[Intro Here]

Affective Polarization: Trends and Theories of its Origins

101-point feeling thermometer ratings are the most widely-used affective polarization measure (Iyengar et al. 2019). The archetypical illustration of affective polarization's rise in the US is shown in Figure 1, where average inparty/outparty ratings and differences in these ratings are plotted using American National Election Study (ANES) Time Series data (Iyengar, Sood, and Lelkes 2012). Partisans' average inparty ratings have been high and mostly stable since the 1980s, while outparty ratings have trended from a neutral 47-points in 1980 to a cold 19-points in 2020. Due to stable inparty ratings and declining outparty ratings, the difference in inparty-outparty ratings—affective polarization—increased from 26-points in 1980 to 52-points in 2020.

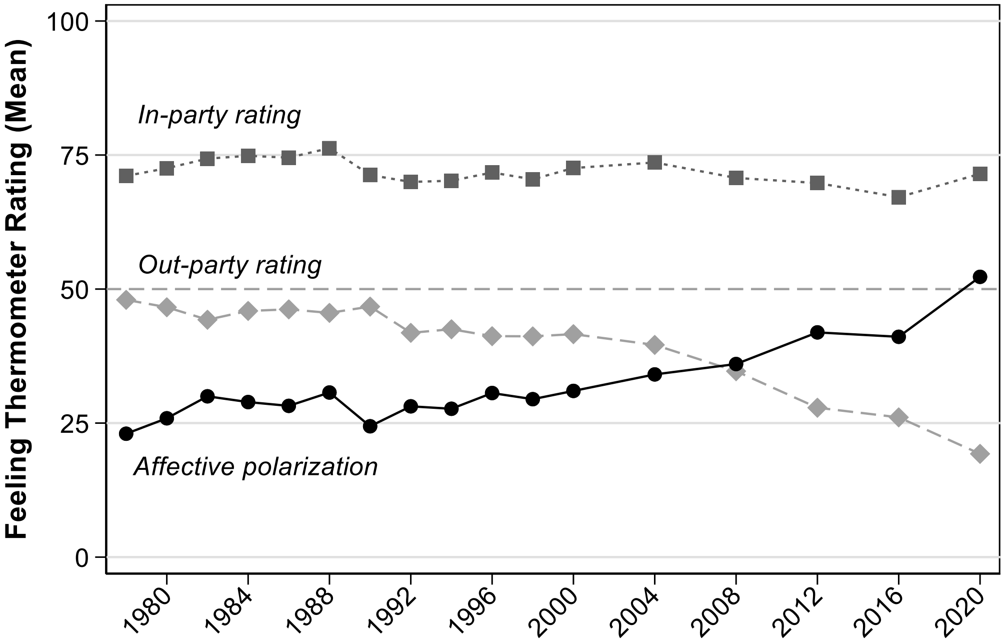

Figure 1—Trends in Party Ratings and Affective Polarization (1978-2020). Points are mean inparty and outparty 101-point feeling thermometer ratings and the differences between these means (affective polarization) over time. Data are weighted. Estimates include partisan-leaning independents. Source: American National Election Study Time Series Cumulative Data File.

Examining feeling thermometer ratings and other measures (Iyengar, Sood, and Lelkes 2012; Iyengar and Westwood 2015; Druckman and Levendusky 2019), a general consensus has emerged among political scientists that affective polarization has risen the last half century, with trends being driven mostly by rising outparty animosities (Iyengar et al. 2019). Today, debates in the affective polarization literature largely center on its causes and consequences. In terms of its causes, extant research points to declining cross-cutting social cleavages (Mason 2018), partisan-ideological sorting (Levendusky 2010; Mason 2015), ideological polarization (Rogowski and Sutherland 2016), and changes in the information environment, particularly regarding the rise of online and social media (Lelkes, Sood, and Iyengar 2017).

Further, predominant theories of affective polarization propose that few, if any, guardrails curtail Americans' expressions of outparty hate. In a recent review, Iyengar et al. (2019) contend that "Unlike race, gender, and other social divides where group-related attitudes and behaviors are subject to social norms (Maccoby & Maccoby 1954), there are no corresponding pressures to temper disapproval of political opponents" (see also Iyengar and Westwood 2015). Lelkes (2016) similarly proposes that "social norms seem to pressure individuals to overstate their feelings of [outparty] antipathy." Thus, although scholars may disagree about the relative weight of various factors that might explain the rise in affective polarization, there is near-universal agreement that the trends in Figure 1 reflect real changes in partisans' feelings toward political groups over time.

The Measurement of Affective Polarization and Survey Mode

Feeling thermometer affective polarization measures have been the subject of significant scrutiny given their ubiquitous usage. For example, Druckman and Levendusky (2019) show that respondents often bring a party's elites, not voters, to mind when rating parties, while Klar, Krupnikov, and Ryan (2018) raise the issue that affective polarization measures may mistake an increased distaste for partisan politics for increased outparty animosity. Most germanely, Iyengar and Tyler (n.d.) assess the robustness of ANES feeling thermometer measures to three possible survey artifacts that risk confounding the observed increase in affective polarization with this measure: oversampling of strong partisans, selection bias of the affectively polarized conditional on strong partisan identity, and priming effects from gauging party ratings after respondents have answered many questions about politics. Iyengar and Tyler (n.d.) conclude “ANES estimates of affective polarization have negligible upward bias” and that “the level of out-party animus is real and not an artifact of selection bias or priming effects.” Thus, while not necessarily perfect for all use cases, feeling thermometer measures are taken as capturing real trends in partisan hostilities.

Although recent work has helped validate the feeling thermometer measure of affective polarization, the role of *survey mode* (e.g., in-person, online, phone, etc.) in shaping measures of affective polarization is largely unstudied. Given ongoing trends in the survey research landscape away from live interviews towards self-administered online surveys, it is important to understand whether and how mode affects the measurement of affective polarization. This is especially true for comparisons over time like those in Figure 1 because the share of ANES interviews fielded online was 0% in 2008 and all years prior, but increased to 65% in 2012, 72% in 2016, and 94% in 2020. And although the high affective polarization between 2012 and 2020 could be consistent with increased affective polarization, these highs may also be partly attributable to *mode effects.*[1]

Mode could matter in two ways when assessing affective polarization. First, mode could affect selection into the sample because the mode a survey is fielded in determines who appears in the sample frame (i.e., coverage) and the probability of respondents taking the survey if they

[1] Differences in mode across surveys analyzing affective polarization over time are not limited to the ANES; Iyengar, Sood, and Lelkes (2012), for example, note their comparison of Americans’ feelings about their child marrying an outpartisan (i.e., a social distance measure) from an online 2008 poll to a face-to-face 1960 poll could also be affected by these differences in survey mode.

are invited to do so (i.e., unit non-response). Selection effects matter when making population-level inferences about affective polarization because surveys fielded using different modes risk sampling Americans who differ distributionally in terms of affective polarization (see also Cavari and Freedman 2023). For example, while not an assessment of differences across modes, Iyengar and Tyler (n.d.) test whether the general overrepresentation of strong partisans in the ANES Time Series has caused it to overestimate increases in affective polarization and find a small degree of sampling bias. In similar fashion, the ongoing shift towards online surveys could lead researchers to overestimate increases in affective polarization over time due to mode-based sampling effects.

Mode could also matter when assessing affective polarization due to *measurement effects*. Whereas sampling effects pertain to differences in the respondents who take a survey based on its mode, measurement effects pertain to differences in how respondents answer questions based on a survey's mode (Groves et al. 2009). To offer a quintessential example of measurement bias, white Americans report less liberal racial attitudes when surveyed online compared to in-person, consistent with social pressures against the expression of inegalitarian racial attitudes (Abrajano and Alvarez 2019). Other measurement effects related to mode include acquiescence, attentiveness, and response order effects (Bowyer and Rogowski 2017; McClendon 1991). And while some work has examined whether sampling bias inflates estimates of affective polarization (Iyengar and Tyler, n.d.; Cavari and Freedman 2023), it remains wholly unclear whether mode causes measurement effects. This is an important gap in the literature because the main approach for addressing sampling effects—weighting—is inappropriate for handling measurement effects.

In-Person vs. Online Distributions of Affective Polarization

Do the observed distributions of affective polarization measured by feeling thermometers differ across the two main ANES modes: live, in-person interviews and self-administered online surveys? Unambiguously, yes. In Figure 2, I display the distributions of affective polarization in the 2016 ANES for respondents interviewed in-person (n=1,029) and online (n=2,581), binning the -100 to 100 scale in tens (with the leftmost bins grouping together all respondents who rated their inparty colder than their outparty, i.e., those between -100 and -1 on the scale).[2]

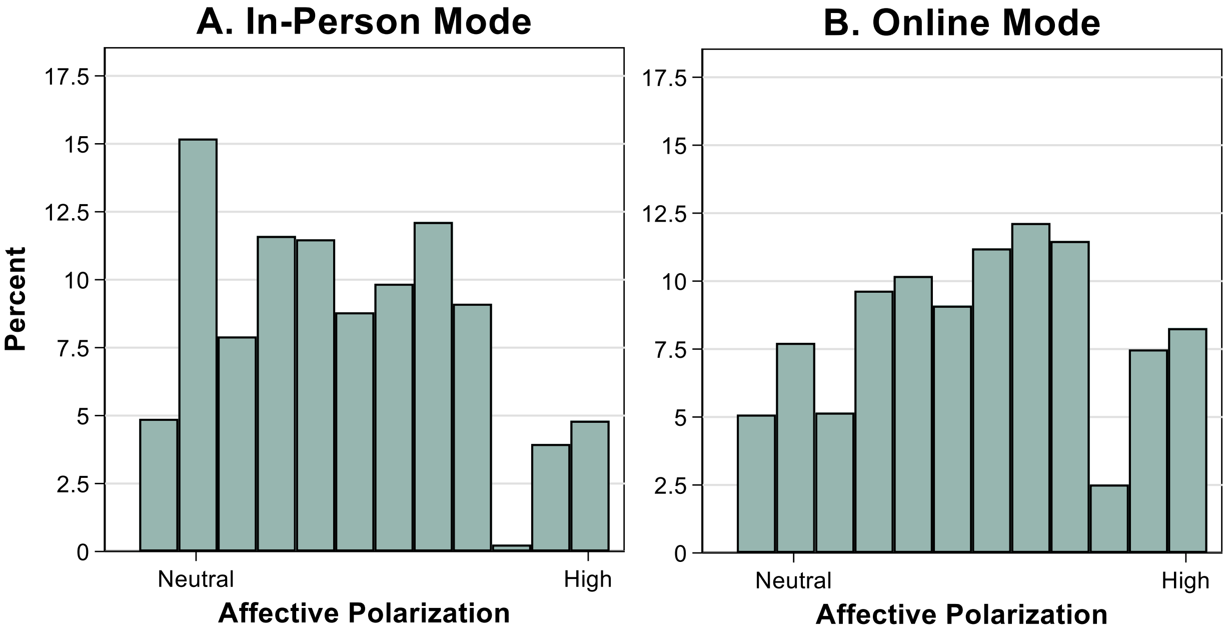

Figure 2—Distributions of Affective Polarization by Survey Mode (2016 ANES). Leftmost bins group all respondents who rate their inparty colder than their outparty. The second neutral bin includes respondents who rate the parties equally. Each subsequent bin is a 10-point range of affective polarization. Data are weighted. Includes partisan-leaning independents. Source: 2016 American National Election Study Time Series.

There are three immediate takeaways from Figure 2. First, the median level of affective polarization is substantially higher online than in-person (45- vs. 30-points), as are the means (44- vs. 34-points). Second, the share of respondents rating the two parties exactly equally is smaller online than in-person (7.7% vs. 15.2%). Third, the share of respondents rating the parties

[2] The 2012 ANES also includes large in-person and online samples. Similar results emerge in the 2012 ANES; affective polarization is more pronounced among online respondents (Appendix []).

at opposite scale extremes is larger online than in-person (8.2% vs. 4.8%). Overall, 2016 ANES online respondents report significantly higher affective polarization than in-person respondents.

Disentangling the Sampling and Measurement Effects of Survey Mode

Cross-sectional comparisons of mixed-mode survey data such as those in Figure 2 cannot distinguish between selection effects and measurement effects. Although cross-sectional studies often attempt to disentangle these effects, standard approaches are methodologically fraught. The typical cross-sectional approach uses regression or matching to model selection into modes and extract the residual measurement effect. However, this approach is unbiased only if the variables used to model selection are *mode insensitive* because they are measured in different modes (i.e., post-treatment). In practice, cross-sectional studies trade-off between allowing selection bias and conditioning on mode sensitive variables. A priori, it is difficult to weigh these biases, and there is unlikely to be a model or matching scheme that avoids both selection and post-treatment bias.

I instead use the 2016-2020 ANES panel and *difference-in-differences* (DiD) to identify the measurement effects of mode on affective polarization (Ollerenshaw 2023). The 2016-2020 ANES includes 2,670 panelists, 639 of whom were interviewed in-person in 2016 then switched online in 2020, and 2,031 of whom were surveyed online in both 2016 and 2020. The 2016-2020 ANES is thus quasi-experimental with respect to mode; panelists who switch from in-person to online modes across waves constitute a “treatment” group, and those in the online modes in both waves are a “control” group. DiD compares average changes in affective polarization between 2016 and 2020 for the treatment and control groups; the difference in the groups’ changes is an estimate of the measurement effect.[3] Notably, the DiD approach does *not* assume respondents’

[3] More formally, DiD imputes the treated group’s unobserved counterfactual trend in untreated outcomes using the observed trend in untreated outcomes for the control group.

mode assignments in 2016 are unrelated to their levels of affective polarization; instead, the DiD approach assumes mode assignments are mean-independent of characteristics related to *trends* in affective polarization—i.e., the parallel trends identification assumption.

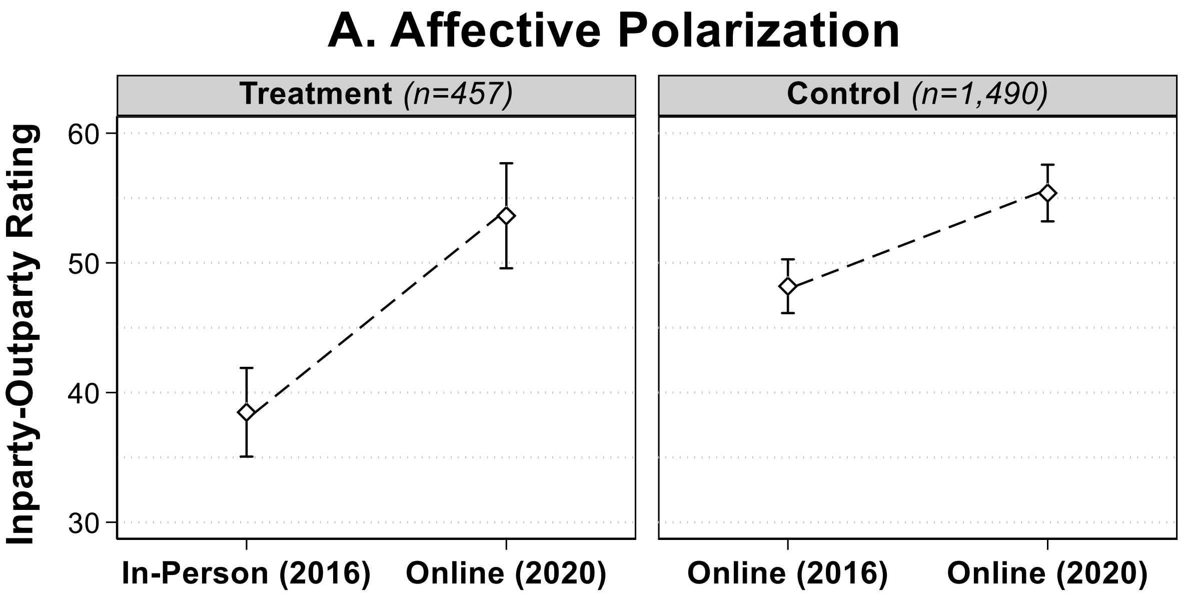

Figure 3—Difference-in-Differences of Affective Polarization by Modes (2016-2020 ANES). Points are mean differences in inparty-outparty ratings (affective polarization) by panel waves with 95 percent confidence intervals. Data are weighted. Estimates include partisan-leaning independents. Source: American National Election Study Time Series Cumulative Data File.

In Figure 3, I plot changes in affective polarization for partisans who switched from in-person interviews in 2016 to online surveys in 2020 (treatment) against partisans surveyed online in 2016 and 2020 (control). Affective polarization increased 15-points for treated partisans but 7-points for control partisans. The 8-point (p<0.001) difference-in-difference-in-means between the groups provides an estimate of the average measurement effect of gauging affective polarization online instead of in-person in 2016. Notably, the measurement effect appears better explained by differences in partisans' outparty ratings than inparty ratings. The estimated effects of online vs. in-person mode are 3-point increases in inparty ratings and 5-point decreases in outparty ratings. Further, the difference between the groups in 2020, when all respondents are surveyed online, provides an estimate of the selection mode effect—2-points (p=0.454). Consistent with the cross-sectional analysis, this quasi-experimental DiD analysis shows partisans report higher affective polarization online than in-person. Additionally, this analysis shows most of this discrepancy is explained by measurement effects, not sampling effects.

Data limitations preclude many DiD robustness checks, such as pre-trend placebo tests, because these require more than two-waves (which are unavailable). However, the parallel trends assumption can still be bolstered by obviating covariate imbalances between the treatment and control groups since imbalances could be related to outcome trends (Imai, Kim, and Wang 2021). While covariate balancing is fraught in cross-sectional studies due to possible mode sensitivity, I can balance on covariates measured online for all respondents in 2020. Specifically, I use entropy balancing to reweight the sample to the third moments of race/ethnicity, gender, age, education, income, political interest, political knowledge, partisanship, ideological identity, and internet access (Hainmueller 2012). The estimated measurement effect is 7-points when using the entropy balancing weights, a negligible difference relative to 8-points using the ANES sampling weights.

Discussion

My analyses show mode affects the measurement of affective polarization. However, the cause of this effect remains ambiguous. One possibility is that respondents are underreporting outparty animosity in live interviews relative to self-administered online modes due to sensitivity bias. Mode effects when assessing attitudes toward social outgroups, such as racial outgroups, are frequently interpreted as evidence of social desirability bias (e.g., Abrajano and Alvarez 2019). However, most affective polarization research contends there are few social norms against expressing outparty hatred; indeed, a study comparing implicit and explicit measures of partisan animus suggests, if anything, there is a social norm to express *more* outparty animosity (Iyengar and Westwood 2015). My findings thus may be consistent with, but are not dispositive of, there being contexts where outparty animosity is considered socially-undesirable to publicly express. If correct, the substantive implication is that surveys fielded with live interviews underestimate affective polarization, and therefore that the trend toward increased affective polarization shown in Figure 1 is partially confounded by the ANES's growing use of online surveys since 2012.

A second explanation is that respondents are offering less constrained (or more dispersed) party ratings in online modes relative to in-person interviews (Homola, Jackson, and Gill 2016). Increased response dispersion would inflate the differences in inparty/outparty ratings offered by respondents without any corresponding increase in the level of affective polarization respondents intend to "imbue" in their ratings. For example, a respondent wishing to convey the exact same level of partisan animosity might offer a "30" outparty rating in-person but a "25" rating online. Measurement variance can often be corrected with anchoring vignettes (Chevalier and Fielding 2011); although the ANES already uses anchoring vignettes when introducing the thermometer ratings (Appendix A), these vignettes may not be detailed enough to fully address measurement variance since they only emphasize the interpretations of ratings at 0, 50, and 100. If correct, this explanation implies that increases in observed affective polarization when comparing older in-person ANES interviews to recent online surveys are partially confounded by measurement bias.

Conclusion

To be clear, I am by no means claiming affective polarization has not increased in the last half century—a 7-point mode-based measurement effect cannot explain the 26-point increase in affective polarization between 1980 and 2020. However, measurement bias of this magnitude is not trivial.

It remains unclear whether changes in mode extend to other measures of affective polarization. For example, [] finds an increase in social distance affective polarization from [] in

[] to [] in []. Notably, however, the [] survey relied on [] interviewing, whereas the [] survey used self-administered online surveys.

[One More Short Paragraph Here]

References

Abrajano, Marisa, and R. Michael Alvarez. 2019. "Answering Questions About Race: How Racial and Ethnic Identities Influence Survey Response." *American Politics Research* 47 (2): 250–74. https://doi.org/10.1177/1532673X18812039.

Bowyer, Benjamin T., and Jon C. Rogowski. 2017. "Mode Matters: Evaluating Response Comparability in a Mixed-Mode Survey*." *Political Science Research and Methods* 5 (2): 295–313. https://doi.org/10.1017/psrm.2015.28.

Cavari, Amnon, and Guy Freedman. 2023. "Survey Nonresponse and Mass Polarization: The Consequences of Declining Contact and Cooperation Rates." *American Political Science Review* 117 (1): 332–39. https://doi.org/10.1017/S0003055422000399.

Chevalier, Arnaud, and Antony Fielding. 2011. "An Introduction to Anchoring Vignettes." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 174 (3): 569–74.

Druckman, James N, and Matthew S Levendusky. 2019. "What Do We Measure When We Measure Affective Polarization?" *Public Opinion Quarterly* 83 (1): 114–22. https://doi.org/10.1093/poq/nfz003.

Groves, Robert M., Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2nd edition. Hoboken, N.J: Wiley.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20 (1): 25–46. https://doi.org/10.1093/pan/mpr025.

Homola, Jonathan, Natalie Jackson, and Jeff Gill. 2016. "A Measure of Survey Mode Differences." *Electoral Studies* 44 (December): 255–74. https://doi.org/10.1016/j.electstud.2016.06.010.

Imai, Kosuke, In Song Kim, and Erik H. Wang. 2021. "Matching Methods for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science*. https://doi.org/10.1111/ajps.12685.

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 (1): 129–46. https://doi.org/10.1146/annurev-polisci-051117-073034.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76 (3): 405–31. https://doi.org/10.1093/poq/nfs038.

Iyengar, Shanto, and Matthew Tyler. n.d. "Testing the Robustness of the ANES Feeling Thermometer Indicators of Affective Polarization."

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. https://doi.org/10.1111/ajps.12152.

Klar, Samara, Yanna Krupnikov, and John Barry Ryan. 2018. "Affective Polarization or Partisan Disdain?: Untangling a Dislike for the Opposing Party from a Dislike of Partisanship." *Public Opinion Quarterly* 82 (2): 379–90. https://doi.org/10.1093/poq/nfy014.

Lelkes, Yphtach. 2016. "Mass Polarization: Manifestations and Measurements." *Public Opinion Quarterly* 80 (S1): 392–410. https://doi.org/10.1093/poq/nfw005.

Lelkes, Yphtach, Gaurav Sood, and Shanto Iyengar. 2017. “The Hostile Audience: The Effect of Access to Broadband Internet on Partisan Affect.” *American Journal of Political Science* 61 (1): 5–20. https://doi.org/10.1111/ajps.12237.

Levendusky, Matthew. 2010. *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans*. Chicago: University of Chicago Press.

Mason, Lilliana. 2015. “‘I Disrespectfully Agree’: The Differential Effects of Partisan Sorting on Social and Issue Polarization.” *American Journal of Political Science* 59 (1): 128–45. https://doi.org/10.1111/ajps.12089.

———. 2018. *Uncivil Agreement: How Politics Became Our Identity*. The University of Chicago Press.

McClendon, Mckee J. 1991. “Acquiescence and Recency Response-Order Effects in Interview Surveys.” *Sociological Methods & Research* 20 (1): 60–103. https://doi.org/10.1177/0049124191020001003.

Ollerenshaw, Trent. 2023. “A Difference-in-Differences Approach for Estimating Survey Mode Effects.” APSA Preprints. https://doi.org/10.33774/apsa-2023-p5r6j-v2.

Rogowski, Jon C., and Joseph L. Sutherland. 2016. “How Ideology Fuels Affective Polarization.” *Political Behavior* 38 (2): 485–508. https://doi.org/10.1007/s11109-015-9323-7.