Letter

Testing the Robustness of the ANES Feeling Thermometer Indicators of Affective Polarization

MATTHEW TYLER   *Rice University, United States*
SHANTO IYENGAR   *Stanford University, United States*

A ffective polarization (AP)—the tendency of political partisans to view their opponents as a stigmatized "out group"—is now a major field of research. Relevant evidence in the United States derives primarily from a single source, the American National Election Studies (ANES) feeling thermometer time series. We investigate whether the design of the ANES produces overestimates of AP. We consider four mechanisms: overrepresentation of strong partisans, selection bias conditional on strong identification, priming effects of partisan content, and survey mode variation. Our analysis uses the first-ever collaboration between ANES and the General Social Survey and a novel experiment that manipulates the amount of political content in surveys. Our tests show that variation in survey mode has caused an artificial increase in the mixed-mode ANES time series, but the general increase in out-party animus is nonetheless real and not merely an artifact of selection bias or priming effects.

INTRODUCTION

There is general agreement that affective polarization (AP)—the tendency of partisans to view their opponents as a stigmatized "out group"— has increased significantly in the United States (US). The relevant survey evidence comes primarily from the American National Election Studies (ANES) feeling thermometer time series. In this note, we consider the possibility that the finding of increased polarization is attributable to distinctive features of the ANES design. Specifically, we investigate the role of selection bias resulting in the oversampling of more polarized partisans, priming effects that elevate the salience of respondents' party affiliation, and changes in survey mode.

The possibility of survey artifacts in the thermometer data has clear implications for the study of American politics. To the extent the data overstate polarization, concerns over any resulting spillover effects, such as the erosion of support for democratic norms and the potential for political violence (e.g., Kalmoe and Mason 2022; Westwood et al. 2022), are premature. Our results show, however, that the broad trends in feeling thermometer results are likely robust to both selection bias and the priming of partisan identity. While the use of an online survey mode has led to a moderate inflation of the mixed-mode ANES time series, the phenomenon of increased animus toward political opponents is real.

Corresponding author: Matthew Tyler [ID], Assistant Professor, Department of Political Science, Rice University, United States, mdtyler@rice.edu.

Shanto Iyengar [ID], William Robertson Coe Professor and Professor of Political Science and of Communication, Department of Political Science, Stanford University, United States, siyengar@stanford.edu.
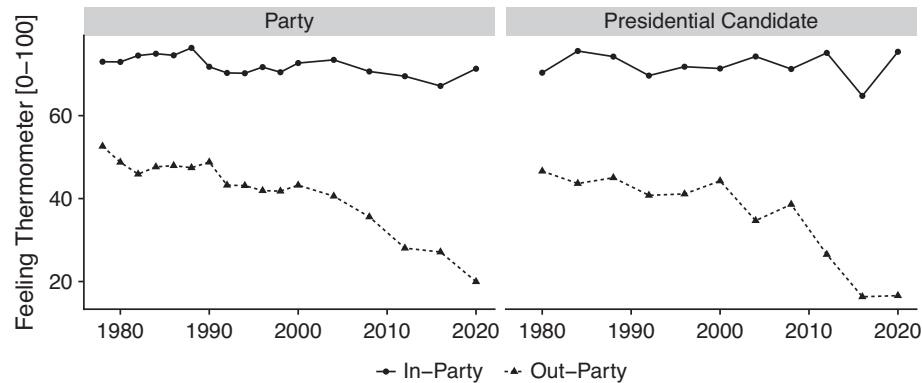
BACKGROUND

Survey self-reports are the most widely used measure of AP. What some have called the "workhorse" indicator (Iyengar et al. 2019) is the feeling thermometer question from the ANES time series. The feeling thermometer battery dates back to 1970 and asks respondents to rate a variety of political leaders and groups on a scale ranging from cold (0) to warm (100). Over the years, ANES respondents have used the thermometer to rate Democrats and Republicans (or the Democratic and Republican Parties). Scholars typically compute the difference between the thermometer score given to the party of the respondent and the score given to the opposing party as the operational measure of AP. In some instances, scholars calculate the corresponding difference in the thermometer ratings of the presidential candidates rather than the parties. As shown in Figure 1, both measures show a substantial increase in AP since 1978; the in-group versus out-group difference for the party thermometer rises from 24 degrees in 1980 to 51 degrees in 2020. The corresponding difference for the in-group and out-group candidate thermometer ratings (e.g., rating for Bush/Gore in 2000) increases from 24 degrees in 1980 to 59 degrees in 2020.

As Figure 1 makes clear, the rise in AP is not attributable to any major movement in partisans' positive feelings for their side, but rather, to the intensification of their dislike for the opposition. The share of ANES partisans expressing extreme negativity for the out party (ratings of 0) remained quite small leading up to and during 2000. Since 2000, however, the size of this share has increased dramatically—from 8% in 2000 to 40% in 2020. Thus, over the first two decades of this century, partisans' mild dislike for their opponents metastasized into a deeper form of animus.

FIGURE 1. ANES Feeling Thermometer Time Series

DESIGN-BASED EXPLANATIONS FOR INCREASED PARTISAN ANIMUS

Survey responses are notoriously responsive to properties of the survey instrument including question wording and ordering, the mode in which the survey is administered, and possible biases in sample composition. There is a vast literature documenting these design effects on a variety of political attitudes (e.g., Palmer and Duch 2001).

Throughout, we benchmark the ANES against the General Social Survey (GSS), treating the GSS as a non-political gold standard survey. The GSS is a now-biennial nationally representative survey of non-institutionalized adults in the US administered by NORC at the University of Chicago. GSS interviews are predominantly conducted face-to-face with some of the survey refusals instead interviewed over the phone (<10%). The GSS is used extensively throughout the social sciences and is held in high regard by academic researchers and government agencies. Because the GSS is far less explicitly political in content, it is more likely to be robust to the potential problems faced by the ANES. In summary, it is the best available benchmark for the ANES.

The Selection Bias Explanation

Participation in surveys is a voluntary act and, as the number of surveys has proliferated, the level of respondent compliance has correspondingly declined. The ANES traditionally maintained relatively high response rates exceeding 60%. However, in recent years, despite repeated contacts with potential respondents and the provision of significant financial incentives for participation, the response rate has declined to around 50%.[1]

The ANES is an unusually lengthy survey with two waves, each requiring approximately 60 minutes to complete, focusing almost exclusively on elections, voting behavior, and political opinions. The time commitment and subject matter are thus likely to discourage individuals who are relatively uninterested in politics; and indeed, ANES is aware that nonresponders are less likely to be affiliated with a political party, are less educated, and less attentive to politics than responders. One possible bias resulting from this increasing phenomenon of nonresponse is the overrepresentation of relatively opinionated or politically involved individuals. This is the first design effect we test.
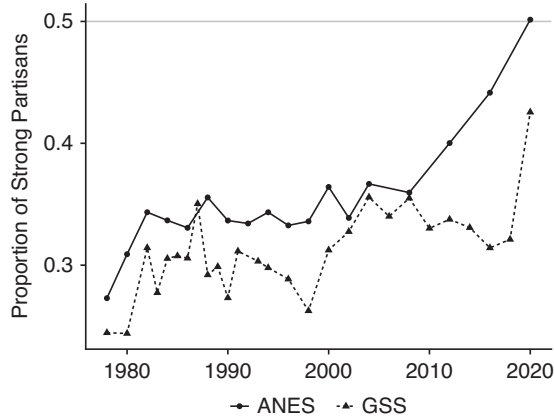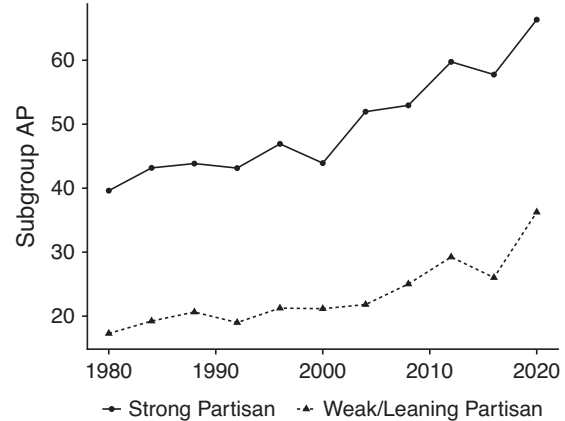
The Priming Explanation

A second possible source of inflation in the ANES estimates of partisan affect is the potential for priming effects induced by repeated exposure to questions with clear partisan valence. In the case of the ANES party thermometers (which are administered during the pre-election wave), respondents have already answered some seventy questions concerning their voting preferences and evaluations of parties/candidates before providing their thermometer scores. In keeping with self-perception theory (Bem 1972), respondents' answers to this lengthy sequence of questions with obvious partisan valence will tend to heighten self-awareness of their partisan affiliation, thereby encouraging polarized responses to the feeling thermometers.

Changes in Survey Mode

The last potential source of bias we consider is the variation in ANES survey modes over time. Prior to 2012, every iteration of the ANES used a face-to-face (FTF) mode with an in-person interviewer. Then, in 2012 and 2016, the ANES used a mixture of FTF and Internet modes. Finally, in 2020, the ANES temporarily dropped the FTF mode (as a response to COVID-19). These changes in survey mode are a potentially significant over-time confounder for any ANES time series analysis that uses all available respondents, which we refer to as the mixed-mode ANES time series.

---

[1] The ANES response rate has declined from a high of around 80% in 1964 to just under 50% in 2020.

FIGURE 2. Proportion of Partisans Classified as Strong Partisans

FIGURE 3. Subgroup AP by PID Strength (ANES Only)

To be clear, we do not take a firm stance on whether any particular mode is superior to another. It is plausible that respondents provide *less* sincere feeling thermometer measures in the FTF mode because of the presence of an interviewer. In that case, the Internet mode may be more accurate. However, the Internet mode is not observed prior to 2012, so our estimate of the long-run increase in AP will be confounded when we compare the 1980 ANES FTF to the 2020 ANES Internet if the two modes are systematically different. This potential over-time confounding is the principal reason we investigate mode effects.

TESTING THE POSSIBLE DESIGN EFFECTS

Selection Mechanisms

We begin our analysis by testing the selection bias explanation. Because of the availability of GSS data, it is useful to decompose this bias into two distinct mechanisms. The first mechanism, discussed immediately below, concerns the overrepresentation of strong partisans in the ANES relative to the GSS benchmark. The second mechanism, discussed further below, concerns the selection of more polarized respondents *conditional* on strong party identification.

Figure 2 compares the strong party identifier time series in the ANES and GSS, making it clear that the ANES often overstates the prevalence of strong partisans relative to the GSS benchmark.[2] Since the mid-1990s, ANES respondents are drawn increasingly from the ranks of strong partisans, so much so that strong partisans are the modal response category in 2020. The increasing number of strong partisans appears to be distinctive to the ANES surveys. The GSS—which

features many fewer questions with obvious political content[3]—shows a markedly different distribution of party identification even though both surveys use identical question wording. The figure shows that the ANES has seen a significant over time increase in the proportion of strong partisans. In contrast, the proportion of strong partisans found in the GSS has remained approximately stable—and drops relative to the early 2000s—over the same period until a surge in 2020–21.[4]
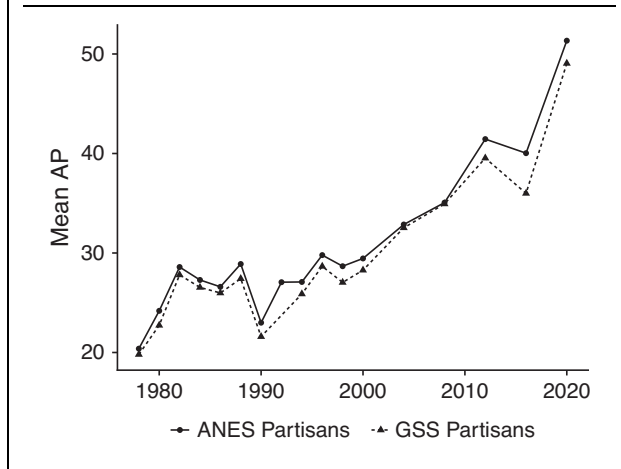
Since strong partisans are significantly more polarized than weak/leaning partisans, this overrepresentation mechanism is potentially important. On the one hand, Figure 3 shows that strong partisans are much more polarized (have higher AP levels) on the feeling thermometer measure than weak partisans in each wave of the ANES. On the other hand, both groups have polarized substantially and proportionally between 1978 and 2020.

We cannot know the full impact of this overrepresentation of strong partisans until we translate the benchmark distribution of strong partisanship onto a counterfactual time series. That is, we need to calculate what the ANES time series would look like under the GSS distribution of strong partisanship. Figure 4 plots the AP time series for the ANES alongside the counterfactual time series "GSS Partisans" that uses the GSS distribution of strong partisanship. This counterfactual time series is not a causal counterfactual; it is

---

[2] We omit non-citizen GSS respondents from all analyses so that the ANES and GSS estimates reported here correspond to similar target populations.

[3] For instance, the 2021 GSS questionnaire included only three items with clear partisan valence (party identification and two questions about the 2016 election). In contrast, the 2020 ANES includes over 40 items with clear partisan valence. More generally, the GSS staff estimates that approximately 30% of the survey content references broadly defined political subject matter. The corresponding figure for the ANES survey is in excess of 60%.

[4] We speculate that the sudden increase in the share of strong partisans in the 2020–21 GSS may be related to mode effects. Because of the pandemic, the study was fielded online instead of via FTF interviews. To support this hypothesis, Figure B6 in Appendix B of the Supplementary Material shows that there are significant mode differences for the proportion of strong partisans in the ANES.

FIGURE 4. Mean AP Time Series with Alternative Partisan Compositions

merely a weighted average of the ANES strong partisans and weak/leaning partisans, weighting by their prevalence in that year's GSS survey. Formally, "GSS Partisans" = $\mathbf{E}$[AP | ANES, Strong] $\mathbf{P}$[Strong | GSS] + $\mathbf{E}$[AP | ANES, Weak/Leaning] $\mathbf{P}$[Weak/Leaning | GSS] within a given year.

We can use this figure to test the overrepresentation of strong partisans mechanism in isolation. If the subgroup AP levels are similar across the ANES and GSS —an assumption we test below—then this figure shows that compositional differences between the ANES and GSS have only a minor impact on the overall mean AP. In fact, the increase from 1978 to 2020 is only about 5.5% less using the GSS partisan composition (29.3 vs. 30.9). This indicates that, if strong partisans are overrepresented in the ANES, then the bias that overrepresentation introduces by itself in the AP time series is relatively small in magnitude.

Having documented differences in the partisan composition of the ANES and GSS samples, we turn to the second mechanism: selection bias conditional on strong partisanship. This mechanism stipulates that strong and/or weak/leaning partisans in the ANES are more polarized than the equivalent groups in the GSS. More specifically, they appear more polarized while exposed to the same questionnaire (particularly, the ANES questionnaire).[5] If this mechanism holds, then the ANES would be overstating AP for reasons uncorrelated with strong partisanship.

To test for this mechanism, we compare the AP levels across 2020 ANES and 2020 GSS respondents, a subset of whom answered the ANES post-election questionnaire. Since we already know the ANES has a higher share of strong partisans, we focus on whether the strong or weak/leaning partisans in the ANES are more polarized than those in the GSS. Typically, the GSS does not include political feeling thermometer items

but, in 2020, the ANES and GSS conducted a collaborative study in which eligible GSS 2016–20 panel respondents were invited to fill out the ANES post-election questionnaire. The GSS respondents were re-recruited by NORC so as to replicate the GSS selection process as closely as possible.[6] The GSS respondents were interviewed concurrently with the 2020 ANES post-election respondents during November 2020 to January 2021. A total of 1,164 GSS respondents completed the ANES post-election survey out of 1,734 invited GSS panelists (67%).

This first-ever reinterview of GSS respondents by ANES gives us a direct test of the degree to which strong partisans or weak/leaning partisans in the ANES sample are more polarized than those in the GSS. While the standard GSS questionnaire does not include the feeling thermometer battery, the 2020 ANES post-election battery includes feeling thermometer measures for the presidential candidates (Biden and Trump, in this case).[7] This allows us to test whether the level of AP is different across studies conditional on strength of party identification. If we find no such differences, then that is consistent with no selection bias conditional on strong party identification.

Figure 5 shows the average AP levels among strong and weak/leaning partisans for both the 2020 ANES and the joint study respondents recruited from the GSS 2016–20 panel. In fact, the differences across the samples are *negative* and/or small in absolute terms: –3.7 (2.6) for strong partisans and 2.2 (3.1) for weak/leaning partisans. Statistically, we can reject differences across the samples that would be large enough to cast doubt on the ANES feeling thermometer time series.

These results suggest that, although the ANES has a noticeably higher share of strong partisans than the GSS, strong partisans exhibit similar levels of polarization whether they originate from the ANES or GSS sample. Similarly, weak/leaning partisans look quite similar across the two samples. Thus, we find no evidence indicating that strong or weak/leaning partisans who participate in the ANES are more polarized than their counterparts in the GSS. After accounting for strong partisanship, selection effects do not appear to have a meaningful impact on the feeling thermometer scores.
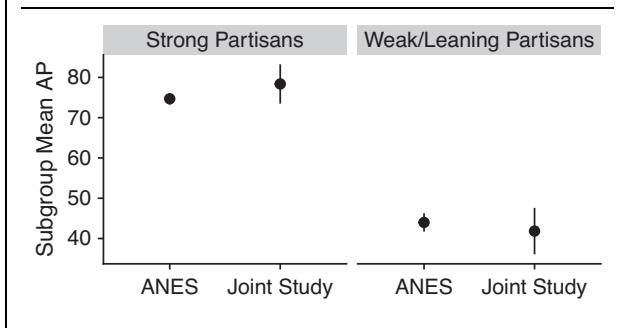
Priming of Partisanship as an Explanation

Priming effects are the third mechanism we test. While the selection bias mechanisms describe a difference in the average ANES and GSS respondent, priming effects are the result of varying the questionnaire respondents are exposed to. A priming effect in this case would cause respondents to self-report higher levels of AP after being exposed to survey content with

---

[5] We revisit whether it is necessary to stress the role of the questionnaire with the third mechanism.

[6] Figure B5 in Appendix B of the Supplementary Material confirms that the joint study, just like the regular GSS samples, contains fewer strong partisans than the 2020 ANES.

[7] Party and candidate affect are highly correlated in surveys where both are observed (0.77 for Biden and 0.78 for Trump, weighted). Figures B3 and B4 in Appendix B of the Supplementary Material plot the conditional relationships.
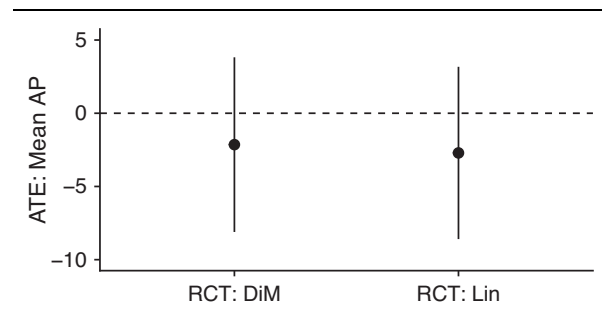
FIGURE 5. Subgroup Mean AP (2020–21)

FIGURE 6. The Effects of Political Content on AP Levels

Note: See models 3 and 4 of Table B1 in Appendix B of the Supplementary Material for the full results.

clear partisan valence. If this holds, then the ANES measure of AP could be interpreted as artificially inflated relative to an individual's "resting" AP level.

Ideally, we would test this mechanism by randomly assigning respondents to take a highly political survey like the ANES (treatment) or a survey with no political content (control) and then record their feeling thermometer outcomes. This would tell us if differences in the questionnaire has an effect on AP levels. Rivers and Nail (2020) performed such an experiment to examine the priming effects of lengthy political surveys on polarized economic evaluations. We replicated the design of Rivers and Nail (2020), but significantly increased the amount of exposure to political content so as to more closely approximate the experience of ANES respondents.[8] We assigned participants to complete a 10–12 minute survey focusing on either political or non-political subject matter.

In the political condition, respondents answered approximately 40 questions used in the ANES concerning evaluations of Donald Trump and Joe Biden, past presidential voting choices, confidence in the two parties' ability to handle major national problems, and positions on several policy issues (including gun control, climate change, health care, and immigration). At the end of the survey, respondents encountered the feeling thermometer questions asking for their evaluations of the two parties. The survey instrument in the non-political condition focused exclusively on consumer products and respondents' recent and anticipated future purchasing behavior. We asked respondents about their cell phone manufacturer, their cell phone carrier, and the amount of time they spend on their phone. They also answered a battery of questions about their leisure activities, in addition to their culinary and beverage preferences.

In both conditions, the standard ANES seven-point party identification question appeared at the 5-minute mark, approximately 7 minutes before the appearance

of the standard feeling thermometer questions.[9] We preregistered this survey experiment, our hypotheses, and our analyses in advance on the Open Science Framework.[10]

We fielded the experiment in May 2022 using the YouGov platform. YouGov interviewed 1,639 respondents who were then matched down to a sample of 1,000 (500 in each condition) to produce the final dataset as part of YouGov's standard algorithm. The respondents were matched to a sampling frame drawn from the 2019 ACS based on gender, age, ethnicity/race, and education.
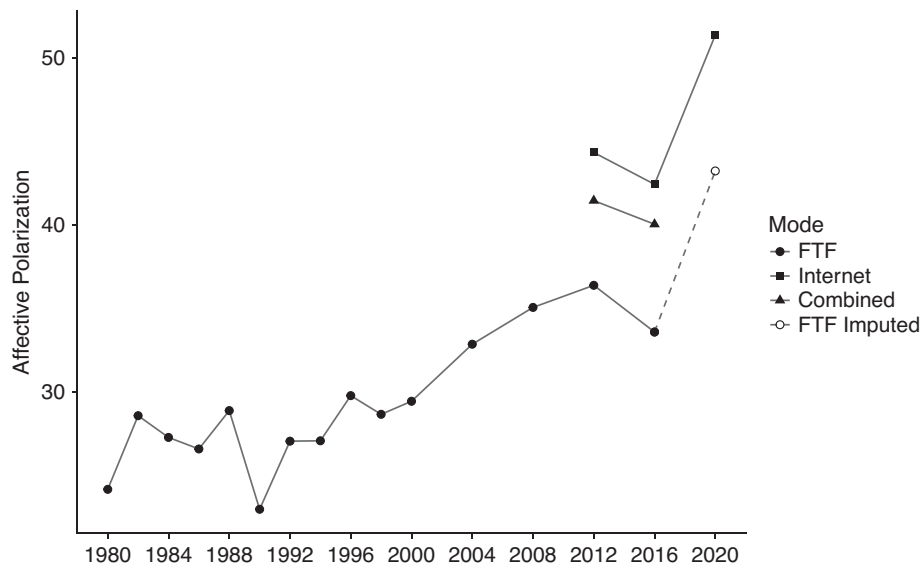
Figure 6 provides a coefficient plot for the treatment effects on AP levels. We plot both the difference-in-means (DiM) and the Lin (2013) covariate-adjustment estimate (Lin). Neither estimate is positive—the Lin estimate is −2.7 (3.0)—and we can reject substantively large positive effects.

As an additional analysis, Appendix E of the Supplementary Material compares feeling thermometer measures taken from a new ANES module inserted into the 2020 wave of the 2016–20 GSS panel with thermometer measures from the subsequent ANES post-election survey. In theory, this allows us to test for priming differences between the ANES and GSS questionnaires by looking at within-respondent changes between surveys. We refer readers to the Supplementary Material for a discussion of the potential flaws of this comparison as a test for priming effects given that the questionnaires were administered with differing proximity to a potentially polarizing election. With that said, the confidence intervals from this analysis overlap with the experimental confidence intervals of Figure 6 in the region associated with substantively small priming effects (1.2–3.3).

---

[8] We do not replicate the ANES questionnaires exactly due to space constraints (the ANES is an exceptionally long survey) and the difficulty in replicating the logical flow of the ANES questionnaire. For the full list of questions with short descriptions included in the two conditions, see Appendices D.1 and D.2 of the Supplementary Material.

[9] Figure B2 in Appendix B of the Supplementary Material shows that we detect no effect of the treatment on strength of party identification using this measure. However, note that, to define respondent-level AP, we use the party identification variable asked by YouGov before our survey was administered. This guarantees that the treatment and control groups are comprised of (on average) equivalent groups of partisans.

[10] https://osf.io/4b92u/?view_only=3f755d19992c4890b92e7d668babb5bf.

FIGURE 7. Affective Polarization Estimates Conditional on Mode Analyzed

All told, respondents who encountered an almost exclusively non-political survey before providing their thermometer scores appear basically just as polarized as those who answered a lengthy series of political evaluations with clear partisan valence. There is no evidence of large priming effects.

Mode Effects

Up to this point, we have ignored survey mode and used all ANES respondents in a given year. Our final analysis examines whether variation in survey mode has confounded the ANES time series. Taking advantage of the fact that both FTF and Internet modes are present in 2012 and 2016, we can assess whether there are differences in average AP levels across modes. Figure 7 plots the FTF-only, Internet-only, and combined estimates of average AP levels over time. The estimates are based on mode-specific weights to account for potential mode-related unit nonresponse. Therefore, in the absence of mode effects, the three estimates should converge on the same quantity.

Figure 7 shows that there are substantively large mode differences in 2012 and 2016. The gap between the FTF-only and Internet-only estimates is 7.9 in 2012 and 8.8 in 2016. Explaining what exactly is causing this mode difference is beyond the scope of this analysis, but it is clear that our estimate of AP does depend, at least to some degree, on the survey mode. Polarization appears lower in the FTF mode but heightened in the online mode. Figure B6 in Appendix B of the Supplementary Material shows that the Internet mode similarly finds a higher share of strong partisans than the FTF mode.

To assess the potential contribution of mode effects to the historical trend in AP, we can leverage the 2016–20 ANES panel respondents. A large number of Democrats and Republicans who completed the 2016 ANES

online subsequently completed the 2020 ANES online ($n = 1,747$). Because we observe their AP levels with the same survey mode—the online component of the 2020 ANES closely matches the online component of the 2016 ANES—we have measures of within-respondent 2016–20 changes in AP levels which are unaffected by 2016–20 mode changes. The average within-respondent increase in AP levels between 2016 and 2020 among this sample is a substantial 9.0. Incorporating these data into an imputation procedure detailed in Appendix F of the Supplementary Material, we estimate that the 2016 *FTF* respondents, if reinterviewed in 2020, would have provided AP levels that were 9.6 points higher in 2020 than 2016. For visual comparison, we depict this imputed FTF 2016–20 trend in Figure 7 as "FTF Imputed." Stepping back, we find that the 1980–2020 increase in this combined FTF and "FTF Imputed" AP time series amounts to 18.9 points, which is smaller than the mixed-mode increase over the same period of 27.1 but still substantial and, importantly, still increasing.

CONCLUSION

We have subjected the standard feeling thermometer measure of AP to a sequence of robustness tests. These tests show that the ANES feeling thermometers provide largely reliable measures of AP over time. Of the first three design-related mechanisms we enumerated, we only find support for the selection of partisans with stronger party attachments into the ANES. We observe this in both the time series data and in the ANES–GSS joint study. However, the impact of this selection effect on the overall increase in AP is likely to be small because both strong and weak/leaning partisans have polarized at similar rates over time. Moreover, the

selection effects on strength of party identification are relatively modest. Therefore, we observe only a 5.5% reduction in the long-run AP trend when we substitute the GSS composition of strong partisans (Figure 4).

We do not find supporting evidence for meaningful priming effects. Magnifying the political content of surveys does not appear to elevate the level of AP. Participants exposed to minimal political content respond basically no differently to the thermometer questions than their counterparts exposed to a lengthy battery of political questions with clear partisan valence.

Yet we do find significant differences in estimated AP across survey modes. In overlapping years, the Internet-only estimate of average AP is about 8–9 points higher than the FTF-only estimate of average AP. We are unable to adjudicate over the accuracy of the FTF versus online estimates without some behavioral indicator of out-party animus. Online respondents may be overstating their hostility toward the out group, whereas FTF respondents may be doing the opposite. All we can conclude is that with the advent on online interviews, there is an upward shift in the long-run AP trend for the mixed-mode ANES time series.

Overall, our results demonstrate that the ANES feeling thermometers are largely robust to concerns over both selection bias and priming effects associated with exposure to a lengthy political survey. Mode effects probably shift the long-run AP trend upward when the Internet mode is included in the analyses but, fortunately, mode effects can be obviated by leaving Internet respondents out of the full time series analysis. In all, the feeling thermometers constitute an invaluable time series on the state of political polarization. Based on the evidence presented here, we encourage their continued use in the years ahead.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055423001302.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/G9OKGB.

REFERENCES

Bem, Daryl J. 1972. "Self-Perception Theory." In *Advances in Experimental Social Psychology*, Vol. 6, 1–62. New York: Elsevier.
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.
Kalmoe, Nathan P., and Lilliana Mason. 2022. *Radical American Partisanship: Mapping Violent Hostility, Its Causes, and the Consequences for Democracy*. Chicago, IL: University of Chicago Press.
Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *Annals of Applied Statistics* 7 (1): 295–318.
Palmer, Harvey D., and Raymond M. Duch. 2001. "Do Surveys Provide Representative or Whimsical Assessments of the Economy?" *Political Analysis* 9 (1): 58–77.
Rivers, Douglas, and Stephanie A. Nail. 2020. "Partisan Polarization and Retrospective Voting." Paper presented at the Annual Meeting of the American Political Science Association.
Tyler, Matthew, and Shanto Iyengar. 2023. "Replication Data for: Testing the Robustness of the ANES Feeling Thermometer Indicators of Affective Polarization." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/G9OKGB.
Westwood, Sean J., Justin Grimmer, Matthew Tyler, and Clayton Nall. 2022. "Current Research Overstates American Support for Political Violence." *Proceedings of the National Academy of Sciences* 119 (12): e2116870119. https://doi.org/10.1073/pnas.2116870119.