the media. Similarly, Settle (2018) argues that the features of social media—and the way we use it—make identity salient and lead to partisan biases and stereotypes, further fueling polarization.

These findings illustrate how both the news media and social media emphasize polarization, and thus why people would believe the partisan public to be polarized—and in turn perceive partisan pressure to be polarized. Indeed, Levendusky and Malhotra (2015, 2016) show how this focus on polarization by the media influences the public's reports of polarization. As they explain, "The mass media depict polarization as widespread, occurring across many issues, and

[1] Although see also Orr and Huber (2019) who find that issue polarization can further fuel affective polarization and Abramowitz and Webster (2016) who examine correlations between these two types of polarization.

accompanied by incivility and dislike of the opposition, not simply issue-based disagreement…When citizens read or watch stories about polarized politics, they observe individuals who are divided and take extreme positions, who eschew compromise, and display incivility toward one another" (pg. 286).

Thus, even though not all partisans are polarized, the public perceives much more polarization than actually exists (Ahler 2014; Ahler and Sood 2018; Levendusky and Malhotra 2015). Perhaps because the media focus on those from the extremes, the public believes that the vast majority of Democrats and Republicans *are* affectively polarized. This descriptive norm of partisans being affectively polarized likely shapes the perception of partisan social pressure to also be affectively polarized: the perception that partisans are polarized (i.e., the descriptive norm) suggests to partisans that they too should be polarized (i.e., the injunctive norm; see, for example, Cialdini et al. 1990; Rimal and Real 2005).

Social Influence. People want to present themselves well to others (Cosmides and Tooby 1992; Goffman 1967) and work at this type of "self-presentation" almost constantly (Holtgraves 1992). Because people practice politics in a social world, this social motivation naturally extends to how ordinary people practice politics. The social, in essence, is in the political (see Sinclair 2012). Thus, to present oneself well, people misreport political attitudes (Kuran 1997; Zaller and Feldman 1992); conform their political views and behaviors to follow elites (Zaller 1992), to fit into their group (Huckfeldt et al. 2013; Mutz 1998), or to follow general norms (Cialdini et al. 1990); change how they describe their partisan identities (Klar and Krupnikov 2016); suppress unpopular or contentious opinions (Carlson and Settle 2016; Klar 2014); and endorse their party's political values (Connors 2019). They are, to borrow terminology from Carlson and Settle (2016), "political chameleons."

People's tendency to morph themselves to fit into social contexts suggests that when partisans believe that there is partisan social pressure to be affectively polarized, they will themselves act affectively polarized—reporting higher levels of out-party antipathy and in-party favorability. However, we would not expect everyone to engage equally in this chameleon-like behavior. Indeed, the tendency to misrepresent oneself to fit into social contexts varies based on individuals' level of self-monitoring (Berinsky 2004; Berinsky and Lavine 2012; Gangestad and Snyder 2000; Snyder 1974). At its core, self-monitoring measures one's desire to impress others, where higher levels of self-monitoring indicate greater tendency to change oneself to appease others.[2] That is, as individuals' self-presentation desires increase, their likelihood of adapting themselves to fit into social contexts also increases. This trait is continuous—from low to high self-monitoring—where those at the lowest end are the *least* likely to misrepresent themselves to impress others and those at the highest end are the *most* likely to do so.

Within the context of my argument—that there is partisan social pressure to be affectively polarized and that this influences self-reports of affective polarization because people have a desire to impress others—the likelihood of reporting affective polarization to acquiesce to partisan social pressure should increase with higher levels of self-monitoring. For partisans who are especially attuned to social pressure and who respond by acquiescing (i.e., those who are higher on the self-monitoring trait), self-reports of affective polarization should be shaped by what is desirable. Thus, in our current climate—if indeed partisans perceive partisan social pressure to report affective polarization—affective polarization should be positively correlated with self-monitoring.

[2] Like a willow, high self-monitors bend to the wind.

Similarly, when the perception of what is desirable regarding affective polarization shifts, partisans who are high in self-monitoring should be most likely to follow that shift and change their reports of affective polarization. From the perspective of a high self-monitor—whose goal is to look good to others—if reporting affective polarization is socially desirable, this is what they will do. Thus, if there *is* social pressure to report affective polarization, then we should see self-monitoring and affective polarization positively correlated (controlling for other important variables). When that same high self-monitor has privacy—and thus has no audience to look good to—these high self-reports of affective polarization should then decrease. Lastly, that same high self-monitor, in a (theoretical) context where affective polarization is socially *un*desirable, should report *less* polarization.

My Argument. My argument builds on previous research that suggests that information about polarization can affect people's responses to survey questions. Levendusky and Malhotra (2016), for example, examine the effects of media coverage of polarization on political attitudes. Among other things, they find that media coverage of polarization increases affective polarization.[3]

My argument extends these findings, proposing that one important consequence of the polarized state of America is that it encourages people to report affective polarization because there is a partisan social pressure to do so.[4] The informational context of polarization—including the exaggerated accounts of polarization by the media and overestimations of polarization by the public—should (first) suggest to partisans that there is social pressure to be affectively polarized (or at least report being so) and (second) motivate partisans (especially those who care to fit into social contexts) to report disliking out-partisans while liking in-partisans. In surveys, this segment of the public may seem to be those who indeed *do* hate out-partisans and love in-partisans, but such self-reports may be as much a function of social pressure as of genuine animus.[5]

My argument that there is partisan social pressure to report affective polarization and that this influences partisans' self-reports of affective polarization because people have a desire to impress others leads to three main expectations—all looking only at partisans (i.e., those who choose "Republican" or "Democrat" when first asked about their partisanship; see Druckman and Levendusky 2019 for a similar approach). First, that partisans perceive social pressure to report affective polarization. Second, that altering the privacy respondents perceive they have then changes self-reports of affective polarization, where reports of affective polarization should be highest in the most public settings and lowest in the most private settings—and that this effect is moderated by self-monitoring, where the change in privacy should be increasingly effective at changing reports of affective polarization as self-monitoring increases. And third, that we see this play out in the real world—that, since those who are especially attuned to social contexts (i.e.,

---

[3] This same condition, however, also increases negative emotions towards *in-partisans* (or, towards the exemplar of an in-partisan), but to a lesser degree. This makes sense when we consider findings that some are turned off by polarization and thus react negatively to *both* in-partisans and out-partisans (see Klar, Krupnikov, and Ryan 2018). Yet there are others in their study who *only* react negatively towards *out-partisans*—this would fit with my argument that some are encouraged by reports of polarization to exaggerate their *own* levels of polarization. Note, though, that Levendusky and Malhotra (2016) suggest a different mechanism than the one proposed here, arguing that coverage of polarization is unflattering for partisans and makes out-partisans seem less similar to the respondent. Unfortunately, their research cannot distinguish between the mechanism they suggest and the mechanism proposed here.

[4] For a comprehensive discussion of other consequences of polarization, see Iyengar et al. 2018.

[5] Note that disentangling these self-reports from behavior (or "genuine" animus) is beyond the scope of this manuscript (but see Discussion and Conclusion section).

those who are high self-monitors) would be most influenced by partisan social pressure, reporting of affective polarization should increase with higher levels of self-monitoring (controlling for other important variables).

I examine my argument's expectations with three sets of analyses. First, I run a survey experiment on Prolific (N=920) that tests whether there is partisan social pressure to report affective polarization. Second, I run another survey experiment with a nationally representative sample from NORC's AmeriSpeak panel (N=1,771) to attempt to alter reports of affective polarization by varying the privacy respondents perceive they have (and, again, to examine if this effect is moderated by self-monitoring). And third, I conduct a set of analyses with nationally representative 2008 American National Election Studies (ANES) data to see if we see the relationships we would expect to see—a positive correlation between self-monitoring and affective polarization, controlling for important variables—reflected in the real world.

Testing my argument with both observational and experimental data is ideal as I can confront causal inference with random assignment and test my proposed mechanism (self-monitoring; i.e., self-presentation desires), as well as examine if the relationships we would expect to see are reflected in nationally representative observational data (Shadish, Cook, and Campbell 2002). Further, across the three sets of analyses, I use three different measures of affective polarization, which serves as an additional check on the results and broadens the applications of my findings. Note, again, that given the argument motivating this research, only those who associate with a party are included in all three sets of analyses (see Druckman and Levendusky 2019 for a similar approach).

Alternative Story: Partisan Social Identity. I argue that there is social pressure *among* partisans to report affective polarization, but that this is driven by broader self-presentation desires rather than partisan identity. We know, however, that partisanship is a strong social identity (Green et al. 2002; Huddy et al. 2015; Mason 2015), and that it is thus possible that the story here is more about partisan social identity than it is about self-presentation desires (although see West and Iyengar 2020 who temper claims that affective polarization is largely a result of social identity). Indeed, partisan strength (a proxy for partisan identity) is a predictor of affective polarization (see Appendix D, Table D.1). Acknowledging this possibility, I test for partisan strength as a moderator between the various treatments and measures of affective polarization. I do not find support for this alternative story.

Survey Experiment 1 (Prolific, N=920)

The first step is examining if partisans perceive social pressure to report affective polarization. For partisan social pressure to influence reports of affective polarization, partisans must first perceive it. Thus, I ran a survey experiment with partisans on Prolific, an online convenience sample (N=920).[6] I randomly assigned participants to be asked about either impressing or disappointing in-partisans. Republicans were told: "Please answer the following 2 questions as you think a Republican wanting to [impress / disappoint] other Republicans would" and Democrats were told:

[6] The sample was 80.17% Democrat and 19.83% Republican (with 60.76% of these strong partisans and 39.24% of these weak partisans); with a mean of 5.24 and standard deviation of 1.78 from extremely conservative (1) to extremely liberal (7), 56.03% women and 43.97% men, a mean of age 33.24 with a standard deviation of 33.24, and 69.35% white and 30.65% either mixed or full minority.

"Please answer the following 2 questions as you think a Democrat wanting to [impress / disappoint] other Democrats would."

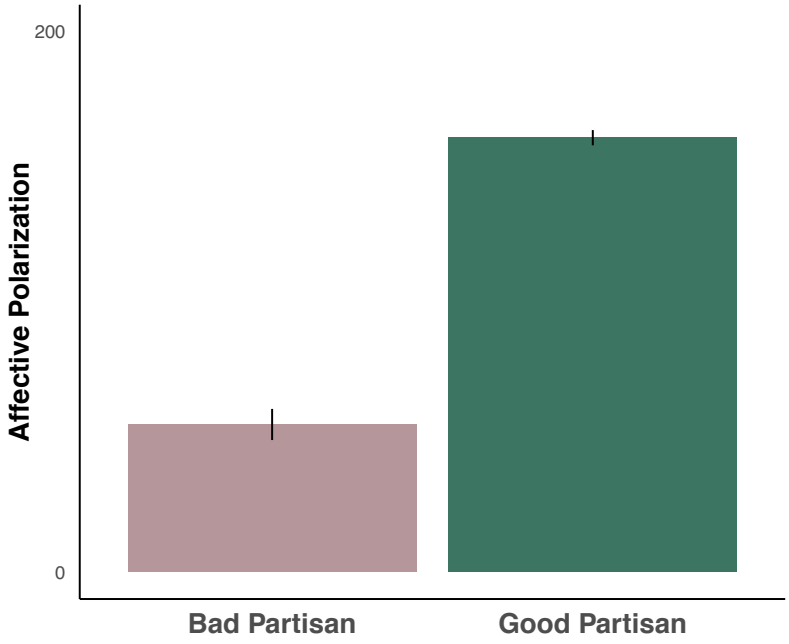
Then, all respondents were asked two questions to measure affective polarization. These questions asked respondents to place each group of partisans on a scale from 0 to 100, where ratings from 50 to 100 indicate favorable feelings and ratings from 0 to 50 indicate unfavorable feelings. Affective polarization is then measured as the difference between feeling thermometers toward in-partisans and feeling thermometers toward out-partisans, where greater numbers indicate greater polarization. This feeling thermometer measure is one of the most often used measures of affective polarization (see Iyengar et al. 2019 for discussion).

Randomly assigning these "impress" versus "disappoint" prompts allows me to compare two important groups: partisans wanting to impress other in-partisans versus partisans wanting to disappoint other in-partisans. A gap between these two would suggest that there is partisan social pressure to report affective polarization (see Klar and Krupnikov 2016 for a similar approach to measuring social desirability).

Results. The results show that indeed partisans feel partisan social pressure to report affective polarization. Respondents said that partisans wanting to impress in-partisans would report high levels of affective polarization (160.76 on a scale from 0 to 200) and partisans wanting to *disappoint* inpartisans would report low levels of polarization (54.61 on a scale from 0 to 200). Note that this difference between the impress ("good partisan") and disappoint ("bad partisan") prompts is 106.16 points—more than half the 200-point scale from least to most affectively polarized—and that this difference is highly significant (p=.000; see Figure 1).

Further, the gap between good and bad partisans is driven by *both* in-party or out-party feelings—the gap between good and bad for in-party feelings is 53 and for out-party feelings is 54—as well as both Democrats and Republicans (although the gap for Democrats is a bit larger)—the gap for Democrats is 108.55 and for Republicans is 98.55. Thus, the gap we see between the two conditions demonstrates that partisans indeed perceive partisan social pressure to report affective polarization, a necessary component to my argument.

Figure 1. Reported Affective Polarization by "Impress" versus "Disappoint" Prompts



Partisans (N=920) were asked to report affective polarization based on wanting to disappoint ("bad partisan") or impress ("good partisan") in-partisans. The difference between the two conditions is 106.16 points (p=.000), which is more than half the 200-point scale. Lines represent 95% confidence intervals.

Survey Experiment 2 (NORC, N=1,771)

Next I test my second expectation: that altering the privacy respondents perceive they have also changes self-reports of affective polarization, where reports of affective polarization should be highest in the most public settings and lowest in the most private settings—and that this is also moderated by self-monitoring (or individual susceptibility to social pressure), where the change in privacy should be increasingly effective at changing reports of affective polarization as self-monitoring increases. If reports of affective polarization are partly guided by self-presentation desires, then they should vary based on whether or not partisans have an audience as well as whether or not partisans will alter themselves to impress others. In a survey context, I can vary this perception by reminding participants of either their privacy or their lack of privacy (as well as measuring their self-monitoring tendencies and modeling it as a moderating variable).

To do this, I conduct a second survey experiment with a nationally representative (but partisan) sample from NORC's AmeriSpeak panel (N=1,771), where I randomly assigned participants to one of three conditions that vary the perception of privacy: 1) the control condition (participants are given no reminders); 2) the public condition (participants are given a reminder that results from the research may be published); or 3) the private condition (participants are given a reminder that their responses are completely private). These treatments were motivated by past research on self-presentation desires and survey responses that vary the perception of an "audience" in a survey. For example, Connors et al. (2019) find that reminding participants that their de-identified data will be publicly posted can alter survey responses because it reminds them that "someone is watching"—and that this public reminder effect is moderated by self-monitoring.

Again, I argue that people's reports of affective polarization are partly driven by self-presentation desires. Thus, the aim of the experimental treatments is to randomly vary respondents'

perception of an audience or lack of an audience—whether their responses are being shared with the general public and whether they are kept confidential—and thus the affective polarization responses partisans give.[7] If people's responses to the affective polarization questions are shaped by self-presentation desires, then they should vary by these conditions.

Importantly, though, not everyone should be equally influenced by the experimental treatments—again, the treatment effects should depend on participants' self-monitoring levels (i.e., their susceptibility to social pressure). If it is the case that reports of affective polarization are driven by self-presentation desires, the treatment effects should be moderated by self-monitoring. We should see, for example, that a high self-monitor reports the greatest affective polarization in the public condition—when they are encouraged to misrepresent themselves because they are presenting their positive "face" (Goffman 1967) to others—and the least in the private condition—when they have less motivation (or opportunity) to impress others.[8] This should be the case because social pressures are strongest in public settings and weakest in private settings and because susceptibility to social pressure increases with higher levels of self-monitoring. In sum, we should see a positive interaction of the public treatment and self-monitoring and a negative interaction of the private treatment and self-monitoring on reports of affective polarization.

Note, though, that we are dealing with strong pre-treatment here, where partisans have already been "treated" with partisan social pressure in the real world (see Levendusky and Malhotra 2015, 2016 for evidence of this pretreatment; see Druckman and Leeper 2012 for discussion of the effects of pre-treatment). Indeed, as we saw in study 1, partisans perceive strong social pressure to report affective polarization. This pre-treatment was one of the motivations behind the design for this survey experiment, as manipulating the *direction* of social pressure would be difficult—potentially impossible—given this pre-treatment context.

Yet, while it is all but impossible to successfully manipulate the direction of social pressure in this context within a quick survey experiment, it *is* possible to manipulate the extent to which people *feel* that social pressure. In other words, while we cannot manipulate the direction of the social pressure, we can manipulate the *strength* of that social pressure—or the level of privacy or lack thereof in participants' responses, with the assumption that social pressure should be strongest in public settings and weakest in private settings.

Although this design allows me to test my argument within the context of pre-treatment, pre-treatment will still influence my results. What this could mean is that—even with this design—we simply cannot shift affective polarization in a quick survey experiment, leading to null findings. It also could mean that if we *do* see an effect, we will see it for treatments trying to *mitigate* affective polarization reports rather than those trying to *exaggerate* them. Lastly, it suggests that if we *do* see treatment effects, these are likely conservative estimates of the effect of partisan social pressure on self-reports of affective polarization, as pre-treatment will bias effects towards null results (Druckman and Leeper 2012).

Design. Thus, after answering questions about demographics, political affiliations, and self-monitoring tendencies, participants were assigned to one of the three conditions.[9] The only difference between the control condition and the two treatment conditions is one sentence in each

[7] It also does this without compromising ethics, as all treatments' statements are true.

[8] Again, this trait is continuous, where those at the lowest end are the *least* likely to misrepresent themselves to impress others and those at the highest end are the *most* likely to do so.

[9] There was a fourth condition (a "friends" condition), the findings of which are beyond the scope of this manuscript (but are mentioned here for transparency).

treatment condition prior to the polarization questions. In the public condition they are told: "Just a reminder, the results based on your responses may be published." In the private condition they are told: "Just a reminder, your responses are completely private."

Then, respondents were asked four questions to measure affective polarization—the same two feeling thermometer questions that were used to measure affective polarization in study 1 and two questions about trust in Democrats and Republicans to do what is "right" for the country (also to measure affective polarization—see Druckman and Levendusky 2019; question wording in Appendix B).[10] Like the feeling thermometers measure, this trust measure is then coded as the difference between trust toward in-partisans and trust toward out-partisans, where greater numbers indicate greater affective polarization. Both measures are continuous.

Main Effect Results. First I examine main treatment effects (see Table 1). Comparing the public condition to the control condition illustrates the opposite of what I predicted—that affective polarization reports are lower (rather than higher, as I had predicted) in the public condition than in the control, although this is not significant (feeling thermometers: p=.167; trust: p=.632). Comparing the private condition to the control condition, we see affective polarization is lower, as predicted, but this is also insignificant (feeling thermometers: p=.302; trust: p=.706).

Table 1. Affective Polarization by Treatment

	Feeling Thermometers			***Trust***		
Condition	**N**	**Mean**	**SE**	**N**	**Mean**	**SE**
Public	605	49.68	1.42	643	1.81	0.05
Control	566	52.46	1.41	604	1.84	0.05
Private	600	50.35	1.47	637	1.82	0.05

Interaction Results. Next I examine the potential interaction of the treatments with self-monitoring. This trait is measured using Berinsky's (2004) 3-item questionnaire, where people are asked how good or bad an actor they would be, how often they put on a show to entertain others, and how often they are the center of attention—they are given four response options to each of these three questions (see Appendix B for question wording). Reverse coding and combining individuals' responses to these three questions creates a 12-point scale from 0 (the lowest self-monitor) to 12 (the highest self-monitor).[11] This measure has been validated and used in political science research (e.g., Berinsky and Lavine 2012; Connors 2019; Connors, Krupnikov, and Ryan 2019).

First, as compared to the control, I find no significant interaction of the public treatment and self-monitoring on either of the two affective polarization measures (coefficient of public treatment and self-monitoring on affective polarization measured by feeling thermometers: α=-1.24, p=.169 and by trust: α=-0.054, p=.103), indicating—again—that the public treatment was ineffective at changing respondents' reported polarization. This makes sense when we think of the control group as pre-treated—people have already reported as much polarization as they are willing to report to fit in. A public reminder will not influence participants' affective polarization

[10] There were also other post-treatment measures, the findings of which are also beyond the scope of this manuscript.

[11] As is typical in the population, the sample was skewed towards low self-monitors with a mean of 4.07 and standard deviation of 2.22 and the Cronbach's alpha for the 3-item scale was .65.

reports because they are *already* driven by social pressure in the control condition.[12]

Next I examine the effect of the private treatment with self-monitoring. Again, given pre-treatment, we should see the real shift in reports of affective polarization when trying to mitigate reports of affective polarization (i.e., when participants' perception of privacy *increases*). Importantly, because the public condition did not work as expected, the following analyses compare the private condition to the control condition, rather than to the public condition, as I had discussed in my expectations. Doing this, we indeed see the predicted negative interaction of the private treatment and self-monitoring on affective polarization with both measures (coefficient of private treatment and self-monitoring on affective polarization measured by feeling thermometers: α=-2.18, p=.020 and by trust: α=-0.10, p=.005; see Figure 2).

In fact, while the marginal treatment effect of the private treatment for the lowest self-monitor on the feeling thermometers measure is 6.83 (p=.114), for the highest self-monitor it is -19.28 (p=.013)—a notable difference of 26.11 points. For the trust measure, this is 0.37 (p=.013) for the lowest self-monitor and -0.78 (p=.005) for the highest self-monitor. In other words, when being reminded that one's responses are completely private, people depress their levels of affective polarization—and this is moderated by the self-monitoring trait, where this negative effect gets stronger as people become higher self-monitors (i.e., as they become more likely to misrepresent themselves in public settings). This interactive treatment effect is notable given both pre-treatment and the fact that the treatment is a simple one-line reminder.

Figure 2. Effect of Private Treatment on Affective Polarization by Self-Monitoring

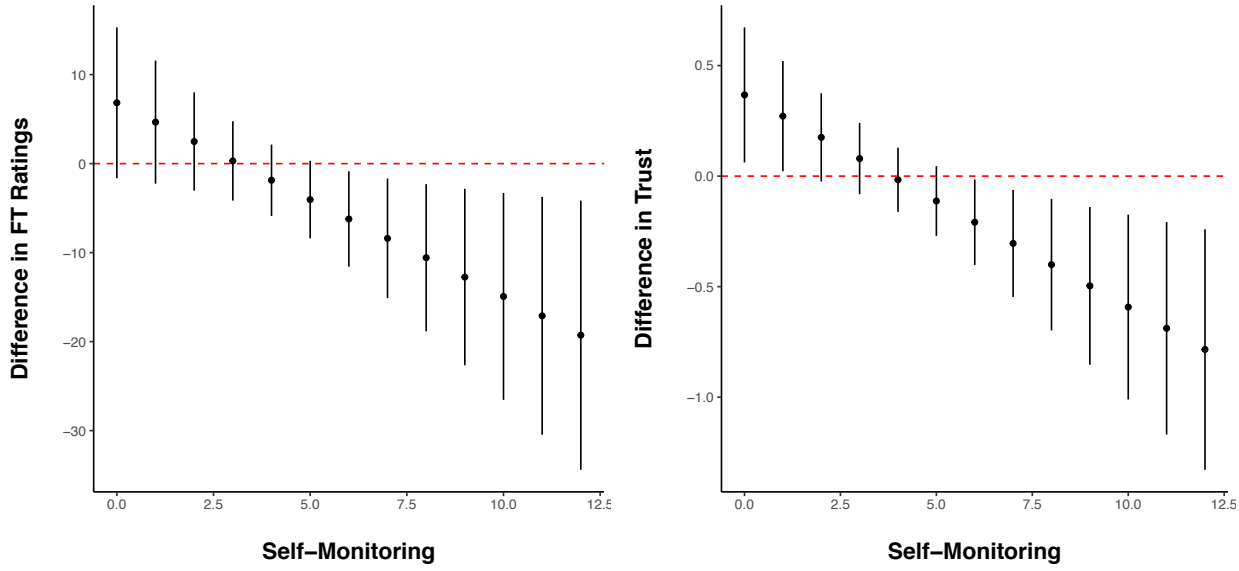


Figure 2 shows the marginal effects of the private treatment on affective polarization moderated by self-monitoring based on OLS regression models with 95% confidence intervals (see Appendix B, Table B.1). On the left, affective polarization is measured as the difference between in-party and out-party feeling thermometers. On the right, affective polarization is measured as the difference between in-party and out-party trust.

[12] 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 affective polarization (similar to the effect of the private treatment)—although this explanation is post-hoc.

Importantly, we see no interaction of either of the treatments and partisan strength (coefficient of public treatment and partisan strength dummy on affective polarization as measured by feeling thermometers: α=-3.22, p=.411 and by trust α=0.00, p=.987; coefficient of private treatment and partisan strength dummy on affective polarization as measured by feeling thermometers: α=-5.81, p=.136 and by trust α=-0.10, p=.481). Although this deviates from what we would expect with the alternative partisan social identity story, it makes sense if we consider that strong partisans potentially have more crystallized, perhaps "truer"—and therefore more immovable—feelings about in-partisans and out-partisans.

Thus, it is possible that weak partisans are more malleable in their affective polarization attitudes. This is indeed what we see: when we limit the sample to weak partisans, we see an even stronger treatment effect of the private treatment (coefficient of private treatment and self-monitoring on affective polarization measured by feeling thermometers: α=-2.84, p=.023 and by trust: α=-0.11, p=.011). In fact, for weak partisans, the marginal treatment effect of the private treatment for the lowest self-monitor on the feeling thermometers measure is 12.73 (p=.028), for the highest self-monitor it is -21.37 (p=.037)—a notable difference of 34.1 points. For the trust measure this respective difference is 0.49 (p=.016) and -0.87 (p=.017).

Next, I run various checks. First, given that self-monitoring is measured rather than randomly assigned, I rerun the main models from the experiment adding controls (party dummy, partisan strength, age, gender, race, ethnicity, and education; see Kam and Trussler 2017 for use of controls in experiments) and find almost the exact same results (see Appendix B, Table B.2). Second, I run a replication of this experiment on a sample from Amazon's Mechanical Turk and find the same results (see Appendix C).[13]

Third, I check if the results are driven by in-party or out-party feelings (see Boxell, Gentzkow, and Shapiro 2017 who operationalize affective polarization with just out-party antipathy). The interactive effect of the private treatment and self-monitoring is quite similar for increasing positive feelings toward out-partisans (feeling thermometers: α=1.08, p=.065; trust: α=0.03, p=.196) and decreasing positive feelings toward in-partisans (feeling thermometers: α=-1.26, p=.019; trust: α=-0.07, p=.002), although more robust for the latter. Lastly, I check if the results are driven by Democrats or Republicans. For Democrats, the marginal effect of the private treatment and self-monitoring is -3.74 (p=.003) for feeling thermometers and -0.15 (p=.001) for trust, whereas for Republicans it is -0.49 (p=.718) for feeling thermometers and -0.04 (p=.409) for trust, demonstrating that the treatment works in the same direction for Democrats and Republicans, but is only significant for Democrats. Although more research is needed to understand why this would be the case, I will discuss this finding in the final discussion.

Given what we know about self-monitoring, these findings suggest that in ordinary circumstances (e.g., on public opinion surveys or when talking with others) high self-monitors are exaggerating their levels of affective polarization. When told they have privacy and thus are less motivated (or have less opportunity) to impress others, their reported levels of affective polarization are more tempered. In conjunction with the null findings of the public treatment, these results imply that in ordinary circumstances we are witnessing overreports of affective polarization (i.e., that people are already treating surveys as "public," and thus those who want to impress others are overreporting affective polarization in surveys and other similarly public settings). Simply reminding them that their responses are private mitigates these reports.

---

[13] This survey experiment was done *prior* to the main survey experiment and was used as pre-test data for the Time-Sharing Experiments for the Social Sciences (TESS) proposal.

American National Election Studies (ANES) Analysis

Lastly, I examine my third expectation: that we see the relationships we would expect to see played out in the real world—i.e., that these results are externally valid. I expect that those who are especially attuned to social contexts (i.e., high self-monitors) will be most influenced by partisan social pressure—and thus reporting of affective polarization should increase with higher levels of self-monitoring (controlling for other important variables). To examine this, I make use of the American National Election Studies (ANES). As my independent variable of interest—self-monitoring—was only measured in the 2008 ANES, I limit my analyses to this year.

Although affective polarization has increased *since* this time, 2008 still demonstrated extremely high levels of affective polarization. On a scale from least affectively polarized (0) to most affectively polarized (200) as measured by feeling thermometers, the mean of partisans in 2008 was 145.50 with a standard deviation of 31.78. Similarly, 38.69% of partisan respondents in 2008 were the most affectively polarized as measured by likes and dislikes. Thus, 2008 is well-suited to test my first expectation, as indeed there was a climate of affective polarization.

In this set of analyses I am able to use two different measures of affective polarization. The first measure is that which I used in studies 1 and 2 (the "feeling thermometers" measure). The second measure uses respondents' reporting of likes and dislikes of in-partisans and out-partisans (the "likes and dislikes" measure; see Iyengar et al. 2019; Levendusky and Malhotra 2015). Using these questions, I create a binary variable that is coded as 1 when respondents report likes of in-partisans but no likes of out-partisans *as well as* dislikes of out-partisans but no dislikes of in-partisans (i.e., when they report liking in-partisans and disliking out-partisans). It is coded 0 otherwise (i.e., when they report liking *both* in-partisans *and* out-partisans or *disliking* both). Again, using multiple measures of affective polarization serves as an additional check on the results.

Model Estimation. Given the observational nature of this analysis, I control for gender, age, race, education, partisanship, ideology, partisan strength, political interest, income, and media consumption, as these variables should drive polarization (Boxell 2020; Mason 2015).[14] For the feeling thermometer dependent variable, which is continuous, I estimate OLS and include robust standard errors (N=441).[15] For the likes and dislikes dependent variable, which is binary, I estimate logistic regression (N=634). Again, if it is the case that partisan social pressure exaggerates reports of affective polarization, then we should see that one's susceptibility to conform to social pressures is related to the affective polarization they report. In particular, we should see that as self-monitoring increases among partisans—meaning that as partisans' desire to impress others increases—people report higher affective polarization.

---

[14] In the likes and dislikes analysis, because two controls (media and interest) are missing data (cutting the sample in half), these missing values are held at their means for the model estimation. Note that Mason (2015) also includes southern residence (dummy), urban residence (dummy), church attendance, and evangelicalism (dummy). Although adding these controls limits the sample (N=224 for feeling thermometers and N=330 for likes and dislikes), self-monitoring remains marginally significant for feeling thermometers (α=1.260, p=0.091) and significant for likes and dislikes (α=.121, p=0.028), although the likes and dislikes model eliminates the evangelical control.

[15] As is demonstrated later on, the same general results are found without robust standard errors.

Results. As predicted, self-monitoring is positively correlated with a greater likelihood of reporting affective polarization measured by both feeling thermometers (self-monitoring coefficient in model predicting feeling thermometer affective polarization with controls: α=1.160, p=.034) and partisan likes and dislikes (self-monitoring coefficient in model predicting likes and dislikes affective polarization with controls: α=0.077, p=.037). These models are depicted in Figure 3 and the full model estimations can be found in Appendix D, Tables D.1 and D.2.

Figure 3. Predicted Affective Polarization by Self-Monitoring

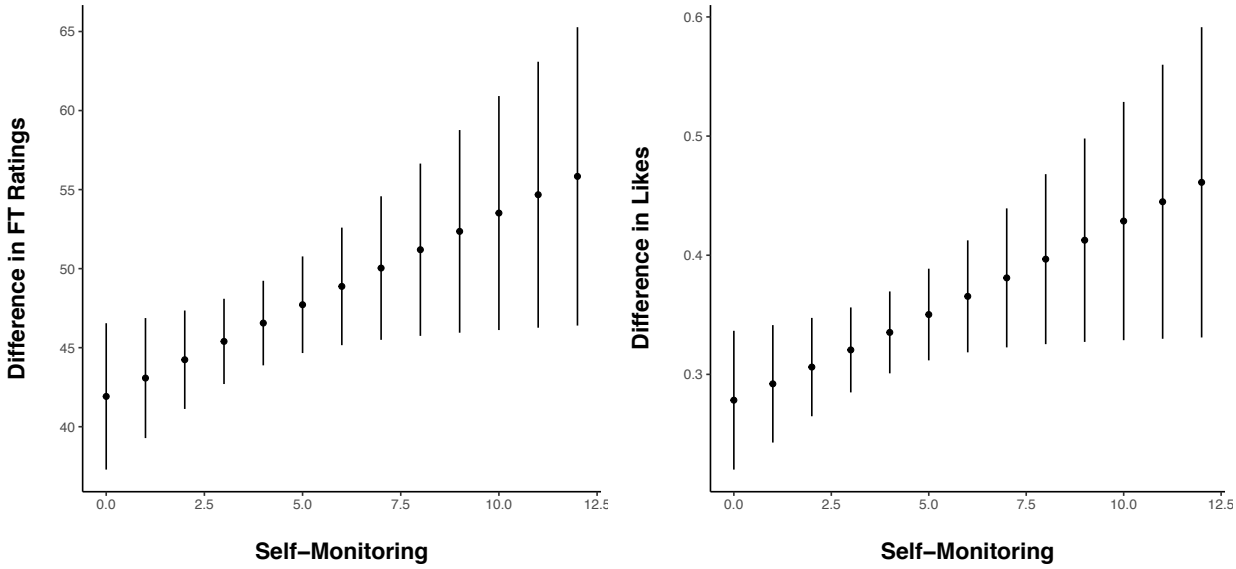


Figure 3 shows the effect of self-monitoring on reports of affective polarization. On the left, affective polarization is measured as the difference between feeling thermometers towards in-partisans versus out-partisans (continuous). On the right, affective polarization is measured as reporting only likes of in-partisans and dislikes of out-partisans versus not (binary). Controls included: party dummy, gender, age, income, race, education, ideology, partisan strength, political interest, and media consumption. The model on the left (feeling thermometers) is based on OLS regression with robust standard errors and shows the predicted difference of feeling thermometer ratings. The model on the right (likes and dislikes) is based on logistic regression and shows the predicted probability of affective polarization. Confidence intervals are 95%.

Various robustness checks were conducted on these results. First, I conduct the same analyses sequentially removing controls. For the first measure of affective polarization (feeling thermometers), altering the model specifications does not change the positive direction. Further, with each of these estimations the effect of self-monitoring on affective polarization remains between 1 and 1.5 and the significance level in all but 3 models remains below p=.05 (removing strength: p=.061, removing ideology: p=.062, and removing age: p=.058). For the second measure of affective polarization (likes and dislikes), altering the model specifications also does not greatly change the positive direction, size, or significance: with each of these estimations the effect of self-monitoring on affective polarization remains between .015 and .017 and in all but 3 models the significance level remains below p=.05 (removing interest: p=.057, removing education: p=.054, and removing age p=.051).

Second, for the feeling thermometers measure, I conduct the same analysis removing robust standard errors and find no substantial change in the results. I do not do so for the likes and dislikes measure, as the dependent variable here is binary and thus I did not originally estimate

robust standard errors. Next, I split the in-party and out-party feeling thermometers to see if one is driving the results. Although the results are more robust with in-party favorability than with out-party antipathy, we see the same general trend with both (see Appendix D, Table D.1).

Then, I see if the results are driven by Democrats or Republicans. The self-monitoring coefficient is in the same *direction* for both Democrats and Republicans, although it is a bit larger and for Democrats (coefficient in feeling thermometers model for Democrats: α=1.78, p=.003 and Republicans: α=0.53, p=.631; coefficient in likes and dislikes model for Democrats: α=.078, p=.066 and Republicans: α=.065, p=.413). Note that the change in significance for Republicans could be a result of the small sample of Republicans (N=176 in feeling thermometers and N=246 in likes and dislikes). I will return to this differential effect by partisanship in the general discussion.

Lastly, to be sure that self-monitoring is not simply predictive of reporting likes or dislikes (i.e., that high self-monitors are simply more likely to cooperate with the survey and give a response to this question), I examine if self-monitoring is correlated with likes of Democrats, dislikes of Democrats, likes of Republicans, and dislikes of Republicans. This is not the case—self-monitoring is far from significantly related to each of these four outcomes. That is, self-monitoring does not predict whether one lists likes or dislikes of partisans, but it *does* predict affective polarization using these variables. Only when they are coded as likes of in-partisans but no likes of out-partisans and no dislikes of in-partisans but dislikes of out-partisans (i.e., when this affective polarization measure is created) does self-monitoring matter (see Appendix D, Table D.2).

Overall, these findings indeed demonstrate what we would expect to see from my argument: that when there is partisan social pressure to report affective polarization, those who are most willing to alter their stated preferences to impress others are *also* those most likely to report affective polarization, while controlling for partisanship, gender, age, income, race, education, ideology, partisan strength, political interest, and media consumption. We see from the analyses that among those who are willing to associate with a party, the more one acquiesces to social pressure, the more they report antipathy toward out-partisans and favorability toward in-partisans. This is exactly what we would expect to see given my argument.

Discussion & Conclusion

The results from three studies—using both experimental data and observational, nationally representative data—suggest that there is partisan social pressure to report affective polarization and that, because of self-presentation desires, this is likely inflating self-reports of affective polarization. It is important to note that if partisans are exaggerating their reports of affective polarization in surveys, it is possible that these exaggerations are even stronger in social settings with in-partisans—although this extension of my findings is left for future research.

Importantly, I also found unexpected nuance in my results. First, I discovered differential treatment effects and observed effects among Democrats and Republicans. All three studies found greater effects among Democrats—Democrats reported slightly higher partisan social pressure, reacted more strongly to the private treatment, and were more influenced in observational data by self-monitoring. It is possible that this reflects stronger pre-treatment among Democrats who apparently observe greater partisan social pressure regarding affective polarization and then reflect this in the real world (studies 1 and 3). If this were true, the stronger effects among Democrats in study 2 would be a reflection of the fact that they had more room to move—they were already pre-

treated in the real world, and thus a private treatment could more effectively depress their affective polarization reports down to their perhaps "truer" levels. Of course, this theoretical analysis is post-hoc and more research is needed to examine this.

Second, I discovered that the treatment in study 2 was slightly more successful at altering in-party feelings and that self-monitoring in study 3 was similarly slightly more predictive of in-party feelings than of out-party feelings. This might suggest some support for a partisan identity story, although other results do not support this same perspective. Remember, *none* of the treatments were moderated by partisan strength in the direction that a social identity perspective would have predicted. A strong partisan was not any more influenced by the treatments than a weak partisan was—in fact, we saw the opposite. In adjusting self-reports of affective polarization, then, it didn't matter how *strong* a partisan one was (in fact, weak partisans were more influenced by the treatments), but it *did* matter how much that partisan changes themselves to impress others (i.e., how high of a self-monitor they were). This not only reinforces my argument that partisans are reporting higher affective polarization because of the partisan social pressure to do so, but it also suggests that perhaps the "true" highly affective polarized partisans are the strong partisans—as they were more resistant to change in the experiments. The *weak* partisans may be the ones who are simply following along in response to partisan social pressure.

Nuances aside—and left for future research—the findings of this piece are important beyond just worries about the measurement of polarization, although this *is* quite important.[16] Given the importance of perception here—that the pressure to report polarization is ultimately driven by the perceived composition of the partisan population *as* polarized—this research suggests a dangerous snowball effect: the more people pretend to be polarized, the more the public *thinks* everyone is polarized, and the more people will pretend to be polarized. Further, these overreports of polarization may then drive opposing partisans to genuinely increase their out-party hostility, reciprocating what they think is true animus. Thus, it does not necessarily matter if these reports are cheap talk: they have many of the same negative consequences.[17]

Further, this research offers a potential way to alleviate reports of polarization—by increasing the perception of response privacy. My research also suggests that—if possible—changing the perception of polarization among the public could also alleviate reports of affective polarization. Respondents *think* the public is deeply affective polarized, and hence they too report higher levels of polarization. But these perceptions are based on misleading media reports. If individuals knew the true level of affective polarization in the public, then they may report lower levels of affective polarization (for a related argument, see Ahler and Sood 2018). Just like perceptions of high polarization could increase reports of polarization, perceptions of *low* polarization could *decrease* reports of polarization.

While these findings suggest that reports of affective polarization are partly driven by social pressure, they certainly do not suggest that affective polarization is merely a social construct. It is important to consider these findings with a broader perspective in order to understand partisanship in America. To do so, it is helpful to imagine the public in a more nuanced way, with different segments of Americans each having their own reactions to the political world. Rather than imagining one trend among the public (e.g., "Democrats hate Republicans" or "everyone is

[16] Note, again, that this research uses three often-used measures of affective polarization (feeling thermometers, likes and dislikes, and trust).

[17] Although I would argue that even calling these reports "cheap talk" artificially diminishes their importance given how common it is that people deal with social influence pressures in both social and political settings.

faking polarization" or "people hate politics"), let us imagine three overarching segments of Americans.

The first group is outlined by Mason (2015, 2018) and Iyengar and Westwood (2015), among others—they are the genuinely polarized and sorted partisans. This group is highly important and concerning (see Iyengar et al. 2019). The second group is responding to this first group—they are turned off by these extreme partisans and thus are withdrawing from partisanship (Klar and Krupnikov 2016) and even politics (Klar et al. 2018). The third group—the segment of the public of which I'm theorizing—is *also* responding to this first group, but in a different manner. Rather than withdrawing from partisanship or politics, this segment is becoming *more partisan*—or at least reporting to be so.

Thus, my findings do *not* show that these first two groups do not exist. Instead, they add a broader perspective by bringing attention to this third group—those who believe partisans are polarized and respond by acquiescing. My findings suggest that social identity and deep-seated animosity are not the full story in explaining reports of affective polarization, but that these measures are partly expressive, perhaps a sort of partisan cheerleading (Bullock et al. 2015; Prior et al. 2015).[18]

In terms of lingering theoretical questions, two big stones are left unturned. First, as previously mentioned, we need a broader picture of how self-presentation desires influence partisanship and partisan expression. What differentiates the high self-monitor that becomes an independent from the one who becomes an affectively-polarized partisan? Second, we are left wondering about self-reports of affective polarization. In particular, what do these self-reports *mean* given that they are relatively costless? Are they cheap talk? Do they lead to political behavior that mimics these reports? A next step would be to examine this, testing the effects of both the self-monitoring trait and social pressure on political polarization *behavior*. Getting to this next step would help us to situate these findings better—we could see if people are "simply" lying about how much they are polarized or if they are instead following the social pressure to the extreme and adopting the same partisan biases that the genuinely polarized express.

Examining polarization behavior could also help us to better understand the normative consequences of these findings. If it is indeed the case that these polarization reports are cheap talk, then we have evidence of arbitrary overreporting of polarization, which could be optimistic (i.e., we are not as polarized as we may seem) or pessimistic (i.e., people are lying about hating each other on surveys), although—again—either way cheap talk is consequential. If we find, however, that these reports actually lead to polarized political behavior, we would have further evidence of a "snowball effect" of political climates—the idea that a climate of political polarization will, on its own, produce more polarization, as well as further evidence of just how strong social pressure can be in influencing politics. At this point, though, the findings from this piece should give us pause and motivation to better understand the measurement of political polarization, the context surrounding that polarization, and more broadly how social influence affects the expression of partisanship.

[18] Note, though, that partisan cheerleading is conceived of as "knowingly distorting" (see Peterson and Iyengar, forthcoming) survey responses to defend one's party, which differs from my conception that partisans are (perhaps unknowingly) distorting survey responses to seem like a good partisan.

References

Abramowitz, Alan I., and Steven Webster. 2016. "The Rise of Negative Partisanship and the Nationalization of U.S. Elections in the 21st Century." *Electoral Studies* 41: 12–22.

Ahler, Douglas J. 2014. "Self-Fulfilling Misperceptions of Public Polarization." *The Journal of Politics* 76(3): 607-620.

Ahler, Douglas J. and Gaurav Sood. 2018. "The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences." *The Journal of Politics* 80(3): 964-981.

Barbera, Pablo. 2014. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1): 76-91.

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.

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

Boxell, Levi. 2020. "Demographic Change and Political Polarization in the United States." *Economic Letters*.

Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro. 2017. "Greater Internet Use is Not Associated with Faster Growth in Political Polarization Among US Demographic Groups." *Proceedings of the National Academy of Sciences* 114(40): 10612-10617.

Bullock, John, Alan Gerber, Seth Hill and Gregory Huber. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10: 519-578.

Carlson, Taylor N., and Jaime E. Settle. 2016. "Political Chameleons: An Exploration of Conformity in Political Discussions." *Political Behavior* 38(4): 817-859.

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(6): 1015.

Connors, Elizabeth C. 2019. "The Social Dimension of Political Values." *Political Behavior*, doi: 10.1007/s11109-019-09530-3.

Connors, Elizabeth C., Yanna Krupnikov, John Barry Ryan. 2019. "How Transparency Affects Survey Responses." *Public Opinion Quarterly* 83(S1): 185-209.

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. 2020. "How to Survey Citizens' Compliance with COVID-19 Public Health Measures: Evidence from Three Survey Experiments." *Journal of Experimental Political Science*, first view.

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(4): 875–896.

Druckman, James N. and Matthew S. Levendusky. 2019. "What Do We Measure When We Measure Affective Polarization?" *Public Opinion Quarterly*, 83(1): 114-122.

Gangestad, Steven W., and Mark Snyder. 2000. "Self-Monitoring: Appraisal and Reappraisal." *Psychological Bulletin* 126(4): 530-555.

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.
Holtgraves, Thomas. 1992. "The Linguistic Realization of Face Management: Implications for Language Production and Comprehension, Person Perception, and Cross-Cultural Communication." *Social Psychology Quarterly*: 141-159.
Huckfeldt, Robert, Jeffrey J. Mondak, Matthew Hayes, Matthew T. Pietryka, and Jack Reilly. 2013. "Networks, Interdependence, and Social Influence in Politics." In *The Oxford Handbook of Political Psychology*, edited by Leonie Huddy, David O. Sears, and Jack S. Levy. Oxford University Press.
Huddy, Leonie, Liliana Mason, and Lene Aarøe. 2015. "Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity." *American Political Science Review* 109(1): 1–17.
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(4): 1326-1338.
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; 129-146.
Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76(3): 405–431.
Iyengar, Shanto, and Sean J. Westwood. 2015. "Fear and Loathing across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* 59(3): 690-707.
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(4): 789-815.
Klar, Samara. 2014. "Partisanship in a Social Setting." *American Journal of Political Science* 58(3): 687-704.
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(2): 379-390.
Kuran, Timur. 1997. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press.
Levendusky, Matthew. 2009. *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans*. University of Chicago Press.
Levendusky, Matthew and Neil Malhotra. 2015. "(Mis)perceptions of Partisan Polarization in the American Public." *Public Opinion Quarterly* 80(S1): 378-391.
Levendusky, Matthew and Neil Malhotra. 2016. "Does Media Coverage of Partisan Polarization Affect Political Attitudes?" *Political Communication* 33(2): 283-301.
Mason, Lilliana. 2015. "'I Disrespectfully Agree': The Differential Effects of Partisan Sorting on Social and Issue Polarization." *American Journal of Political Science* 59(1): 128-145.
Mason, Lilliana. 2018. *Uncivil Agreement: How Politics Became Our Identity*. University of Chicago Press.
McGregor, Shannon C. 2019. "Social Media as Public Opinion: How Journalists Use Social Media to Represent Public Opinion." *Journalism* 20(8): 1070-1086.

Mutz, Diana C. 1998. *Impersonal Influence: How Perceptions of Mass Collectives Affect Political Attitudes*. Cambridge University Press.
Mutz, Diana C. 2002. "Cross-Cutting Social Networks: Testing Democratic Theory in Practice." *American Political Science Review* 96(1): 111–126.
Orr, Lilla V., and Gregory A. Huber. 2019. "The Policy Basis of Measured Partisan Animosity in the United States." *American Journal of Political Science*.
Peterson, Erik, and Shanto Iyengar. 2019. "Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading?" *American Journal of Political Science*, forthcoming.
Prior, Markus, Gaurav Sood, and Kabir Khanna. 2015. "You Cannot be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions." *Quarterly Journal of Political Science* 10(4): 489-518.
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(3): 389-414.
Robison, Joshua, and Kevin J. Mullinix. 2016. "Elite Polarization and Public Opinion: How Polarization is Communicated and its Effects." *Political Communication* 33(2): 261-282.
Settle, Jaime E. 2018. *Frenemies: How Social Media Polarizes America*. Cambridge University Press.
Shadish, William R., Thomas D. Cook, and Donald Thomas Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
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(4): 526.
West, Emily A., and Shanto Iyengar. 2020. "Partisanship as a Social Identity: Implications for Polarization." *Political Behavior*, first view.
Westwood, Sean, Erik Peterson, and Yphtach Lelkes. 2018. "Are There Still Limits on Partisan Prejudice?" *Public Opinion Quarterly* 83(3): 584-597.
Zaller, John, and Stanley Feldman. 1992. "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences." *American Journal of Political Science* 36(3): 579- 616.
Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Appendix A: Prolific Survey Experiment

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 <u>Democrats</u> (again, as a [Democrat / Republican] wanting to impress other [Democrats / Republicans])? [0 to 100 degrees]
        ii. How would you rate <u>Republicans</u> (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: [_____]

Appendix B: AmeriSpeak Survey Experiment

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]:

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 B.1. Effect of Public and Private Treatments and Self-Monitoring Interactions on Affective Polarization (Measured by Feeling Thermometers and Trust)

	Thermometers	**Trust**
Public Treatment	2.229	0.181
	(4.156)	(0.152)
Private Treatment	6.835	0.367
	(4.298)	(0.157)
Self-Monitoring	0.474	0.018
	(0.673)	(0.024)
Public*Self-Monitoring	-1.240	-0.054
	(0.901)	(0.033)
Private*Self-Monitoring	-2.176	-0.096
	(0.932)	(0.034)
Constant	50.670	1.771
	(3.054)	(0.111)

Regression table based on two OLS models (the first predicting affective polarization as measured by feeling thermometers and the second predicting affective polarization as measured by trust). All treatment effects in comparison to control condition. Standard errors in parentheses.

Table B.2. Effect of Private Treatment and Self-Monitoring Interactions on Affective Polarization with Controls

	Thermometers	**Trust**
Private	6.699	0.381
	(4.081)	(0.145)
Self-Monitoring	0.891	0.038
	(0.649)	(0.023)
Private*SM	-2.085	-0.096
	(0.885)	(.031)
Democrat	-1.605	0.059
	(2.168)	(0.077)
Strength	22.912	0.918
	(1.954)	(0.070)
Age	0.157	0.008
	(0.065)	(0.002)
Gender	-1.204	0.098
	(2.001)	(0.072)
Education	-0.096	-0.026
	(0.637)	(0.023)
Constant	36.096	1.771
	(9.394)	(0.111)

Standard errors in parentheses. Race and ethnicity included as controls but not in table to save space.

Appendix C: Mturk Survey Experiment

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]

Figure C.1. Effect of Public Treatment on Affective Polarization by Self-Monitoring

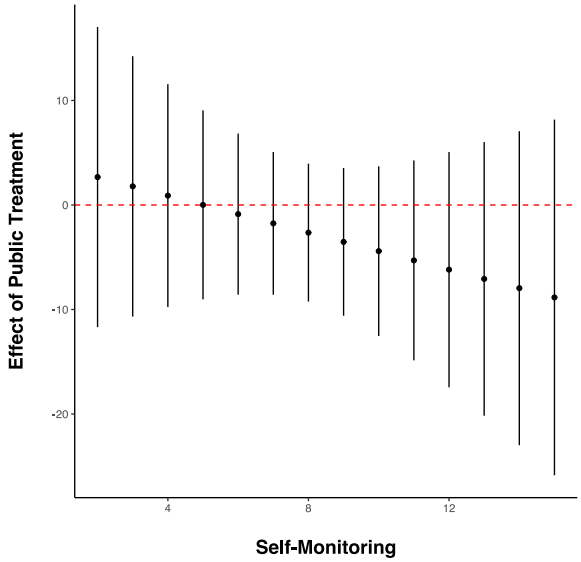


*Figure shows no significant interaction of the public treatment by self-monitoring on affective polarization (*p=*.427). Only weak and strong partisans included. Randomly assigned independent variable is public treatment (as compared to control). Self-monitoring is continuous from low to high—see Berinsky 2004. Dependent variable—affective polarization—is operationalized with the difference between feeling thermometers of in-party versus out-party members—continuous, from low (0) to high (170) affective polarization. Marginal plot is based on OLS regression.*

Figure C.2. Effect of Private Treatment on Affective Polarization by Self-Monitoring

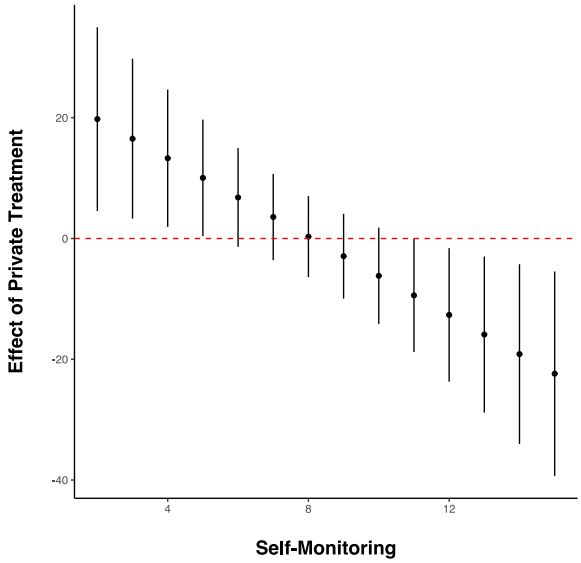


*Figure shows a significant negative interaction of the public treatment by self-monitoring on affective polarization (*p=*.005). Only weak and strong partisans included. Randomly assigned independent variable is private treatment (as compared to control). Self-Monitoring is continuous from low to high—see Berinsky 2004. Dependent variable—affective polarization—is operationalized with the difference between feeling thermometers of in-party versus out-party members—continuous, from low (0) to high (170) affective polarization. Marginal plot is based on OLS regression.*

Appendix D: ANES Analysis

Table D.1. Affective Polarization (Feeling Thermometers), In-Party & Out-Party Thermometers by Self-Monitoring

	Main Model	In-Party Continuous	In-Party Binary	Out-Party Continuous	Out-Party Binary
Self-Monitoring	1.16	0.793	0.196	-0.371	0.114
	(0.546)	(0.298)	(0.0512)	(0.432)	(0.0499)
Media	1.834	-1.141	0.439	-2.976	-0.443
	(7.000)	(4.494)	(0.706)	(5.404)	(0.713)
Interest	0.261	-2.906	0.102	-3.260	0.109
	(7.616)	(4.238)	(0.653)	(5.295)	(0.670)
Strength	20.39	10.56	1.389	-9.902	1.002
	(2.915)	(1.813)	(0.320)	(2.167)	(0.316)
Liberal	-5.351	-6.693	-0.960	-1.282	-0.131
	(4.900)	(2.704)	(0.390)	(3.204)	(0.390)
Extremity	18.19	2.720	0.0790	-15.36	1.382
	(8.455)	(5.216)	(0.773)	(5.832)	(0.743)
Education	16.34	-0.490	-0.136	-16.92	1.251
	(6.893)	(4.213)	(0.624)	(4.752)	(0.607)
Male	2.780	-2.409	-0.202	-5.149	0.344
	(2.731)	(1.605)	(0.267)	(1.964)	(0.262)
Age	8.276	3.913	1.584	-4.318	0.543
	(8.247)	(4.473)	(0.739)	(5.986)	(0.732)
Income	-16.63	-11.49	-2.712	5.237	-1.403
	(6.681)	(3.911)	(0.662)	(5.003)	(0.642)
White	0.609	-0.840	-0.342	-1.683	0.200
	(4.165)	(2.530)	(0.450)	(3.526)	(0.469)
Black	-1.701	1.150	0.537	2.629	0.170
	(5.358)	(3.098)	(0.475)	(4.224)	(0.511)
Democrat	17.90	10.02	1.024	-7.859	0.501
	(5.250)	(2.879)	(0.398)	(3.339)	(0.394)
Constant	105.4	72.84	-2.526	67.64	-4.103
	(9.073)	(5.405)	(0.940)	(7.084)	(0.943)
N	441	442	443	441	443

Table 1 shows the effect of self-monitoring (continuous from 0, low, to 12, high) on affective polarization measured as the difference in feeling thermometers towards the in-party minus feeling thermometers towards the out-party (main model), in-party thermometers (as both continuous and binary), and outparty thermometers (as both continuous and binary). The in-party continuous dependent variable is coded from 0 to 100, where higher numbers indicate more favorability toward the in-party and lower numbers indicate less favorability toward the in-party. The in-party binary dependent variable is coded as 1 if the respondent is the most favorable and 0 otherwise. The out-party continuous dependent variable is also coded from 0 to 100, but where higher numbers indicate more favorability toward the out-*party. The out-party binary dependent variable is coded as 1 if the respondent is the least favorable toward the out-party and 0 otherwise. In the three models with a continuous dependent variable, OLS is estimated and robust standard errors are calculated. In the two models with a binary dependent variable, logistic regression is estimated.*

The liberal coefficient is in comparison to conservatives—moderates were included in the regression but not the table.

Table D.2. Predicted Affective Polarization (Likes and Dislikes) by Self-Monitoring

	Main Model	Like Democrat	Dislike Democrat	Like Republican	Dislike Republican
Self-Monitoring	0.077	0.045	-0.029	-0.022	0.024
	(0.037)	(0.034)	(0.032)	(0.035)	(0.033)
Media	-0.075	0.042	0.477	0.484	0.147
	(0.672)	(0.689)	(0.599)	(0.641)	(0.593)
Interest	-1.627	1.075	1.811	0.631	1.282
	(0.668)	(0.635)	(0.574)	(0.593)	(0.546)
Strength	0.582	0.318	-0.033	0.120	0.329
	(0.208)	(0.196)	(0.173)	(0.190)	(0.169)
Liberal	-0.358	0.475	-0.040	-0.577	0.563
	(0.296)	(0.292)	(0.229)	(0.241)	(0.235)
Extremity	0.857	-0.149	-0.108	-0.543	-0.430
	(0.517)	(0.506)	(0.432)	(0.473)	(0.434)
Education	-1.256	1.619	1.696	1.397	1.757
	(0.447)	(0.456)	(0.388)	(0.420)	(0.402)
Male	-0.629	-0.011	0.568	0.368	0.396
	(0.189)	(0.184)	(0.162)	(0.176)	(0.162)
Age	0.376	-0.218	0.190	-0.017	-0.138
	(0.547)	(0.537)	(0.463)	(0.510)	(0.461)
Income	-0.670	0.730	0.738	0.381	0.613
	(0.444)	(0.445)	(0.382)	(0.416)	(0.377)
White	0.178	0.074	0.335	0.092	0.347
	(0.326)	(0.318)	(0.254)	(0.280)	(0.263)
Black	0.731	0.109	-0.519	-0.036	0.205
	(0.351)	(0.372)	(0.291)	(0.311)	(0.291)
Democrat	0.727	2.579	-0.944	-2.385	0.946
	(0.300)	(0.260)	(0.215)	(0.235)	(0.217)
Constant	-0.108	-2.897	-2.266	0.354	-2.713
	(0.748)	(0.752)	(0.630)	(0.661)	(0.642)
N	634	845	838	844	841

Self-Monitoring: *Continuous from low (0) to high (12).* ***Likes and Dislikes:*** *The dependent variable here is coded as binary with 1 as the most extreme polarization (reporting likes of in-party, but no likes of out-party, dislikes of out-party, but no dislikes of in-party) and 0 otherwise (reporting liking both in-party and out-party or reporting disliking both in-party and out-party). Because two of the controls (media and interest) are missing large amounts of data, they are held at their means for this model estimation. Note that self-monitoring does not predict reporting likes or dislikes on their own—only when these are coded as liking the in-party and disliking the out-party does self-monitoring matter.* ***Like Democrat:*** *Asked if there is anything respondent likes about the Democratic party, with 1 as yes and 0 as no. Logistic regression estimated.* ***Dislike Democrat:*** *Asked if there is anything respondent dislikes about the Democratic party, with 1 as yes and 0 as no. Logistic regression estimated.* ***Like Republican:*** *Asked if there is anything respondent likes about the Republican party, with 1 as yes and 0 as no. Logistic regression estimated.* ***Dislike Republican:*** *Asked if there is anything respondent dislikes about the Republican party, with 1 as*

yes and 0 as no. Logistic regression estimated. ***Self-Monitoring:*** *Self-monitoring is continuous from low (0) to high (1).* ***Liberal Coefficient:*** *The liberal coefficient is in comparison to conservatives; moderates were included in the regression but not the table.*