A Difference-in-Differences Approach for Estimating Survey Mode Effects

Trent Ollerenshaw
Ph.D. Candidate
Duke University
Department of Political Science
trent.ollerenshaw@duke.edu

Version: July 22nd, 2023
Word Count: 7,247

Abstract: A survey's mode can influence both who takes the survey (selection) and how they respond to its questionnaire (measurement). To distinguish selection and measurement effects, most studies of mode effects use cross-sectional designs. However, cross-sectional designs risk omitted variable bias when the selection process is not fully modeled, but post-treatment bias if the selection process is modeled with variables measured in different survey modes. To address these shortcomings, I propose using *difference-in-differences* with mixed-mode panel surveys to identify measurement effects. Difference-in-differences compares changes in survey responses over time among panelists who switch modes to panelists who do not switch modes. Difference-in-differences can help reduce omitted variable bias without introducing post-treatment bias. I demonstrate the difference-in-differences approach by estimating the effects of completing live interviews vs. online surveys on the measurement of racial attitudes and political knowledge in the 2016-2020 American National Election Study and cognitive functioning in the 1992-2020 Health and Retirement Study.

Acknowledgements: The author thanks Christopher D. Johnston, D. Sunshine Hillygus, Talbot Andrews, Elizabeth Mitchell Elder, and Andrew Trexler for helpful feedback on this manuscript.

A survey's mode can influence both who participates in the survey and how they respond to its questionnaire (Olson et al. 2021; Voogt and Saris 2005). In recent years, the loss of near-universal landline coverage, a precipitous decline in response rates, and the advent of self-administered online surveys have shifted survey research into new modes to maintain viability. Indeed, many widely-used surveys for research on health, economics, culture, and politics have adopted online modes, including the Census Bureau's American Community Survey, the General Social Survey, and the American National Election Study. The consequences of changing modes, and especially the effect of shifting away from live interview modes to online surveys, must be understood to sustain inferences drawn in this rapidly evolving landscape for survey research.

In practice, identifying mode effects is challenging because mode can influence both selection into the survey and participants' responses (i.e., measurement). To decompose selection and measurement effects, studies have generally employed either cross-sectional designs that try to model the selection processes in mixed-mode surveys or experiments that randomize assigned modes. Unfortunately, both approaches have major shortcomings. To avoid bias, cross-sectional designs must account for confounding variables, but they must do so without conditioning on mode sensitive variables measured in different modes (i.e., post-treatment). This is often a catch-22 because many variables related to selection are or could be mode sensitive, so in practice, researchers must weigh omitted variable bias and post-treatment bias. Experiments often suffer similar biases when there is differential selection based on assigned modes, and experiments that try to ensure participation after randomizing modes are rare because this design requires tightly-controlled lab settings (Chang and Krosnick 2010; Endres et al. 2022; Gooch and Vavreck 2019).

In this paper, I propose a difference-in-differences (DD) approach for identifying mode effects that improves on cross-sectional designs in terms of bias reduction while providing a

cost-free, feasible alternative to lab experiments. In Section 1, I review extant cross-sectional and experimental approaches for identifying mode effects. In Section 2, I outline a DD approach for estimating mode effects using mixed-mode panel data where some panelists switch modes across waves while others do not. In Section 3, I apply DD to the simplest case (two waves, two modes) by estimating the effects of in-person vs. online mode on racial attitudes and political knowledge in the 2016-2020 American National Election Study. In Section 4, I extend DD to a more complex case—the 1992-2020 Health and Retirement Study, which switched some respondents from live interviews to online surveys in its 14th and 15th waves. In Section 6, I discuss the benefits and limitations of DD, as well as future applications of DD for identifying mode effects.

1. Surveys, Population Inferences, and Mode Effects

Mixed-mode surveys, particularly those with an online mode, have become ubiquitous (de Leeuw 2005; Olson et al. 2021). Increasingly, surveys use multiple modes to recruit samples that would be difficult to obtain in one mode (e.g., pairing online surveys with phone interviews to recruit respondents without internet access). Further, many longitudinal surveys have adopted mixed-mode designs in the wake of COVID-19, which made in-person interviews infeasible. For example, the American National Election Study, a widely-used survey for research on political behavior, fielded a panel that included in-person and web surveys in 2016, but only web surveys in 2020. Given these ongoing developments, understanding how mode affects survey responses is critical to sustaining the inferences researchers draw from surveys.

Mode Effects: Selection and Measurement

Mode can affect survey responses in two ways: selection effects and measurement effects (Groves et al. 2009). Selection effects describe the myriad factors that make a member of the target population more likely to appear in samples fielded in one mode versus another. Some

individuals, for example, may not appear in an online survey's sample frame if they lack internet access (coverage error; Keeter and Mcgeeney 2015), while others may be less likely to sit down for a live interview than an online survey (unit non-response; Bethlehem 2010). If mode affects respondents' probabilities of appearing in the sample, surveys with different modes can produce diverging inferences unless these imbalances are addressed (e.g., through weighting and/or non-response adjustment; Krueger and West 2014).

Measurement effects, by contrast, pertain to differences in a respondent's behavior *across* modes, i.e., how would the same respondent have answered a question had it been fielded in one mode versus another. For example, a quintessential measurement effect is social desirability bias, which is when respondents provide different responses to the same question asked in different modes due to differing social influences (e.g., underreporting a socially-ostracized attitude in less confidential modes). Other measurement effects include differing respondent acquiescence and attentiveness, interviewer effects, and response order effects like primacy and recency bias (Bowyer and Rogowski 2017; McClendon 1991). Together, selection and measurement effects can cause otherwise identical surveys fielded in different modes to produce different inferences.

As surveys increasingly utilize mixed-modes or fully transition online, identifying mode effects becomes important for contextualizing population-level inferences. This is especially true for longitudinal surveys that risk conflating measurement effects with individual-level change (Cernat and Sakshaug 2021). Unfortunately, measurement and selection effects can be difficult to disentangle. Directly comparing differences in response averages or distributions across modes provides an estimate for *overall* mode effects, but not the share of the effect caused by selection vs. measurement effects. Moreover, the solutions to selection effects (e.g., weighting) are inappropriate for addressing measurement effects, so researchers often seek to separate out

selection and measurement effects. Towards this goal, research on mode effects has used two primary approaches: cross-sectional designs and experiments.

Cross-Sectional Designs

Cross-sectional designs are the most used approach for estimating measurement effects. In the standard cross-sectional design, researchers compare differences in responses to the same survey items fielded in different modes while modeling the selection process with regression (but the logic of cross-sectional designs extends to cross-sectional matching and covariate balancing; e.g., Greenacre 2016; Lugtig et al. 2011; Schonlau et al. 2009). If the selection process can be accurately modeled, cross-sectional designs can in theory identify mode's measurement effects.

Though cross-sectional designs are common because mixed-mode surveys are widely available, these designs rely on arguably untenable assumptions. First, cross-sectional designs must account for all confounding variables. In practice, this is difficult because there are many factors that influence respondents' probability of taking a survey, few of which are measured in surveys. For example, online respondents have different personalities than those who complete live interviews (Valentino et al. 2020), but personality is rarely assessed in surveys and has not been used to model selection despite personality's importance in many areas of social scientific research. Personality is just one of what are many possible unobserved confounds, each of which risks biasing estimated measurement effects in cross-sectional studies.

The solution to omitted variable bias seems straightforward—account for more possible confounding variables. Unfortunately, this solution exists in tension with a second type of bias in cross-sectional designs: post-treatment bias. Confounding variables in cross-sectional designs are usually assessed in surveys fielded in different modes (i.e., post-treatment); the imbalances could thus be caused by selection and/or measurement effects (Vannieuwenhuyze and Loosveldt 2013).

Cross-sectional designs that use post-treatment variables to address omitted variable bias risk introducing post-treatment bias (Montgomery et al. 2018). The dilemma researchers face is that it is difficult to know how best to weigh between omitted variable bias and post-treatment bias. In practice, most cross-sectional designs rely on a strong assumption that the variables used to model selection are *mode insensitive* (i.e., they do not experience any mode-based measurement effects). In this way, cross-sectional estimates risk sensitivity to idiosyncratic modeling decisions and untested assumptions without strong theoretical or empirical justifications (Ho et al. 2007).

Experimental Designs

Given the shortcomings of cross-sectional designs, researchers have used experiments to identify measurement effects. By randomizing modes, experiments can obviate selection bias—an ideal design for isolating mode's measurement effects. In practice, however, experiments also often suffer major deficiencies.

First, as noted by Endres et al. (2022), experimental studies usually randomize modes and then allow respondents to opt to take the survey—a decision respondents will likely make in part based on their assigned mode. Experiments of this sort suffer from selection bias just like cross-sectional designs, despite randomization (Hernán et al. 2013). Few experiments on mode effects have randomized modes *after* respondents agree to participate in the survey; for example, Endres et al. (2022) brought individuals into a lab and then randomly assigned them to an in-person or video interview. There are seemingly just two other experiments that use similar designs: Chang and Krosnick (2010) and Gooch and Vavreck (2019). The dearth of experiments on mode effects where differential selection is not a major concern reflects the costliness of the lab-based design relative to approaches that leverage publicly-available surveys, which are rarely administered in tightly-controlled labs.

However, a concern with lab experiments is that they are essentially infeasible to field on representative samples. The three aforementioned lab experiments were fielded on convenience samples, and while their internal validity is very high, convenience samples often differ from the populations researchers are most interested in (Krupnikov and Levine 2014; Mullinix et al. 2015; Sears 1986). Population-based sampling can improve experimental generalizability (Mutz 2011), but drawing a probability-based sample, inviting respondents into a lab, then randomly assigning modes—all to identify measurement effects—is prohibitively expensive. Lab experiments trade-off a degree of generalizability to maximize internal validity, an understandable decision given the internal validity concerns associated with cross-sectional studies. However, effects derived from lab experiments risk not generalizing if these effects are unevenly distributed as functions of sample characteristics imbalanced vis-à-vis the target population (Druckman and Kam 2011).

2. Identifying Measurement Effects with Differences-in-Differences (DD)

I propose an alternative to cross-sectional and experimental designs for identifying measurement effects—difference-in-differences (DD). DD involves much weaker identifying assumptions than cross-sectional designs and offers a feasible alternative to experiments. In this section, I review the logic of DD for estimating causal effects in observational settings, then outline how DD can be used to identify measurement effects using mixed-mode panel data. I focus on the aspects of DD most germane to identifying measurement effects; see Roth et al. (2023) for a recent synthesis of the rapidly-evolving literature on general applications of DD.

Using panel data, DD compares treated units to untreated units (or units that receive a treatment later; Baker et al. 2022) over time. Although they may start in different places, DD assumes that the treatment and control groups would have shared post-treatment trends absent treatment—the parallel trends assumption (PTA). In the potential outcomes framework for causal

inference (Rubin 2005), control units are used to impute the unobserved post-treatment outcomes for treated units. Any post-treatment divergence in the groups' outcomes thus offers an estimate of the average treatment effect on the treated (ATT). Notably, DD does not assume treatment statuses are orthogonal to characteristics that affect outcome *levels*; thus, it can allow the sorts of selection biases that undermine cross-sectional designs. Instead, DD assumes treatment statuses are mean-independent of characteristics that affect outcome *trends*.

I propose DD can be used to identify mode's measurement effects in surveys where:

(1) Respondents are surveyed in at least two waves (i.e., panel surveys).

(2) Some respondents are surveyed in different modes across waves (i.e., treated units).

(3) Some respondents are surveyed in the same mode across waves (i.e., control units).

I define the target estimand as the ATT of being surveyed in one mode vs. another on average responses. The ATT is estimated by imputing the unobserved untreated outcomes of treated units (those who switch modes) with the observed untreated outcomes of control units (those who do not switch modes).

Several assumptions are necessary to sustain DD inferences. DD assumes that changes in outcomes for control units proxy counterfactual changes in the treated units' untreated outcomes. This assumption can never be directly tested because the treated group's untreated outcomes are unobserved. Parallel pre-treatment trends can bolster a DD design's credibility by showing the treatment and control groups moved together before treatment (Angrist and Pischke 2010), but parallel pre-treatment trends are not necessarily sufficient for the PTA to hold (Kahn-Lang and Lang 2020). Additionally, the PTA is violated if there are exogeneous post-treatment shocks that differentially affect the treatment and control groups since pre-post differences across groups can no longer be fully attributed to the treatment (Angrist and Pischke 2009). DD also assumes stable

unit treatment values (SUTVA) and no anticipatory treatment effects (Malani and Reif 2015). Whether these latter two assumptions hold should be assessed qualitatively using a survey's design. For example, a survey that independently samples individuals from a large population is unlikely to violate SUTVA, and if the survey does not inform respondents of the modes they will later be surveyed in until they have finished the survey at hand, anticipatory effects are unlikely.

Although DD is not a foolproof method for causal inference with observational data, it involves weaker identifying assumptions than cross-sectional designs that often face a trade-off between allowing for omitted variable bias or introducing post-treatment bias (Vannieuwenhuyze and Loosveldt 2013). Additionally, DD provides a cost-free and more generalizable alternative to lab experiments where mixed-mode panel surveys are available. In the following two sections, I illustrate applications of DD in the simplest two-waves, two-modes survey (Section 3) and in a more complex survey involving more than two waves and staggered mode switches (Section 4).

3. 2x2 Difference-in-Differences: Measurement Effects in the 2016-2020 ANES

In this section, I estimate measurement effects in the two-modes, two-waves (2x2) case. I use a widely-used survey for political science research—the American National Election Study (ANES). I examine two outcomes that have been theorized to be mode sensitive: racial attitudes and political knowledge (Appendix A for detailed sample descriptions and question wordings).

Data and Methodology

To use DD, a survey must interview respondents in multiple waves, interviewing some in *different* modes across waves and others in the *same* mode across waves. The 2016-2020 ANES meets these criteria. The 2016-2020 ANES includes 2,670 respondents; 639 were interviewed in-person in 2016 but switched online in 2020, while 2,031 respondents took online surveys in 2016 and 2020. I use DD to estimate the ATT of being surveyed in-person instead of online in 2016.

The outcomes I examine are racial resentment and political knowledge. Racial resentment is a four-item construct ranging from 0 (low resentment) to 1 (high resentment) capturing a mix of anti-Black affect and beliefs that Blacks fail to uphold American values of individualism and hardworkingness. I subset this analysis to non-Hispanic whites. Political knowledge is the sum of correct answers to eight questions: one about US senators' term lengths, two about the partisan majorities in Congress, and five office-recalls for major political figures (e.g., John Roberts).

For racial resentment, I test the hypothesis that white Americans will report higher racial resentment in online surveys than live interviews because anti-Black sentiments can be socially undesirable to express (Abrajano and Alvarez 2019; Krysan 1998). For political knowledge, I test the hypothesis that respondents will get more items correct in online surveys than in-person interviews because online respondents can search for answers (Clifford and Jerit 2014; Graham 2022; Jensen and Thomsen 2014). The target estimands are the ATTs of fielding live interviews vs. online surveys on mean responses. I compare the difference-in-difference-in-means for those switching modes (treatment) to those who stay online (control) between 2016 and 2020. This is equivalent to a two-way fixed effects (TWFE) regression with fixed effects for units and waves.

Results

In Figure 1, I plot estimated means of white racial resentment by wave and 2016 mode (in-person/treatment, online/control). In the treatment group, mean racial resentment was 0.54 in-person in 2016 and 0.52 online in 2020—an insignificant 0.02-point decline ($p=0.086$). Whites in the control group, however, saw a large decline in racial resentment, from 0.59 in 2016 to 0.53 in 2020—a 0.06-point decline ($p<0.001$). The difference-in-differences (i.e., ATT) is -0.04 points ($p=0.001$); i.e., whites who completed the 2016 ANES in-person reported 0.04-points less racial resentment than they would have had they been surveyed online. Notably, the two groups report

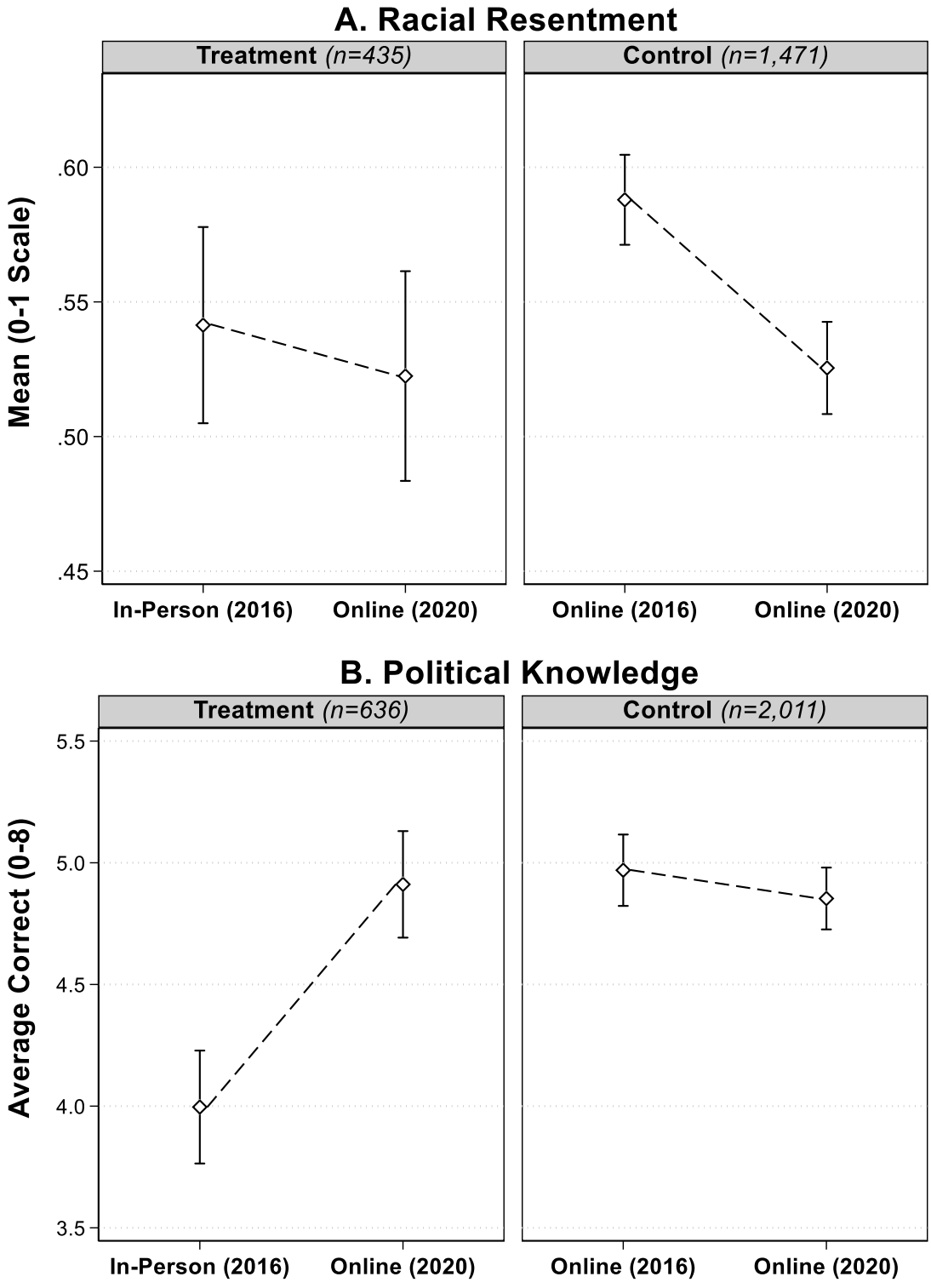

Figure 1—Change in Racial Resentment and Political Knowledge by 2016 Interview Mode. Panel A shows the means of racial resentment by wave (white non-Hispanics). Panel B shows the average of eight political knowledge items answered correctly by wave (full sample). Lines are 95 percent confidence intervals. Data weighted. Source: 2016-2020 ANES.

near-identical racial resentment in the same online mode in 2020, which suggests the difference between the groups in 2016 is largely a product of measurement effects, not sampling effects.

The observed measurement effect for racial resentment has important implications. First, the 2016-2020 ANES understates individual-level decreases in white racial resentment over this period. Thus, the recent decreases in racial resentment noted by Jardina and Ollerenshaw (2022) and others looking at ANES data may understate racial liberalization during this period since the 2020 ANES abandoned in-person interviews for online surveys. Second, these findings bolster conclusions about socially-desirable responding on racial attitude measures which have primarily been tested with cross-sectional designs less well-equipped to simultaneously address omitted variable bias and post-treatment bias (e.g., Abrajano and Alvarez 2019; Krysan 1998; see also Bowyer and Rogowski 2017 for a quasi-experimental design reaching the same conclusion).

Turning to political knowledge, Figure 1 plots the average of eight political knowledge items correctly answered by wave and 2016 mode. Treatment respondents, on average, correctly answered 4.00 items in-person in 2016 and 4.91 items online in 2020—a large 0.91 item increase ($p<0.001$). Control respondents, however, correctly answered an average of 4.97 items in 2016 and 4.85 items in 2020—a 0.12-item *decline* in political knowledge ($p=0.055$). The ATT is -1.03 correct political knowledge items ($p<0.001$). We can conclude from this analysis that on average, respondents who took the 2016 wave in-person got 1.03 fewer political knowledge items correct than they would have had they been surveyed online due to mode-based measurement effects.

Usefully, this analysis can essentially rule out selection effects as an explanation for the differences in political knowledge across 2016 modes (Ansolabehere and Schaffner 2014). If selection had caused differences in political knowledge across modes in 2016, these differences would likely also be seen in 2020. However, the average number of correct items between these

groups are statistically indistinguishable when they both take online surveys in 2020 (p=0.657). Although the degree of measurement effects will differ based on the political knowledge items and the implementation of methods to deter search (Graham 2022), my results support the claim that online respondents appear more politically knowledgeable because at least some of them search for the answers (Clifford and Jerit 2016; Liu and Wang 2014; Shulman and Boster 2014).

Robustness Check #1: Covariate Adjustments

Probing and augmenting the parallel trends assumption (PTA) can bolster the credibility of DD. Unfortunately, many DD robustness checks require more than two waves. Even for 2x2 cases, however, we have some data to help augment the PTA: covariates. Covariate imbalances between treatment and control groups can undermine DD if the covariates are related to outcome trends. Addressing covariate imbalances can therefore bolster the credibility of a DD approach.

There are several ways to make covariate adjustments. Covariates can be used to match treatment respondents to similar control respondents or create weights that correct imbalances across the treatment and control groups (Imai, Kim, and Wang 2021). Importantly, with panel data, covariates can and should be derived *from the same survey wave in the same mode* for all respondents to avoid post-treatment bias when adjusting for potentially mode sensitive variables.

In Table 1, I compare the naïve ATTs to ATTS derived using entropy balancing weights that balance on covariates all measured online in 2020 (Hainmueller 2012; Hainmueller and Xu 2013).[1] The covariates I balance on are age, race/ethnicity, gender, marriage, parent, education, income, urbanicity, religiosity, union membership, unemployment, political interest, ideology,

[1] I entropy balance to covariates' third moments and create separate weights for the white non-Hispanic subsample.

partisanship, and internet access. The ATTs are near-identical, which is unsurprising since the imbalances between the ANES samples by 2016 mode are small (Appendix B). But when there are larger imbalances, covariate adjustments can provide a useful tool for augmenting the PTA.

	Racial Resentment		**Political Knowledge**	
Weight	*Sampling*	*Balancing*	*Sampling*	*Balancing*
ATT of Online Mode	-0.044 (0.013)	-0.050 (0.012)	-1.03 (0.10)	-1.05 (0.10)

Table 1—Measurement Effects by Covariate Adjustments. Entries are ATTs with standard errors in parentheses. Negative estimates indicate lower racial resentment or political knowledge in the in-person mode relative to online mode. Sampling weights drawn from ANES. Balancing weights derived with entropy balancing. Source: 2016-2020 ANES.

4. Heterogeneity-Robust DD: The 1992-2020 Health and Retirement Study

For decades, DD has been performed with two-way fixed effects regression (TWFE). However, recent econometrics work has found TWFE is *not* a valid DD estimator except in the simple 2x2 case (Borusyak et al. 2023; Goodman-Bacon 2021; Imai and Kim 2021). To briefly summarize the issue, TWFE is not robust to heterogeneous treatment effects across groups and time periods; in fact, the estimated treatment effect may be incorrectly signed due to negative weighting (de Chaisemartin and D'Haultfœuille 2020). In this section, I illustrate *heterogeneity-robust* DD designs for identifying measurement effects in cases with more than two panel waves.

Data and Methodology

I examine the 1992-2020 Health and Retirement Study (HRS), a 15-wave, biennial survey by the University of Michigan. Among other uses, the HRS allows researchers to track trends in Americans' cognitive functioning over long periods. I focus on waves 4-15 of the HRS because waves 1-2 used different cognitive functioning measures than the rest of the panel, and a new cohort was added in wave 4. Waves 4-15 of the HRS had 3,843 panelists.

The HRS implemented survey mode changes in waves 14 and 15. In waves 4-13, all HRS panelists were interviewed live, either in-person or over the phone. In waves 14 and 15, however,

the HRS switched some panelists to online modes: 377 panelists shifted online for wave 14, then returned to live interviews in wave 15; 504 panelists completed live interviews in wave 14, but shifted online in wave 15; and 2,962 panelists completed live interviews in both waves 14 and 15. The HRS is ideal for illustrating heterogeneity-robust DD; it has many waves and staggered mode switches that are non-absorbing (i.e., panelists can switch back and forth between modes).

The HRS includes several measures of cognitive functioning; I focus my illustration on immediate word recalls (IWR) and self-rated memory (SRM). To assess IWR in live interviews, interviewers read a list of 10 nouns and ask respondents to report back as many words as they recall. To assess IWR online, the 10 nouns appear one at a time onscreen, and respondents type as many as they recall afterwards. SRM is assessed the same way in both modes on a five-point scale rescaled from 0 (poor) to 1 (excellent).

Research on cognitive functioning and mode effects finds online respondents score better on these measures than live interviewees. Domingue et al. (2023), for example, finds cognitive functioning scores are higher in online than phone surveys using random mode assignments, and Runge et al. (2015) finds IWR is 0.85 words of 10 higher online than in live interviews using a cross-sectional design. Both studies have the standard weaknesses associated with their designs, however; Domingue et al. (2023) note noncompliance and differential selection into the survey based on modes, while Runge et al. (2015) estimates measurement effects for IWR controlling for variables measured in different modes that risk mode sensitivity, such as self-rated memory.

Results

In Figure 2, I plot trends in immediate word recalls (IWR) and self-rated memory (SRM). I plot trends separately for the panelists who switched online in wave 14 (left) vs. 15 (right) and compare both to the panelists that never switched online. We can see a few things from these

trends. First, IWR and SRM decline over time, likely due to an aging effect. Second, in waves 4-13 before any mode switches occur, differences in IWR and SRM between mode switchers and non-switchers can already be seen. Panelists who shift online average better IWR and SRM than those who do not shift online even in live interview waves. This is evidence of selection effects. Third, the groups diverge when mode switches occur in waves 14 or 15 such that panelists who switch online in these waves show increasing IWR and SRM relative to the live interview group.

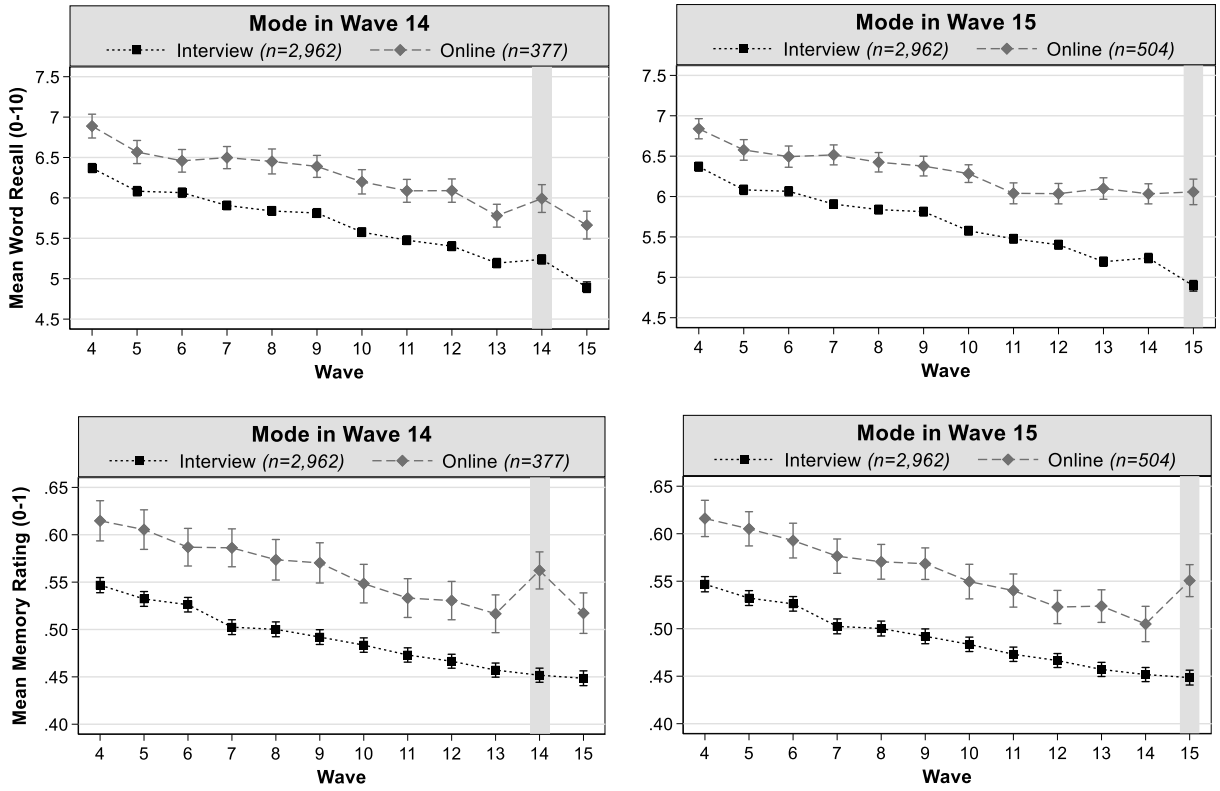

Figure 2—Trends in Immediate Word Recalls and Self-Rated Memory (1998-2020). Points in top panels are average immediate word recalls out of 10. Points in bottom panels are averages of self-rated memory on a 0 (poor) to 1 (excellent) scale. "Interview" groups are surveyed with in-person or phone interviews. "Online" groups are surveyed with live interviews until waves 14 or 15, when they switch online. Biennial waves. Source: Health and Retirement Study.

I next use DD to estimate the measurement effects of switching to the online mode in waves 14 or 15 for the two cognitive functioning measures. I use the heterogeneity-robust *fect* estimator from Liu, Wang, and Xu (LWX) (2022), which is an appropriate estimator for non-

absorbing, staggered treatments. In Figures 3 and 4, I use *fect* to identify the effects of taking the HRS in online vs. live modes on IWR and SRM (Appendix C for tables of results). In Figures 3 and 4, online mode switches occur in period 1; waves -9 to 0 are placebos used to assess whether there are pre-treatment trends indicative of possible PTA violations.

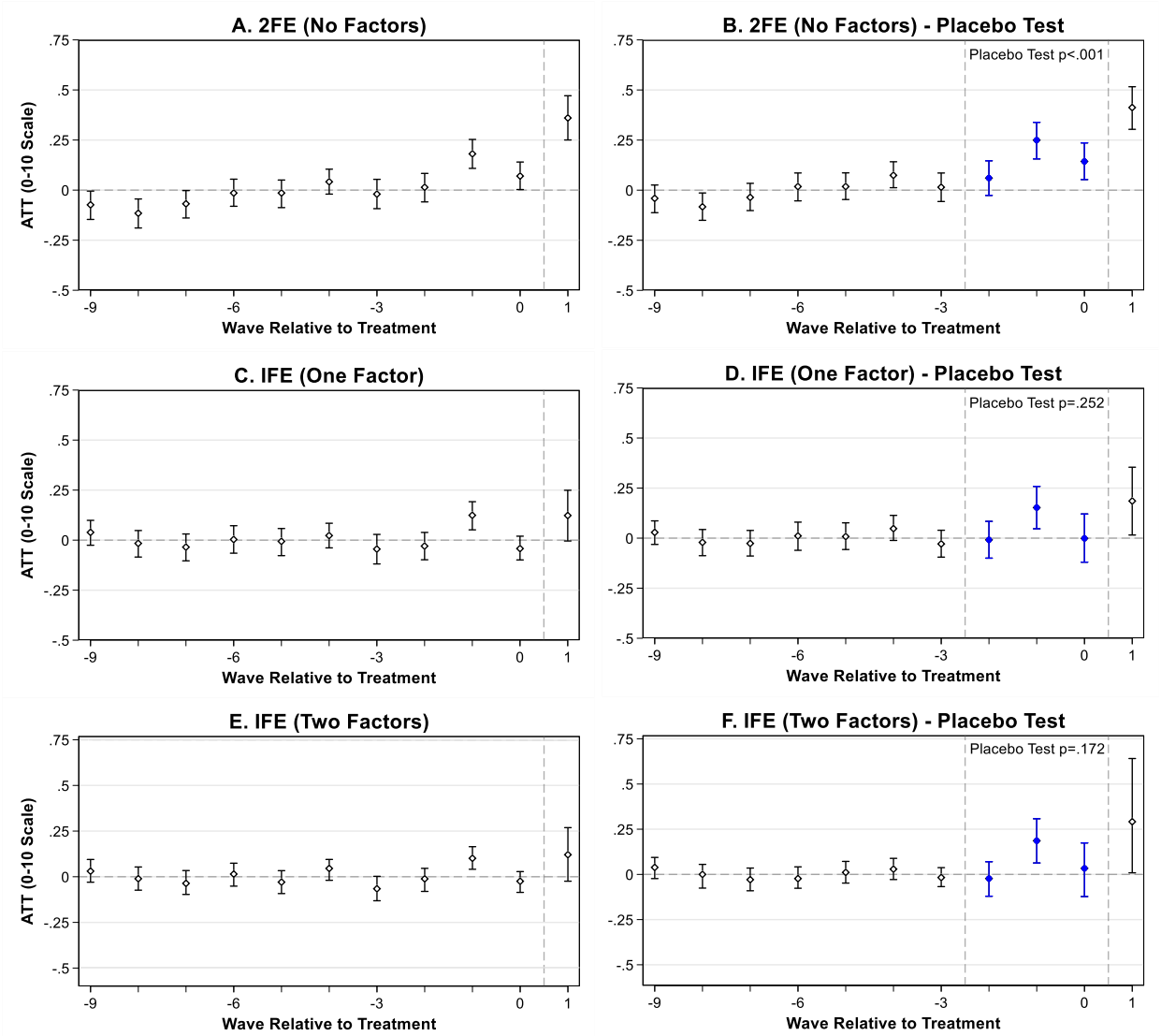

Figure 3—Effect of Switching to Online Mode on Immediate Word Recalls. Points are ATTs of online vs. live mode on immediate word recalls out of ten with 95 percent confidence intervals (bootstrapped standard errors). Panels A, C, and E are two-way fixed effects models with zero, one, and two interactive fixed effects, respectively. Panels B, D, and F are panel placebo tests. Estimator is *fect*. Biennial waves (1998-2020). N=3,843. Source: Health and Retirement Study.

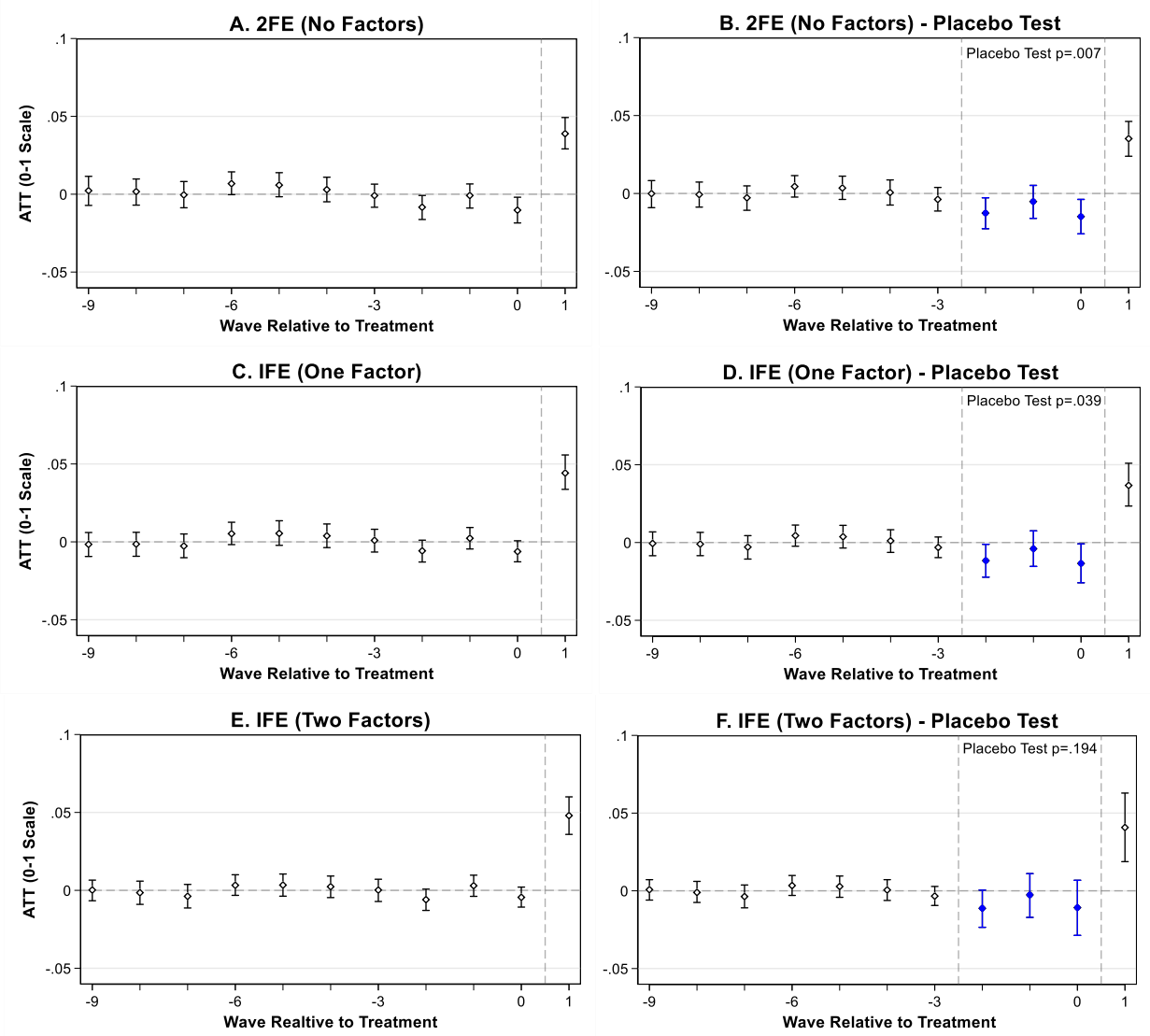

Figure 4—Effect of Switching to Online Mode on Self-Rated Memory. Points are ATTs of online vs. live mode on self-rated memory (0-1 scale) with 95 percent confidence intervals (bootstrapped standard errors). Panels A, C, and E are two-way fixed effects models with zero, one, and two interactive fixed effects, respectively. Panels B, D, and F are panel placebo tests. Estimator is *fect*. Biennial waves (1998-2020). N=3,843. Source: Health and Retirement Study.

In panels A of Figures 3 and 4, I first use the *fect* model similar to heterogeneity-robust TWFE. The estimated effect of completing the HRS online instead of in the live interview mode is a 0.36-word increase out of 10 in IWR ($p<0.001$) and a 0.04-point increase in SRM on a 0-1 scale ($p<0.001$). Thus, respondents do seemingly perform better on these cognitive functioning measures online. These results imply that switching some of its panelists online caused waves 14

and 15 of the HRS to underestimate cognitive declines because online panelists scored better on the immediate word recall task and self-reported better memory, on average.

Robustness Check #2: Identifying and Addressing Parallel Trends Violations

Violations of the parallel trends assumption (PTA) are arguably the foremost challenge to inference with DD (Chiu et al. n.d.; Hassell and Holbein n.d.). Perhaps the best indicator of PTA violations is non-parallel pre-treatment trends. It is therefore best practice to visually inspect pre-trends and use additional statistical placebo tests to marshal evidence for (or against) the PTA.

In panel A of Figure 3, there is clear evidence of PTA violations for the IWR outcome; the pre-trends show the groups are diverging such that IWR is declining faster among untreated respondents. In panel A of Figure 4, we can see likely minor PTA violations for SRM; SRM seems to decline slower for untreated respondents. In addition to visual inspection of the pre-trends, LWX (2022) recommend performing "panel placebo tests." The panel placebo test holds out three waves of observations before treatment onset and uses the remainder of untreated observations to predict the held out outcomes such that the average differences between observed and modeled outcomes should be about zero. The panel placebo tests associated with the TWFE models are shown in panels B; both find evidence of PTA violations ($p_{IWR}<0.001$, $p_{SRM}=0.007$) such that the naively estimated treatment effects are likely biased up for IWR but down for SRM.

One common approach for addressing PTA violations is to use interactive fixed effects (IFE) between units and time to account for pre-trends. Usefully, *fect* easily implements IFE. In Figures 3 and 4, I re-estimate the measurement effects of online vs. live mode using one factor IFE (panel C) and two factor IFE (panel E). I offer the associated panel placebo tests in panels D and F. Visually, we see from Figures 3 and 4 that the IFE models level off the pre-trends such that the pre-trends bounce around zero without any clear direction over time. The panel placebo

tests further suggest IFE addresses PTA violations; adding one factor gives a null panel placebo test for IWR (p=0.252) and a narrowly significant test for SRM (p=0.039). Visually and using statistical placebo tests, we see that adding a second factor makes little difference over the first; although the SRM panel placebo test becomes null with two factors (p=0.194), this has more to do with decreasing precision than meaningful changes to the placebo coefficients.

In Figure 3, the measurement effect on IWR is 0.36-words with zero factors (p<0.001), 0.12-words with one factor IFE (p=0.054), and 0.12-words with two factor IFE (p=0.114). The inclusion of IFEs attenuates the estimated measurement effect for IWR because mode switchers were declining less in IWR over time than non-switchers, i.e., a likely PTA violation. In Figure 4, the measurement effect on SRM is 0.039-points with zero factors (p<0.001), 0.044-points with one factor IFE (p<0.001), and 0.048-points with two factor IFE (p<0.001). Here, including IFEs increases the estimated measurement effect because the mode switchers were declining faster in SRM than non-switchers. Both cases illustrate how PTA violations can bias DD estimates. Here, PTA violations lead naïve TWFE models to overstate the measurement effects of online vs. live modes on immediate word recalls but understate the measurement effects on self-rated memory.

Robustness Check #3: Alternative Heterogeneity-Robust Estimators

Econometricians have not settled on a "standard" heterogeneity-robust estimator in lieu of TWFE regression. Several other estimators would have been as appropriate as LWX (2022), each differing primarily in the control units and/or periods used as comparisons. In Table 2, I compare estimates from one factor IFE with LWX (2022) to those from two other heterogeneity-robust estimators that accommodate staggered, non-absorbing treatments and IFE: Borusyak, Jaravel, and Spiess (BJS) (2023) and de Chaisemartin and D'Haultfœuille (CD) (2020) (see Chiu et al. n.d. for a review of these and other estimators that could be used for different mixed-mode

panel structures; e.g., Imai, Kim, and Wang 2021; Sun and Abraham 2021). I also provide results from TWFE regressions with linear IFEs, acknowledging that this is still a widely-used approach for causal inference with panel data.

I find similar effects across estimators. The effects of online mode on IWR range from one-tenth to just under three-tenths word increased recalls out of 10—all substantively small, but sometimes significant, effects. The effects of online mode on SRM, however, range from 0.04- to 0.05-points on a 0-1 scale—these are large effects equivalent to reversing a decade's worth of declines in SRM. The substantive takeaway for those interested in tracking cognitive functioning is that some cognitive functioning measures are highly sensitive to their mode of administration.

Immediate Word Recalls (0-10 Scale)			
LWX (2022) *fect*	BJS (2023) *did_imputation*	CD (2020) *did_multiplegt*	TWFE Regression w/Linear IFE
0.12 (0.06)	0.23 (0.06)	0.29 (0.06)	0.10 (0.05)
Self-Rated Memory (0-1 Scale)			
LWX (2022) *fect*	BJS (2023) *did_imputation*	CD (2020) *did_multiplegt*	TWFE Regression w/Linear IFE
0.044 (0.006)	0.046 (0.006)	0.051 (0.006)	0.044 (0.006)

Table 2—ATTs on Immediate Word Recalls and Self-Rated Memory by Estimator. Entries are ATTs by estimator with standard errors in parentheses. Estimators use linear interactive fixed effects. Positive estimates indicate increased word recalls or better self-rated memory online vs. live. N=3,843. Source: Health and Retirement Study.

5. Conclusion

In this paper, I proposed using difference-in-differences (DD) to identify mode effects. DD can be employed with panel surveys when some respondents switch modes across waves and others stay in the same mode. DD assumes that outcomes for panelists who did vs. did not switch modes would have moved in parallel absent mode switches. The parallel trends assumption can be gauged and augmented but is not directly testable. However, parallel trends is a much weaker identifying assumption than those for cross-sectional designs, which face a catch-22 of assuming

no omitted variable bias and that variables used to address omitted variable bias measured in different modes are mode insensitive. DD is a credible alternative to cross-sectional designs, and it is often more feasible than lab experiments which are, in practice, rare due to the difficulty of ensuring no differential selection after randomizing mode assignments.

There remains substantial work to do applying DD to understand mode effects. I examine several applications to political science and health research, but future work should apply DD to identify mode effects for other outcomes, especially those that have currently only been studied cross-sectionally. Further, I defined differences-in-means as the target estimands, but mode can affect other response patterns. Homola et al. (2016), for example, find online respondents have more dispersed responses than in-person interviewees, whereas Bowyer and Rogowski (2017) reach the opposite conclusion comparing online and phone respondents. Extensions focusing on mode's distributional consequences (e.g., quantile effects; Callaway and Li 2019) could help to adjudicate these competing claims.

Finally, difference-in-differences has immediate implications and applications for panel survey design. Measurement effects can confound estimates of individual-level change in panel surveys. To identify and account for measurement effects, panel surveys should consider phasing in new modes such that some panelists continue taking the survey in their original modes. Such designs would allow researchers to quantify the bias introduced by mode switches, increasing the validity of inferences derived from panelists switching modes and bolstering the credibility of surveys as tools for population-level inference amidst a changing landscape for survey research.

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.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2): 3–30.

Angrist, Joshua David, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Ansolabehere, Stephen, and Brian F. Schaffner. 2014. "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison." *Political Analysis* 22(3): 285–303.

Baker, Andrew C., David F. Larcker, and Charles C. Y. Wang. 2022. "How Much Should We Trust Staggered Difference-in-Differences Estimates?" *Journal of Financial Economics* 144(2): 370–95.

Bethlehem, Jelke. 2010. "Selection Bias in Web Surveys." *International Statistical Review* 78(2): 161–88.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2023. "Revisiting Event Study Designs: Robust and Efficient Estimation." http://arxiv.org/abs/2108.12419 (May 10, 2023).

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.

Callaway, Brantly, and Tong Li. 2019. "Quantile Treatment Effects in Difference in Differences Models with Panel Data." *Quantitative Economics* 10(4): 1579–1618.

Cernat, Alexandru, and Joseph W. Sakshaug. 2021. “Estimating the Measurement Effects of Mixed Modes in Longitudinal Studies: Current Practice and Issues.” In *Advances in Longitudinal Survey Methodology*, Wiley Series in Probability and Statistics, ed. Peter Lynn. Hoboken, NJ: Wiley, 227–49.

de Chaisemartin, Clément, and Xavier D’Haultfœuille. 2020. “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects.” *American Economic Review* 110(9): 2964–96.

Chang, Linchiat, and Jon A. Krosnick. 2010. “Comparing Oral Interviewing with Self-Administered Computerized QuestionnairesAn Experiment.” *Public Opinion Quarterly* 74(1): 154–67.

Chiu, Albert, Xingchen Lan, Ziyi Liu, and Yiqing Xu. “What To Do (and Not to Do) with Causal Panel Analysis under Parallel Trends: Lessons from A Large Reanalysis Study.”

Clifford, Scott, and Jennifer Jerit. 2014. “Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies.” *Journal of Experimental Political Science* 1(2): 120–31.

———. 2016. “Cheating on Political Knowledge Questions in Online Surveys: An Assessment of the Problem and Solutions.” *Public Opinion Quarterly* 80(4): 858–87.

Domingue, Benjamin W et al. 2023. “The Mode Effect of Web-Based Surveying on the 2018 U.S. Health and Retirement Study Measure of Cognitive Functioning.” *The Journals of Gerontology: Series B*: gbad068.

Druckman, James N, and Cindy D. Kam. 2011. “Students as Experimental Participants.” In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman,

Donald P. Greene, James H. Kuklinski, and Arthur Lupia. Cambridge University Press, 41–57.

Endres, Kyle, D. Sunshine Hillygus, Matthew DeBell, and Shanto Iyengar. 2022. "A Randomized Experiment Evaluating Survey Mode Effects for Video Interviewing." *Political Science Research and Methods*: 1–16.

Gooch, Andrew, and Lynn Vavreck. 2019. "How Face-to-Face Interviews and Cognitive Skill Affect Item Non-Response: A Randomized Experiment Assigning Mode of Interview." *Political Science Research and Methods* 7(1): 143–62.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225(2): 254–77.

Graham, Matthew H. 2022. "Detecting and Deterring Information Search in Online Surveys." *American Journal of Political Science*.

Greenacre, Zerrin Asan. 2016. "The Importance of Selection Bias in Internet Surveys." *Open Journal of Statistics* 06(03): 397–404.

Groves, Robert M. et al. 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.

Hainmueller, Jens, and Yiqing Xu. 2013. *Ebalance: A Stata Package for Entropy Balancing*. Rochester, NY: Social Science Research Network. SSRN Scholarly Paper. https://papers.ssrn.com/abstract=1943090 (November 30, 2021).

Hassell, Hans J.G., and John Holbein. “Navigating Potential Pitfalls in Dierence-in-Dierences Designs: Reconciling Conflicting Findings on Mass Shootings’ Effect on Electoral Outcomes.”

Hernán, Miguel A., Sonia Hernández-Díaz, and James M. Robins. 2013. “Randomized Trials Analyzed as Observational Studies.” *Annals of internal medicine* 159(8): 10.7326/0003-4819-159-8-201310150–00709.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” *Political Analysis* 15(3): 199–236.

Homola, Jonathan, Natalie Jackson, and Jeff Gill. 2016. “A Measure of Survey Mode Differences.” *Electoral Studies* 44: 255–74.

Imai, Kosuke, and In Song Kim. 2021. “On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data.” *Political Analysis* 29(3): 405–15.

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*.

Jardina, Ashley E., and Trent Ollerenshaw. 2022. “The Polls-Trends: The Polarization of White Racial Attitudes and Support for Racial Equality in The U.S.” *Public Opinion Quarterly* 86(S1): 576–87.

Jensen, Carsten, and Jens Peter Frølund Thomsen. 2014. “Self-Reported Cheating in Web Surveys on Political Knowledge.” *Quality & Quantity* 48(6): 3343–54.

Kahn-Lang, Ariella, and Kevin Lang. 2020. “The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications.” *Journal of Business & Economic Statistics* 38(3): 613–20.

Keeter, Scott, and Kyley Mcgeeney. 2015. "Coverage Error in Internet Surveys." *Pew Research Center Methods*. https://www.pewresearch.org/methods/2015/09/22/coverage-error-in-internet-surveys/ (September 8, 2022).

Krueger, Brian S., and Brady T. West. 2014. "Assessing the Potential of Paradata and Other Auxiliary Data for Nonresponse Adjustments." *Public Opinion Quarterly* 78(4): 795–831.

Krupnikov, Yanna, and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1(1): 59–80.

Krysan, Maria. 1998. "Privacy and the Expression of White Racial Attitudes: A Comparison Across Three Contexts." *Public Opinion Quarterly* 62(4): 506.

de Leeuw, Edith D. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics* 21(2): 233–55.

Liu, Licheng, Ye Wang, and Yiqing Xu. 2022. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science*.

Liu, Mingnan, and Yichen Wang. 2014. "Data Collection Mode Effects On Political Knowledge." *Survey Methods: Insights from the Field (SMIF)*.

Lugtig, Peter, Gerty J.L.M. Lensvelt-Mulders, Remco Frerichs, and Assyn Greven. 2011. "Estimating Nonresponse Bias and Mode Effects in a Mixed-Mode Survey." *International Journal of Market Research* 53(5): 669–86.

Malani, Anup, and Julian Reif. 2015. "Interpreting Pre-Trends as Anticipation: Impact on Estimated Treatment Effects from Tort Reform." *Journal of Public Economics* 124: 1–17.

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

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. “How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It.” *American Journal of Political Science* 62(3): 760–75.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. “The Generalizability of Survey Experiments*.” *Journal of Experimental Political Science* 2(2): 109–38.

Mutz, Diana Carole. 2011. *Population-Based Survey Experiments*. Princeton University Press.

Olson, Kristen et al. 2021. “Transitions from Telephone Surveys to Self-Administered and Mixed-Mode Surveys: AAPOR Task Force Report.” *Journal of Survey Statistics and Methodology* 9(3): 381–411.

Roth, Jonathan, Pedro H. C. Sant’Anna, Alyssa Bilinski, and John Poe. 2023. “What’s Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature.” *Journal of Econometrics*.

Rubin, Donald B. 2005. “Causal Inference Using Potential Outcomes: Design, Modeling, Decisions.” *Journal of the American Statistical Association* 100(469): 322–31.

Runge, Shannon K, Benjamin M Craig, and Heather S Jim. 2015. “Word Recall: Cognitive Performance Within Internet Surveys.” *JMIR Mental Health* 2(2): e20.

Schonlau, Matthias, Arthur van Soest, Arie Kapteyn, and Mick Couper. 2009. “Selection Bias in Web Surveys and the Use of Propensity Scores.” *Sociological Methods & Research* 37(3): 291–318.

Sears, David O. 1986. “College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology’s View of Human Nature.” *Journal of Personality and Social Psychology* 51: 515–30.

Shulman, Hillary C., and Franklin J. Boster. 2014. “Effect of Test-Taking Venue and Response Format on Political Knowledge Tests.” *Communication Methods and Measures* 8(3): 177–89.

Sun, Liyang, and Sarah Abraham. 2021. “Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects.” *Journal of Econometrics* 225(2): 175–99.

Valentino, Nicholas A., Kirill Zhirkov, D. Sunshine Hillygus, and Brian Guay. 2020. “The Consequences of Personality Biases in Online Panels for Measuring Public Opinion.” *Public Opinion Quarterly* 84(2): 446–68.

Vannieuwenhuyze, Jorre T. A., and Geert Loosveldt. 2013. “Evaluating Relative Mode Effects in Mixed-Mode Surveys:: Three Methods to Disentangle Selection and Measurement Effects.” *Sociological Methods & Research* 42(1): 82–104.

Voogt, Robert J J, and Willem E Saris. 2005. “Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects.” *Journal of Official Statistics*: 21.

Appendix Materials for "A Difference-in-Differences Approach for Estimating Survey Mode Effects"

Trent Ollerenshaw
Ph.D. Candidate
Duke University
Department of Political Science
trent.ollerenshaw@duke.edu

Contents

Appendix A: Data Descriptions and Question Wordings

2016-2020 American National Election Study (ANES) Panel

Sample Size: 2,670 (full sample), 1,921 (non-Hispanic white subsample).

Field Dates: The 2016 pre-election was fielded September 7-November 7, 2016, the post-election was fielded November 9, 2016-January 8, 2017. The 2020 pre-election was fielded August 18-November 3, 2020, the post-election was fielded November 8, 2020-January 4, 2021.

Sample Recruitment: Data collection was performed by Westat, Inc. "2016 ANES respondents were invited by email where possible, with letters used if there was no email on file or after an initial non-response…All respondents who completed the post-election survey did so in the same mode used for the pre-election survey" (pg. 4). Respondents who completed the 2016 ANES were invited via email or mail to complete the 2020 ANES.

Response Rate and Panel Attrition: The response rate (AAPOR RR1) in the 2016 ANES pre-election wave was 50 percent for the face-to-face sample and 44 percent for the internet sample. Of those who completed the 2016 pre-election wave, 90 percent of the face-to-face sample and 84 percent of the internet sample completed the 2016 post-election wave. The reinterview rate for the 2020 pre-election wave was 77.9 percent. Of those who completed the 2020 pre-election wave, 94.0 percent completed the 2020 post-election wave. Overall, retention was 73.2 percent.

Weights and Sample Design Effects: The 2016-2020 ANES Panel is a probability-based sample collected with a complex sampling design. To generalize to the target population, the ANES recommends using weight variable V200011b for the 2016-2020 sample that completed the post-election 2020 wave. The strata and cluster variables are V200011d and V200011c, respectively.

Question Wordings:
Racial Resentment: A four-item scale recoded to range from 0 to 1. First and fourth items reverse coded. Each item prompts: "Do you agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly with this statement?" Five-point response scale: Agree strongly, Agree somewhat, Neither agree nor disagree, Disagree somewhat, Disagree strongly.

1. 'Irish, Italian, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors.' (Reverse Coded)
2. 'Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class.'
3. 'Over the past few years, blacks have gotten less than they deserve.'
4. 'It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites.' (Reverse Coded)

Political Knowledge: An eight-item summative scale for the number of political knowledge items correctly answered. Don't Know/No Response answers are also coded as incorrect responses.

1. "For how many years is a United States Senator elected - that is, how many years are there in one full term of office for a U.S. Senator?" [6 years]
2. "Do you happen to know which party currently has the most members in the U.S. House of Representatives in Washington?" [The Democratic Party]
3. "Do you happen to know which party currently has the most members in the U.S. Senate?" [The Republican Party]
4. "What job or political office does Mike Pence now hold?" [Vice President]
5. "What job or political office does Nancy Pelosi now hold?" [Speaker of the House]
6. "What job or political office does Angela Merkel now hold?" [Chancellor of Germany]
7. "What job or political office does Vladimir Putin now hold?" [President of Russia]
8. "What job or political office does John Roberts now hold?" [Chief Justice of SCOTUS]

References

American National Election Studies. 2019. ANES 2016 Time Series Study Full Release [dataset and documentation]. September 4, 2019 version. https://electionstudies.org/data-center/2016-time-series-study/

American National Election Studies. 2021. ANES 2020 Time Series Study Full Release [dataset and documentation]. July 19, 2021 version. https://electionstudies.org/data-center/2020-time-series-study/

Health and Retirement Survey (HRS) by National Institute on Aging

Sample Size: 3,843.

Field Dates: The HRS Cognitive Functioning Measures survey is fielded biennially and spans 1992-2020. However, the first two waves use different word sets for the immediate word recall task than the later waves, so I exclude these waves. Additionally, a new cohort was introduced in wave 4, so I opt to subset to panelists who (1) completed wave 4 in 1998 and (2) remained in the panel until wave 15 in 2020.

Sample Recruitment: Data collection was conducted by the University of Michigan at Ann Arbor through the Survey Research Center.

Response Rate and Panel Attrition: The response rate for the new cohort in 1992 was 0.81, and the response rate for the 1998 cohort was 0.72. Retention rates for each panel wave are found on the HRS website: https://hrs.isr.umich.edu/documentation/survey-design/response-rates.

Imputation of Missing Values: The "Imputation of Cognitive Functioning Measures: 1992 – 2020" data file imputes non-response using a "multivariate, regression-based procedure" which "used a combination of relevant demographic, health, and economic variables, as well as prior and current wave cognitive variables to perform the imputations." Imputations on the word recall task are usually about 1% of the sample, and do not exceed 2% of the sample across waves 4-15.

Immediate Word Recall Lists: The immediate word recall randomizes four lists across waves:

1. Hotel, River, Tree, Skin, Gold, Market, Paper, Child, King, Book
2. Sky, Ocean, Flag, Dollar, Wife, Machine, Home, Earth, College, Butter
3. Woman, Rock, Blood, Corner, Shoes, Letter, Girl, House, Valley, Engine
4. Water, Church, Doctor, Palace, Fire, Garden, Sea, Village, Baby, Table

Self-Rated Memory: "Part of this study is concerned with people's memory, and ability to think about things. First, how would you rate your memory at the present time? Would you say it is excellent, very good, good, fair or poor?" [Response Set: Excellent, Very good, Good, Fair, Poor]

References

Health and Retirement Study, HRS Imputation of Cognitive Functioning Measures: 1992 – 2020 public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI. Accessed June 2023: https://hrs.isr.umich.edu/publications/biblio/5760

Appendix B: ANES Covariate Distributions Pre/Post Entropy Balancing

Table B1—Pre-Balancing Covariate Distributions by 2016 Mode (Full Sample)

	In-Person (2016)			Online (2016)		
Covariate	*Mean*	*Variance*	*Skewness*	*Mean*	*Variance*	*Skewness*
Black	0.08	0.08	3.00	0.11	0.09	2.57
Hispanic	0.11	0.10	2.49	0.11	0.10	2.45
White	0.67	0.22	-0.72	0.67	0.22	-0.71
Male	0.47	0.25	0.11	0.49	0.25	0.05
Education	0.56	0.08	-0.10	0.52	0.09	0.12
Age	0.52	0.08	-0.04	0.51	0.07	0.01
Married	0.56	0.25	-0.24	0.55	0.25	-0.18
Parent	0.37	0.23	0.55	0.34	0.22	0.69
Union	0.15	0.13	1.99	0.15	0.13	1.91
Unemployed	0.01	0.01	10.58	0.02	0.02	6.84
Retired	0.17	0.14	1.71	0.18	0.14	1.71
Disabled	0.05	0.05	4.08	0.06	0.05	3.78
Income	0.51	0.10	-0.10	0.51	0.10	-0.09
Democrat	0.44	0.25	0.25	0.46	0.25	0.15
Republican	0.40	0.24	0.40	0.41	0.24	0.36
Liberal	0.27	0.20	1.04	0.27	0.20	1.02
Conservative	0.32	0.22	0.78	0.34	0.23	0.66
Political Interest	0.65	0.06	-0.45	0.63	0.06	-0.39
Religiosity	0.25	0.10	1.01	0.25	0.10	1.02
Internet User	0.98	0.02	-6.18	0.97	0.03	-5.12
Internet Access	0.84	0.06	-1.60	0.84	0.07	-1.74
Rural	0.16	0.13	1.86	0.16	0.14	1.82
Small Town	0.30	0.21	0.86	0.24	0.18	1.24
Suburban	0.29	0.20	0.95	0.29	0.21	0.92

Note: Data are weighted with ANES panel sampling weight. Source: 2016-2020 ANES.

Table B2—Post-Balancing Covariate Distributions by 2016 Mode (Full Sample)

	In-Person (2016)			Online (2016)		
Covariate	*Mean*	*Variance*	*Skewness*	*Mean*	*Variance*	*Skewness*
Black	0.08	0.08	3.00	0.08	0.08	3.00
Hispanic	0.11	0.10	2.49	0.11	0.10	2.49
White	0.67	0.22	-0.72	0.67	0.22	-0.72
Male	0.47	0.25	0.11	0.47	0.25	0.11
Education	0.56	0.08	-0.10	0.56	0.08	-0.10
Age	0.52	0.08	-0.04	0.52	0.08	-0.04
Married	0.56	0.25	-0.24	0.56	0.25	-0.24
Parent	0.37	0.23	0.55	0.37	0.23	0.54
Union	0.15	0.13	1.99	0.15	0.13	1.98
Unemployed	0.01	0.01	10.58	0.01	0.01	10.58
Retired	0.17	0.14	1.71	0.17	0.14	1.71
Disabled	0.05	0.05	4.08	0.05	0.05	4.08
Income	0.51	0.10	-0.10	0.51	0.10	-0.10
Democrat	0.44	0.25	0.25	0.44	0.25	0.25
Republican	0.40	0.24	0.40	0.40	0.24	0.40
Liberal	0.27	0.20	1.04	0.27	0.20	1.03
Conservative	0.32	0.22	0.78	0.32	0.22	0.78
Political Interest	0.65	0.06	-0.45	0.65	0.06	-0.45
Religiosity	0.25	0.10	1.01	0.25	0.10	1.01
Internet User	0.98	0.02	-6.18	0.98	0.02	-6.17
Internet Access	0.84	0.06	-1.60	0.84	0.06	-1.60
Rural	0.16	0.13	1.86	0.16	0.13	1.86
Small Town	0.30	0.21	0.86	0.30	0.21	0.86
Suburban	0.29	0.20	0.95	0.29	0.20	0.95

Note: Data are weighted with entropy balancing (Hainmueller 2012). Source: 2016-2020 ANES.

Table B3—Pre-Balancing Covariate Distributions by 2016 Mode (Non-Hispanic Whites)

	In-Person (2016)			Online (2016)		
Covariate	*Mean*	*Variance*	*Skewness*	*Mean*	*Variance*	*Skewness*
Male	0.46	0.25	0.15	0.49	0.25	0.04
Education	0.60	0.07	-0.07	0.54	0.08	0.12
Age	0.55	0.08	-0.17	0.54	0.07	-0.17
Married	0.60	0.24	-0.42	0.61	0.24	-0.43
Parent	0.32	0.22	0.76	0.30	0.21	0.87
Union	0.15	0.13	1.94	0.15	0.13	1.97
Unemployed	0.01	0.01	12.34	0.01	0.01	8.91
Retired	0.20	0.16	1.50	0.21	0.16	1.45
Disabled	0.05	0.05	4.24	0.05	0.05	4.27
Income	0.55	0.11	-0.29	0.54	0.10	-0.22
Democrat	0.39	0.24	0.45	0.38	0.24	0.47
Republican	0.47	0.25	0.14	0.51	0.25	-0.05
Liberal	0.29	0.20	0.95	0.27	0.20	1.06
Conservative	0.37	0.23	0.55	0.41	0.24	0.39
Political Interest	0.67	0.06	-0.51	0.64	0.06	-0.50
Religiosity	0.26	0.10	0.90	0.23	0.10	1.09
Internet User	0.98	0.02	-6.97	0.97	0.03	-5.28
Internet Access	0.86	0.06	-1.71	0.87	0.06	-2.01
Rural	0.20	0.16	1.53	0.19	0.16	1.55
Small Town	0.30	0.21	0.87	0.26	0.19	1.08
Suburban	0.28	0.20	0.99	0.30	0.21	0.87

Note: Data are weighted with ANES panel sampling weight. Source: 2016-2020 ANES.

Table B4—Post-Balancing Covariate Distributions by 2016 Mode (Non-Hispanic Whites)

	In-Person (2016)			Online (2016)		
Covariate	*Mean*	*Variance*	*Skewness*	*Mean*	*Variance*	*Skewness*
Male	0.46	0.25	0.15	0.46	0.25	0.15
Education	0.60	0.07	-0.07	0.60	0.07	-0.07
Age	0.55	0.08	-0.17	0.55	0.08	-0.17
Married	0.60	0.24	-0.42	0.60	0.24	-0.42
Parent	0.32	0.22	0.76	0.32	0.22	0.75
Union	0.15	0.13	1.94	0.15	0.13	1.94
Unemployed	0.01	0.01	12.34	0.01	0.01	12.31
Retired	0.20	0.16	1.50	0.20	0.16	1.49
Disabled	0.05	0.05	4.24	0.05	0.05	4.24
Income	0.55	0.11	-0.29	0.55	0.11	-0.28
Democrat	0.39	0.24	0.45	0.39	0.24	0.45
Republican	0.47	0.25	0.14	0.47	0.25	0.14
Liberal	0.29	0.20	0.95	0.29	0.20	0.95
Conservative	0.37	0.23	0.55	0.37	0.23	0.55
Political Interest	0.67	0.06	-0.51	0.67	0.06	-0.51
Religiosity	0.26	0.10	0.90	0.26	0.10	0.90
Internet User	0.98	0.02	-6.97	0.98	0.02	-6.95
Internet Access	0.86	0.06	-1.71	0.86	0.06	-1.70
Rural	0.20	0.16	1.53	0.20	0.16	1.53
Small Town	0.30	0.21	0.87	0.30	0.21	0.87
Suburban	0.28	0.20	0.99	0.28	0.20	0.99

Note: Data are weighted with entropy balancing (Hainmueller 2012). Source: 2016-2020 ANES.

Appendix C: Table Results for HRS Analyses

Immediate Word Recalls

Table C1—ATTs of Online vs. Live Mode on Immediate Word Recalls (2FE – No Factors)

Wave Relative to Treatment										
-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1
-0.07	-0.12	-0.07	-0.01	-0.01	0.04	-0.02	0.01	0.18	0.07	0.36
(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)	(0.04)	(0.05)

Table C2—ATTs of Online vs. Live Mode on Immediate Word Recalls (IFE – One Factor)

Wave Relative to Treatment										
-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1
0.04	-0.02	-0.03	0.00	-0.01	0.02	-0.04	-0.03	0.12	-0.04	0.12
(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.03)	(0.06)

Table C3—ATTs of Online vs. Live Mode on Immediate Word Recalls (IFE – Two Factors)

Wave Relative to Treatment										
-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1
0.03	-0.01	-0.04	0.01	-0.03	0.05	-0.07	-0.01	0.10	-0.02	0.12
(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.08)

Self-Rated Memory

Table C4—ATTs of Online vs. Live Mode on Self-Rated Memory (2FE – No Factors)

Wave Relative to Treatment										
-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1
0.002	0.002	-0.000	0.007	0.006	0.003	-0.001	-0.008	-0.001	-0.010	0.039
(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)

Table C5—ATTs of Online vs. Live Mode on Self-Rated Memory (IFE – One Factor)

Wave Relative to Treatment										
-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1
-0.002	-0.001	-0.003	0.005	0.005	0.004	0.001	-0.006	0.002	-0.006	0.044
(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.006)

Table C6—ATTs of Online vs. Live Mode on Self-Rated Memory (IFE – Two Factors)

Wave Relative to Treatment										
-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1
0.000	-0.002	-0.004	0.003	0.003	0.002	0.000	-0.006	0.003	-0.004	0.048
(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.006)