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: June 24th, 2023
Word Count: 7,237

Abstract: As surveys increasingly rely on new modes, it is important that researchers understand how mode influences survey responses. Two common designs for identifying mode effects are cross-sectional approaches and experiments. But cross-sectional designs risk a combination of omitted variable bias and post-treatment bias when conditioned on respondent characteristics that are themselves mode sensitive. In theory, experiments can obviate these biases, but only when the experiment occurs in tightly-controlled settings that avoid differential uptake across modes. Considering the costliness and paucity of such experiments, in this paper, I propose a difference-in-differences approach for estimating mode effects. Leveraging mixed-mode panel surveys, mode effects can be identified by comparing changes in responses for panelists who switch modes across waves to those who remain in the same modes. Difference-in-differences offers a cost-free alternative to experiments and potentially large bias reduction gains vis-à-vis widely-utilized cross-sectional designs. I apply the difference-in-differences approach by estimating the effects of completing live interviews vs. web surveys on racial attitudes and political knowledge in the 2016-2020 ANES and on cognitive functioning measures in the 1992-2020 Health and Retirement Study.

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

The mode a survey is fielded in can affect 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, the advent of self-administered web surveys, and a precipitous decline in response rates have caused researchers to turn to new modes to maintain the viability of survey research. Indeed, many major surveys for research on health, economics, culture, and politics have recently adopted new online modes, including the General Social Survey, the U.S. Census Bureau's American Community Survey, and the American National Election Study. The consequences of changing modes, and especially the effect of shifting away from live interviews towards self-administered online surveys, must be well-understood to sustain inferences drawn in this new landscape for survey research.

In practice, identifying mode effects poses a challenge because mode can affect both selection into the survey and participants' responses in that survey (i.e., measurement). To decompose measurement and selection effects, studies have generally employed either cross-sectional designs that compare responses across modes while adjusting for covariate imbalances or experiments that randomize modes. Unfortunately, both approaches have major shortcomings. To avoid bias, cross-sectional designs must account for all confounds associated with selection without conditioning on mode sensitive (i.e., post-treatment) variables. This is often a catch-22 because many variables related to selection are themselves mode sensitive, so researchers must weigh between allowing omitted variable bias or introducing post-treatment bias. Experiments suffer similar biases when modes are assigned before respondents' participation is assured due to the potential for differential uptake by assigned modes, and experiments that ensure respondents' participation after assigning modes are rare since they require controlled lab settings, which can be cost-prohibitive (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 approaches in terms of bias reduction and provides a feasible, cost-free 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. In Section 3, I apply DD to the simplest case (two waves, two modes) by estimating the effects of in-person vs. online surveys on racial attitudes and political knowledge in the 2016-2020 ANES. 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 modes in waves 14 and 15. In Section 6, I discuss the benefits and limitations of DD, as well as possible 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 developments, understanding how mode affects survey responses is critical.

Mode Effects: Selection and Measurement

Mode is primarily thought to affect survey outcomes in two ways, which can be broadly characterized as "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 a web survey's sample frame if they lack internet access (i.e., coverage error; Keeter and Mcgeeney 2015), whereas others may be less likely to sit down for a live interview than an online survey (i.e., unit non-response; Bethlehem 2010). If characteristics associated with respondents' probability of appearing in different modes are also associated with their responses, surveys fielded in different modes can produce different population-level inferences.

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 (Bowyer and Rogowski 2017; McClendon 1991). Selection and measurement effects can cause otherwise identical surveys using different modes to generate different population-level inferences.

As surveys increasingly utilize mixed-modes or fully transition online, identifying mode effects becomes important for sustaining and 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 since these effects are observationally equivalent. Moreover, the solutions to selection effects (e.g., weighting) can be inappropriate for addressing measurement

effects, so separating out measurement effects is often necessary. Towards this goal, researchers have primarily used two approaches: cross-sectional and experimental designs.

Cross-Sectional Designs

By far the most common approach for estimating measurement effects is the cross-sectional design. In the standard cross-sectional design, researchers use mixed-mode surveys to compare differences in responses across modes while modeling the selection process. Generally, selection is modeled with regression; researchers control for confounds thought to be associated with selection and the outcome (but the logic of cross-sectional designs extends to matching and balancing designs; see Greenacre 2016; Lugtig et al. 2011; Schonlau et al. 2009). If the selection process can be accurately modeled, researchers can estimate measurement effects using mixed-mode, cross-sectional data.

Though cross-sectional designs are frequently employed because mixed-mode surveys are widely available, these designs rely on arguably untenable assumptions. First, all confounds must be accounted for. In practice, this is difficult because many factors influence respondents' probability of taking a survey, and these factors are usually infrequently or imperfectly 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, in turn, used to model selection despite personality's great importance in many areas of social scientific research. Personality is one of what are likely many unobserved confounds, each of which risks biasing the estimates of measurement effects derived using cross-sectional designs.

More concerningly, cross-sectional designs rely on the assumption that the covariates used to model selection are themselves unaffected by measurement effects (i.e., that they are *mode insensitive*). This assumption arises from the fact that covariates in cross-sectional designs

are almost always assessed post-treatment (i.e., in different modes); covariate imbalances could thus be caused by selection *and* measurement effects (Vannieuwenhuyze and Loosveldt 2013). Cross-sectional models that include mode sensitive variables risk introducing post-treatment bias (Montgomery et al. 2018). The dilemma researchers face is that it is difficult to know how to weigh between post-treatment bias and omitted variable bias for mode sensitive variables. Thus, cross-sectional estimates risk sensitivity to idiosyncratic modeling decisions without strong theoretical or empirical justification (Ho et al. 2007).

Experimental Designs

Given the shortcomings of cross-sectional designs, researchers have used experiments to identify measurement effects. By randomizing respondents into modes, experiments can obviate selection effects—an ideal design for isolating measurement effects.

In practice, however, experiments often suffer from major deficiencies. First, as noted by Endres et al. (2022), experimental studies usually randomize respondents into modes *before* they agree to participate. Thus, although mode assignments are random, selection bias is introduced if respondents differentially opt to participate based on their assigned mode. Experimental designs with this design suffer the same limitations as cross-sectional designs, despite "randomization" (Hernán et al. 2013). Few experiments randomize mode *after* respondents agree to participate; for example, Endres et al. (2022) bring individuals into a lab and then have them complete an in-person or video interview. There are seemingly only two other experiments with similar designs: Chang and Krosnick (2010) and Gooch and Vavreck (2019). The dearth of experiments where randomization occurs after ensuring respondents' participation reflects the costliness of this design relative to cross-sectional designs that can often leverage publicly-available survey data.

A second concern with experimental studies of mode effects is that they are infeasible to field on representative samples if researchers intend to randomize respondents post-selection. Well-designed experiments have only been conducted on convenience samples in lab settings, and although their internal validity is without question, convenience samples often differ greatly from the populations of interest (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 to a lab, and then randomly assigning modes—all to identify measurement effects—is prohibitively expensive. In practice, lab experiments have traded a degree of generalizability for decreased costs because they are primarily concerned with maximizing internal validity. Experimental estimates thus risk not generalizing to populations of interest if there are heterogeneous measurement effects as functions of imbalanced sample characteristics 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 that can credibly identify measurement effects—difference-in-differences (DD). DD involves weaker assumptions than cross-sectional designs and offers a cost-free, generalizable 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 the 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

the introduction of treatment—i.e., the "parallel trends" assumption. In the potential outcomes framework for causal inference (Rubin 2005), control units are used to impute the unobserved post-treatment outcomes for treated units. Post-treatment divergence in the groups' outcomes is 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*, and thus it allows for selection biases that often undermine cross-sectional designs; instead, DD assumes treatment assignments are mean-independent of characteristics that affect outcome *trends*.

For the purpose of identifying measurement effects, we can consider modes as treatment statuses. I propose DD can be used to identify the measurement effects of survey modes when:

(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 average treatment effect on the treated (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 with the observed untreated outcomes of control units.

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-intervention 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-trends are neither necessary nor sufficient for the assumption to hold (Kahn-Lang and Lang 2020). Additionally, parallel trends 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 can be assessed using a survey's sampling design. For example, a survey that samples individuals independently 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 after they have completed the survey at hand, there is little reason to be concerned about anticipatory treatment effects.

Although DD is not a foolproof approach for causal inference, it involves weaker assumptions than cross-sectional designs that often face some combination of omitted variable bias and post-treatment bias (Vannieuwenhuyze and Loosveldt 2013). Additionally, DD provides a cost-free alternative to experiments where panel surveys are publicly-available. Usefully, DD is increasingly feasible as major panel surveys turn to mixed-mode designs (Cernat and Sakshaug 2021). In the following sections, I show DD applications in the simple two-wave case (Section 3) and in a more complex case with 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-group, two-wave (2x2) case. I examine a widely-used survey for political science research—the 2016-2020 American National Election Study (ANES). I look at two outcomes theorized to be mode sensitive: racial attitudes and political knowledge (see Appendix A for detailed data descriptions and question wordings).

Data and Methodology

To use DD, a survey must interview the same respondents in multiple waves, interview some in *different* modes across waves, and interview 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 both 2016 and 2020. I can 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) that captures 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 senators' term lengths, two about 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 more 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 get more items correct in online surveys than live interviews because online modes allow respondents to 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 (2FE) regression with fixed effects for units and time.

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 treatment and control groups report near-identical racial resentment in the same mode in 2020, which suggests the observed difference between the groups in 2016 is largely a product of measurement effects.

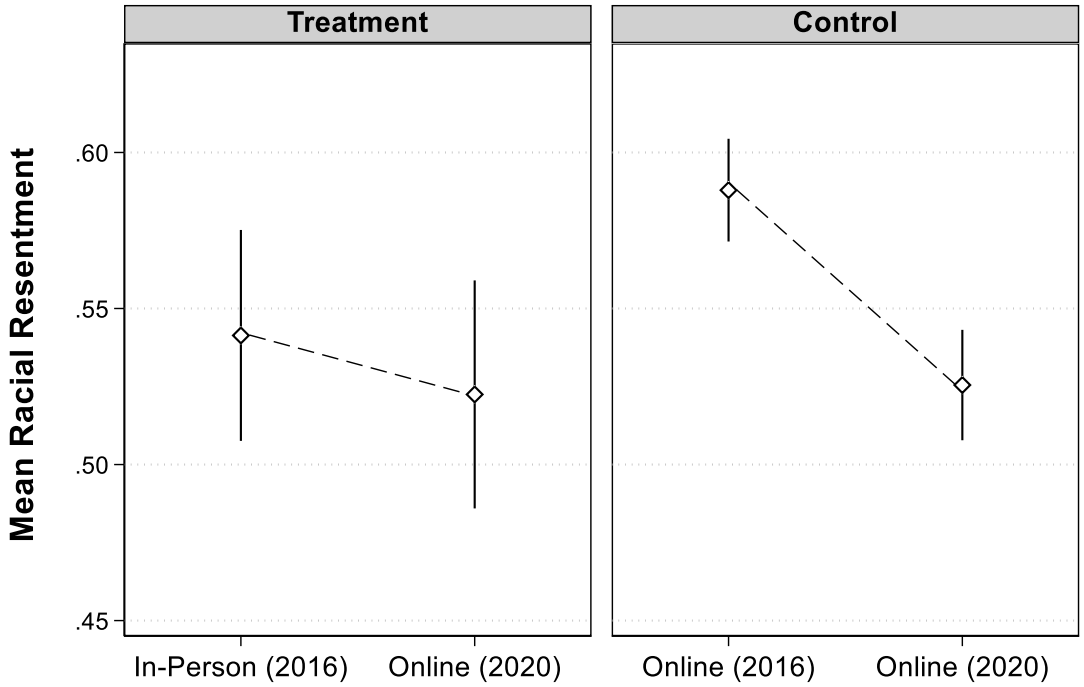


Figure 1—Change in Racial Resentment by 2016 Interview Mode. Figures show mean racial resentment by wave with 95 percent confidence intervals. Treatment respondents switch from in-person in 2016 to online in 2020 (n=435). Control respondents are online in 2016 and 2020 (n=1,471). Data weighted. White non-Hispanic subsample only. Source: 2016-2020 ANES.

The observed measurement effect for racial resentment has important implications. First, the 2016-2020 ANES panel understates individual-level decreases in 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 ANES abandoned in-person interviews for online surveys in 2020. Second, these findings bolster conclusions about socially-desirable responding on racial attitude measures which have primarily

been derived from cross-sectional studies less well-equipped to address confounding and post-treatment bias (e.g., Abrajano and Alvarez 2019; Bowyer and Rogowski 2017; Krysan 1998).[1]

Turning to political knowledge, in Figure 2, I plot the average number of eight political knowledge items correctly answered across waves by 2016 mode. Treatment respondents, on average, correctly answered 4.00 items in-person in 2016, and 4.91 items online in 2020—a 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* ($p=0.055$). The ATT is -1.03 correct knowledge items ($p<0.001$); 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.

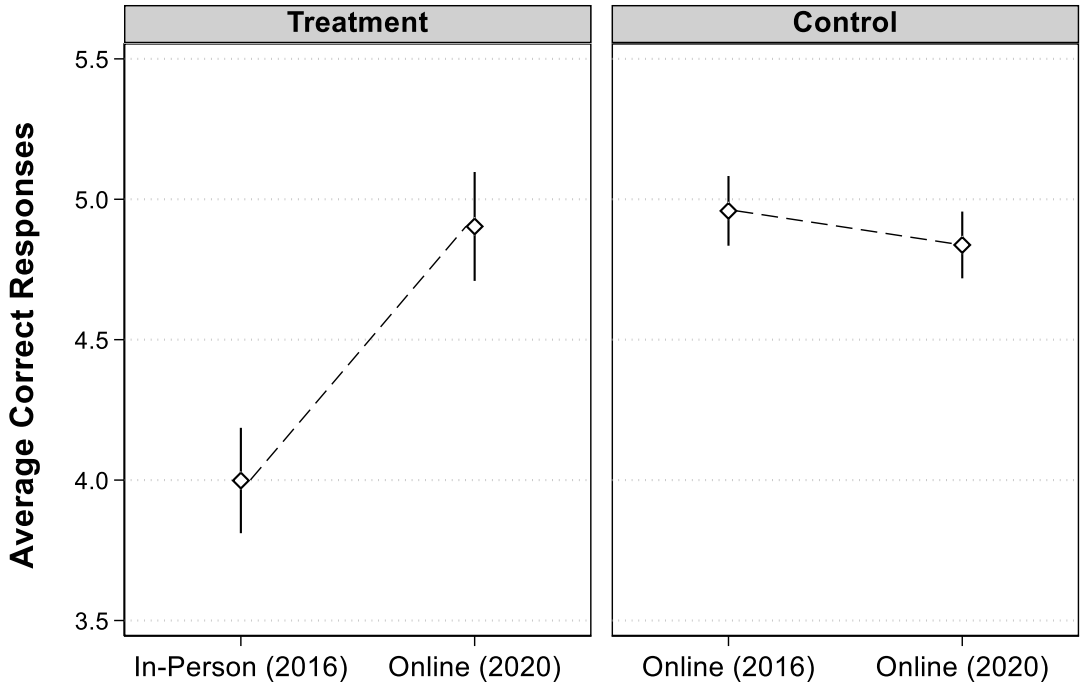


Figure 2—Change in Political Knowledge by 2016 Interview Mode. Figures show the average of eight political knowledge correctly answered by wave with 95 percent confidence intervals. Treatment respondents switch from in-person in 2016 to online in 2020 (n=636). Control respondents are online in 2016 and 2020 (n=2,011). Data weighted. Source: 2016-2020 ANES.

[1] Chang and Krosnick (2010) also show experimentally that whites engage in socially-desirable responding on a question about government assistance to Blacks.

Usefully, this analysis can essentially rule out selection effects as an explanation for the differences in political knowledge across modes in 2016 (Ansolabehere and Schaffner 2014). If selection had caused the difference in political knowledge across modes in 2016, this difference would likely also be observable in 2020. However, the average number of correct items between these groups are statistically indistinguishable when they both complete online surveys in 2020 ($p=0.657$). These results imply that mode primarily matters for assessing political knowledge due to measurement effects. Although the degree of measurement bias will differ based on the political knowledge items asked and the methods used to deter search (Graham 2022), my results support the claim that online respondents often appear more politically knowledgeable because they look up 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 can bolster the credibility of DD. Unfortunately, many robustness checks (e.g., pre-trend placebo tests) require more than two panel waves. Even in 2x2 cases, however, we have some data that can help augment the parallel trends assumption: covariates. Covariate imbalances between treatment and control groups can undermine DD if the covariates are related to outcome trends. Addressing covariate imbalances can thus bolster a DD design's credibility.

There are several ways to make covariate adjustments. Covariates can be used to match treatment respondents to similar control respondents, or to create weights that correct covariate imbalances across treatment and control groups. Importantly, in surveys that meet the criteria for DD, covariates can be derived *from the same wave and from the same mode* for all respondents, avoiding bias when adjusting for potentially mode sensitive variables.

In Table 1, I compare the ATTs without covariate adjustment and with entropy balancing weights that balance the treatment and control groups (Hainmueller 2012; Hainmueller and Xu 2013).[2] The covariates I balance on are measured in 2020 and include age, race/ethnicity, gender, marriage, parent, education, income, urbanicity, religiosity, union membership, unemployment, political interest, ideology, partisanship, and internet access. The ATTs are near-identical, which is unsurprising since the imbalances are small in this case (Appendix B). But when there are larger imbalances, covariate adjustments can augment the critical parallel trends assumption.

	Racial Resentment		**Political Knowledge**	
Weights	*Sampling*	*Balanced*	*Sampling*	*Balanced*
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 Adjustment. Entries are ATTs with standard errors in parentheses. Negative estimates indicate lower racial resentment or political knowledge in the in-person relative to online mode. Sampling weights constructed by the ANES. Balanced weights derived with entropy balancing. Source: 2016-2020 ANES.

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

In the canonical 2x2 case, DD estimation can be accomplished using 2FE regression. Unfortunately, although researchers often use 2FE in applications with more than two periods or staggered treatments, recent work shows that the 2FE model is *not* equivalent to a DD estimator in these common applications when treatment effects are heterogeneous (Borusyak et al. 2023; Goodman-Bacon 2021; Imai and Kim 2021). Many longitudinal surveys have many waves and staggered mode switches. In such cases, 2FE is inappropriate because homogeneous effects should not be assumed. Instead, researchers should use heterogeneity-robust DD methods to identify measurement effects, which I illustrate in this section.

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

Data and Methodology

I examine the Health and Retirement Study (HRS) Cognitive Functioning Measures, a 15-wave, biennial (1992-2020) survey by the University of Michigan Survey Research Center that tracks episodic memory, mental status, and vocabulary (McArdle et al. 2007). Among other uses, the HRS has allowed researchers to track Americans' cognitive functioning.[3] HRS waves 1 and 2 used different questionnaires than the rest of the panel, and a new cohort was added wave 4; for these reasons, I examine the cohort of 3,843 who entered the HRS by wave 4 and participated through wave 15 (of course, the cohort of interest can differ based on researchers' target population).

The HRS implemented design changes in its 14th and 15th waves. In waves 4-13, all respondents were surveyed via telephone or in-person interviews. In waves 14 and 15, however, the HRS switched different sets of panelists online: 377 respondents shifted online for wave 14, only to return to live interviews in wave 15; 504 respondents completed live interviews in wave 14, but shifted online in wave 15; 2,962 respondents completed live interviews in both waves 14 and 15. Thus, the HRS has mode switches that are *staggered* and *non-absorbing* (i.e., the online panelists in wave 14 switch back to live interviews for wave 15).

The HRS includes several measures of cognitive functioning; I focus my illustration on immediate word recalls (IWR), though DD could be applied equally well to other measures. To assess IWR in live interviews, interviewers read a list of 10 nouns and then ask respondents to report as many words as they can recall, in any order. To assess IWR in online surveys, the 10 nouns appear one at a time onscreen, and respondents then type as many as they can recall. The

[3] The HRS addresses item nonresponse using imputation. See Appendix A for more details.

more words correctly recalled, the better the individual's cognitive functioning.[4] Although there is not much research on mode effects for word recalls, a cross-sectional study found IWRs were 0.85 words out of ten higher in online surveys than live interviews (Runge et al. 2015). However, Runge et al. (2015) controls for self-rated memory, which I find in supplemental analyses to be *extremely* mode sensitive (Appendix D). Conditioning on mode sensitive variables risks post-treatment bias. DD is advantageous because it can address selection bias without conditioning on mode sensitive variables.

In Figure 3, I plot trends in IWR for the cohort of 3,843 panelists who participated in the HRS from waves 4 to 15. I plot trends separately for those who switched online in wave 14 (left) and wave 15 (right) and compare both to the control group that never switched online. We can see a few things about IWR from Figure 3. First, IWR declines over time, likely due to a decline in cognitive functioning from aging 22 years (McArdle et al. 2007). Second, there are differences in IWR for respondents who eventually switched online in waves 14 and 15 and those who never switched online. In live interviews between waves 4-13, panelists who would later switch online recalled 0.50-0.75 more words on average than those who never switched online. This is likely a consequence of mode switches occurring more for respondents with characteristics (e.g., youth, internet access) associated with higher IWR. Third, though Figure 3 does not offer a formal test for parallel pre-trends, the groups seem to mostly move in parallel from waves 4 to 13. Finally, the differences between the groups seem to expand in the waves when respondents switch online, though it is not immediately clear from Figure 3 what exactly the effect of switching modes was.

[4] There are four IWR word sets given in a randomized sequence across waves (see Appendix A).

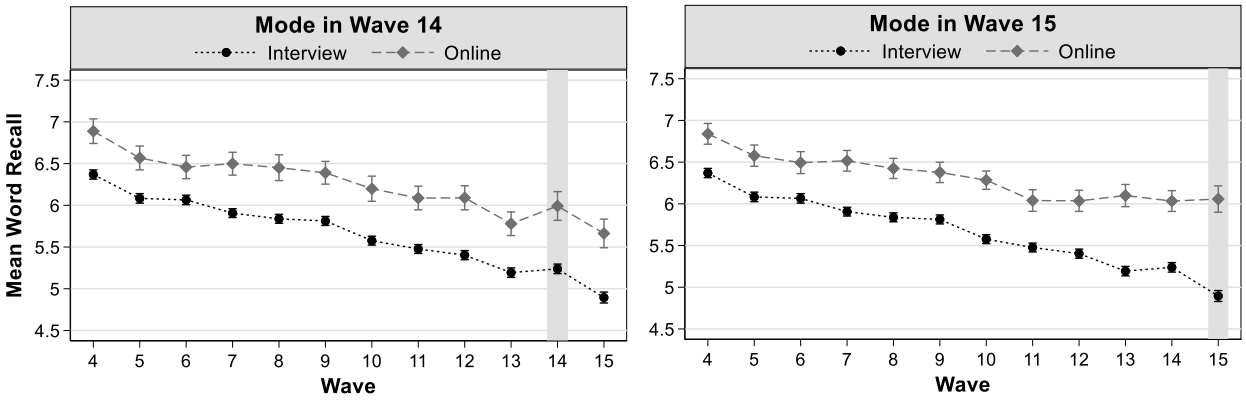


Figure 3—Trends in Immediate Word Recalls (1998-2020). Points are averages of ten nouns recalled with 95 percent confidence intervals. "Interview" groups are surveyed using a mix of in-person and telephone interviews (n=2,962). "Online" groups are surveyed using live interviews until waves 14 (n=377) or 15 (n=504), when they are surveyed online. Biennial waves. Source: Health and Retirement Study.

To summarize, I have one outcome (IWR) measured in 12 waves (waves 4-15) with a staggered, non-absorbing treatment where some respondents shift online instead of completing a live interview in waves 14 and 15. 2FE is inappropriate (Imai and Kim 2021; Sun and Abraham 2021), so I use a heterogeneity-robust estimator from de Chaisemartin and D'Haultfœuille (CD) (2020) designed for staggered, non-absorbing treatments. Usefully, the CD (2020) estimator can probe the parallel trends assumption with placebo tests in pre-treatment intervals, where effects should ideally be non-significant since there is no reason to assume anticipatory effects given the HRS's design (Appendix A).

Results

In Figure 4, I use CD's (2020) heterogeneity-robust DD estimator to estimate the ATT of having completed the IWR task online as opposed to in a live interview in waves 14 or 15. I plot the ATTs such that "0" corresponds to treatment onset in wave 14 or 15, depending on when the respondent switched modes, and the negative waves are placebo tests occurring before treatment

onset (e.g., "-1" is the wave immediately before treatment onset). The primary ATT of interest is the coefficient in the 0th period, while the ATTs in preceding waves are pre-trends placebo tests.[5]

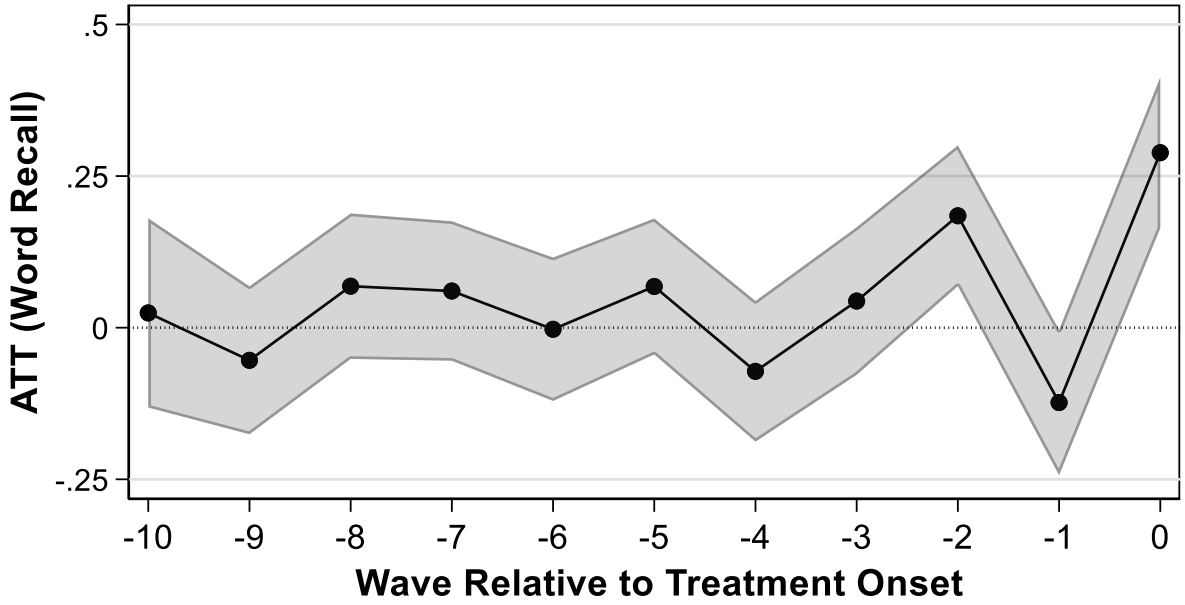


Figure 4—Effect of Switching to Online Mode on Immediate Word Recalls. Points are ATTs of switching from live to online interviews on average word recalls out of ten with 95 percent confidence intervals (bootstrapped standard errors). Treatment occurs in the 0th period. Periods -10 to -1 are placebo tests. Biennial waves (1998-2020). Source: Health and Retirement Study.

The ATT of completing the IWR online instead of in a live interview is a 0.30 increase in word recalls ($p<0.001$). The 0.30-word measurement effect seems small, but an effect of this size is approximately twice the typical biennial IWR decline in the HRS. HRS waves 14 and 15 thus understate declines in IWR because 10% and 13% of the sample, respectively, shifted online for these waves. Shifting more panelists online would be consequential for assessing trends in IWR because the online measurement effect will confound the expected declining biennial trends.

Robustness Check #2: Placebo Tests

The credibility of the DD would be bolstered by a lack of consistent pre-treatment trends. Of ten placebo tests, one is significantly different from zero (Figure 4). Unfortunately, this pre-

[5] The placebo estimates are provided in Appendix Table C1.

treatment divergence occurs just two periods before treatment onset; it could be the case that the treatment and control groups moved in parallel during earlier waves, but began diverging closer to treatment onset. Alternatively, with so many placebo tests, it is not surprising that one ends up significant by chance. Indeed, there is some evidence of mean reversion from this divergence just prior to treatment onset (Wave "-1"). These placebo tests thus mostly bolster the parallel trends assumption. However, it is important to keep in mind that a lack of consistent pre-trends does not confirm the parallel trends assumption holds; parallel trends is ultimately untestable because it involves unobserved potential outcomes.

Robustness Check #3: Alternative Estimators

Econometricians have only recently begun scrutinizing 2FE for DD, and they have not settled on a standard heterogeneity-robust estimator in lieu of it. Several other estimators would have been as appropriate as CD (2020), each primarily differing in the control observations used as comparison units (and/or with what weights). Ideally, the estimated treatment effect would not be sensitive to the choice of equally-appropriate estimators. In Table 2, I compare the ATT from CD (2020) to those derived with two other widely-utilized estimators—Callaway and Sant'Anna (CS) (2021) and Borusyak, Jaravel, and Spiess (BJS) (2023)—and a 2FE model (for a review of the similarities and differences between these estimators, see Roth et al. 2023). Notably, all four estimators show significant effects ranging from approximately three-tenths to four-tenths word recall increases in the online mode. These results offer consistent evidence that completing the IWR task online rather than with a live interviewer modestly increases immediate word recalls.

de Chaisemartin and D'Haultfœuille (2020)	Callaway and Sant'Anna (2021)	Borusyak, Jaravel, and Spiess (2023)	Two-Way Fixed Effects (2FE)
0.30 (0.06)	0.28 (0.06)	0.38 (0.06)	0.36 (0.05)

Table 2—ATTs on Immediate Word Recalls by Estimator. Entries are ATTs with standard errors in parentheses. Positive estimates indicate increased word recalls in the online versus live mode. Source: Health and Retirement Study.

Robustness Check #4: Unit-Specific Trends

Although the placebo tests do not show clear evidence of non-parallel pre-trends, we may want to ensure the results are robust to possible violations of parallel trends since these introduce bias (Hassell and Holbein n.d.). One common approach for addressing non-parallel pre-trends is to account for *unit-specific trends* by including a linear (or higher-order) interaction between unit and time in 2FE models. However, unit-specific trends do *not* obviate the need for heterogeneity-robustness and, not all heterogeneity-robust estimators can incorporate unit-specific trends.[6]

In Table 3, I estimate the ATT of having completed the IWR task online instead of in a live interview using two heterogeneity-robust estimators and a 2FE model, all with linear unit-specific trends. The CD (2020) estimate is essentially unchanged—the ATT of online mode on IWR is 0.29 instead of 0.30 words. However, including linear trends decreases the estimated effect using the BJS (2023) and 2FE estimators from just under four-tenths of a word to just over two-tenths of a word. These attenuations are consistent with minor violations of parallel trends where treated units decline less in IWR over time than untreated units. In cases with more severe divergences, unit-specific trends can be crucial for identifying treatment effects with DD designs.

de Chaisemartin and D'Haultfœuille (2020)	Borusyak, Jaravel, and Spiess (2023)	Two-Way Fixed Effects (2FE) w/Linear Trends
0.29	0.23	0.20
(0.07)	(0.06)	(0.06)

Table 3—ATTs on Immediate Word Recalls with Linear Unit-Specific Trends. Entries are ATTs with standard errors in parentheses. Positive estimates indicate increased word recalls in the online versus live mode. Source: Health and Retirement Study.

5. Conclusion

In this paper, I proposed difference-in-differences (DD) designs can be used to identify mode effects. DD can be employed with panel surveys where some respondents switch modes

[6] Callaway and Sant'Anna's (2021) estimator currently does not support unit-specific trends.

across waves while others stay in the same mode across waves. DD assumes parallel trends in potential outcomes—i.e., that individuals who did vs. did not switch modes would have moved in parallel absent switching modes. Parallel trends is a strong assumption that needs to be gauged and augmented, but it is still weaker than the assumptions of cross-sectional designs which must simultaneously assume no omitted variable bias and that none of the variables used to address omitted variable bias are mode sensitive. DD thus offers a credible alternative to common cross-sectional designs for estimating measurement effects, as well as lab experiments which are, in practice, rare given the cost of randomizing modes *after* researchers have ensured respondents will complete the survey in whichever mode they are assigned to avoid differential selection.

There remains substantial work to do applying DD to understand mode effects. I examine three outcomes that have received scholarly attention as mode sensitive (racial attitudes, political knowledge, immediate word recalls). Future work should apply DD to identify mode effects for other survey outcomes, especially those that have only been studied with cross-sectional designs. Further, I defined the difference-in-means as the target estimand in all three cases, but mode can affect other survey response patterns. Homola et al. (2016), for example, find online respondents have more dispersed responses than in-person interviewees, but Bowyer and Rogowski (2017) reach the opposite conclusion examining online versus phone respondents. Extensions of DD focusing on mode's distributional consequences (e.g., quantile effects; Callaway and Li 2019) could help adjudicate competing claims about mode's effects on survey responses.

Finally, the DD has immediate applications for longitudinal survey design. Measurement effects confound estimates of individual-level change (Cernat and Sakshaug 2021). To identify and, in turn, account for measurement effects, longitudinal surveys should consider phasing in modes such that some panelists continue taking the survey in their original modes. Such designs

will allow researchers to quantify the bias introduced by mode switches, increasing the validity of inferences comparing responses derived from different 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.

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225(2): 200–230.

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 Questionnaires: An Experiment." *Public Opinion Quarterly* 74(1): 154–67.

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.

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 Difference-in-Differences 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.

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).

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, Mingnan, and Yichen Wang. 2014. “Data Collection Mode Effects On Political Knowledge.” *Survey Methods: Insights from the Field (SMIF)*. https://surveyinsights.org/?p=5317 (April 19, 2023).

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.

McArdle, John J., Gwenith G. Fisher, and Kelly M. Kadlec. 2007. “Latent Variable Analyses of Age Trends of Cognition in the Health and Retirement Study, 1992-2004.” *Psychology and Aging* 22(3): 525–45.

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: 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*. https://www.sciencedirect.com/science/article/pii/S0304407623001318 (May 9, 2023).

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 wave was fielded September 7, 2016 to November 7, 2016. The 2016 post-election wave was fielded November 9, 2016 to January 8, 2017. The 2020 pre-election wave was fielded August 18, 2020 to November 3, 2020. The 2020 post-election wave was fielded November 8, 2020 to 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://hrsdata.isr.umich.edu/sites/default/files/documentation/data-descriptions/1686605147/COGIMP9220_dd.pdf?_gl=1*1ypf7d9*_ga*MTg3MzQwNzY3MC4xNjg3MjkxMTgx*_ga_FF28MW3MW2*MTY4NzI5MTE4MS4xLjEuMTY4NzI5MTIxOS4wLjAuMA..&_ga=2.59863213.1875097192.1687291182-1873407670.1687291181

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.

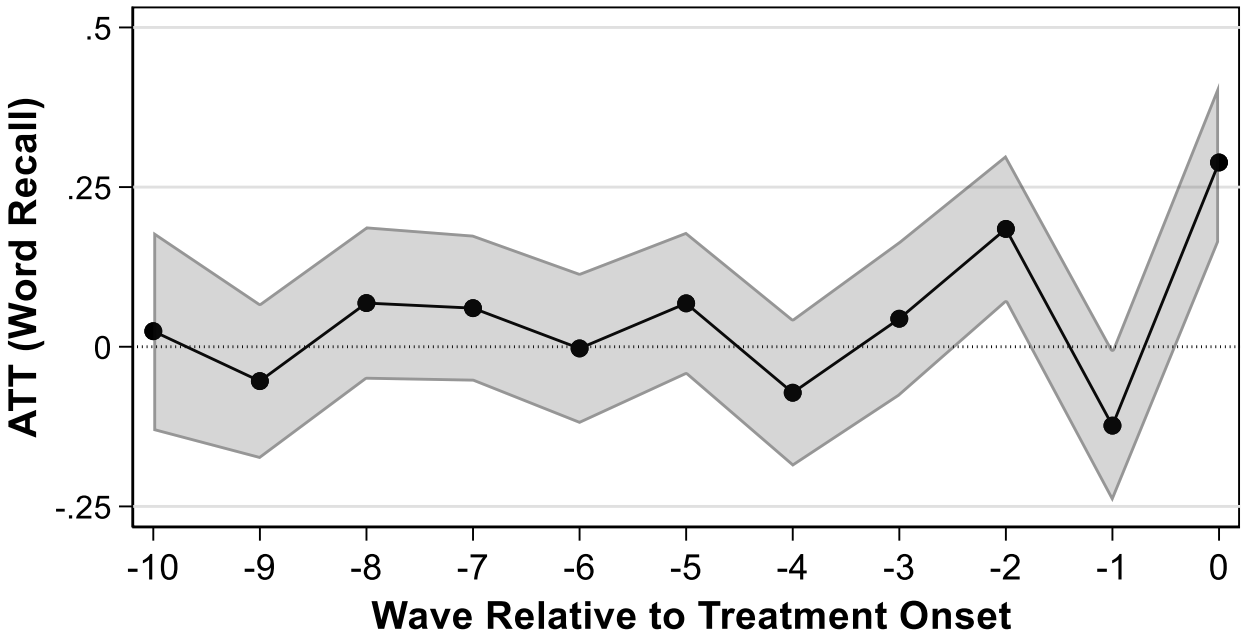
Appendix C: Full Results for HRS Immediate Word Recall Analysis

Table C1—ATTs of Switching to Online Mode on Immediate Word Recalls

Wave	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0
ATT	0.02	-0.05	0.07	0.06	-0.00	0.07	-0.07	0.04	0.18	-0.12	0.29
	(0.08)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)

Note: Entries are ATTs of taking online survey instead of live interview on average word recalls out of ten with bootstrapped standard errors in parentheses. Treatment occurs in the 0th period. Periods -10 to -1 are placebo tests. Biennial waves (1998-2020). Estimator from de Chaisemartin and D'Haultfœuille (2020). Estimates correspond to those plotted in Figure 4 (reproduced below). Source: Health and Retirement Study.

Figure 4 (Reproduced)—ATTs of Switching to Online Mode on Immediate Word Recalls



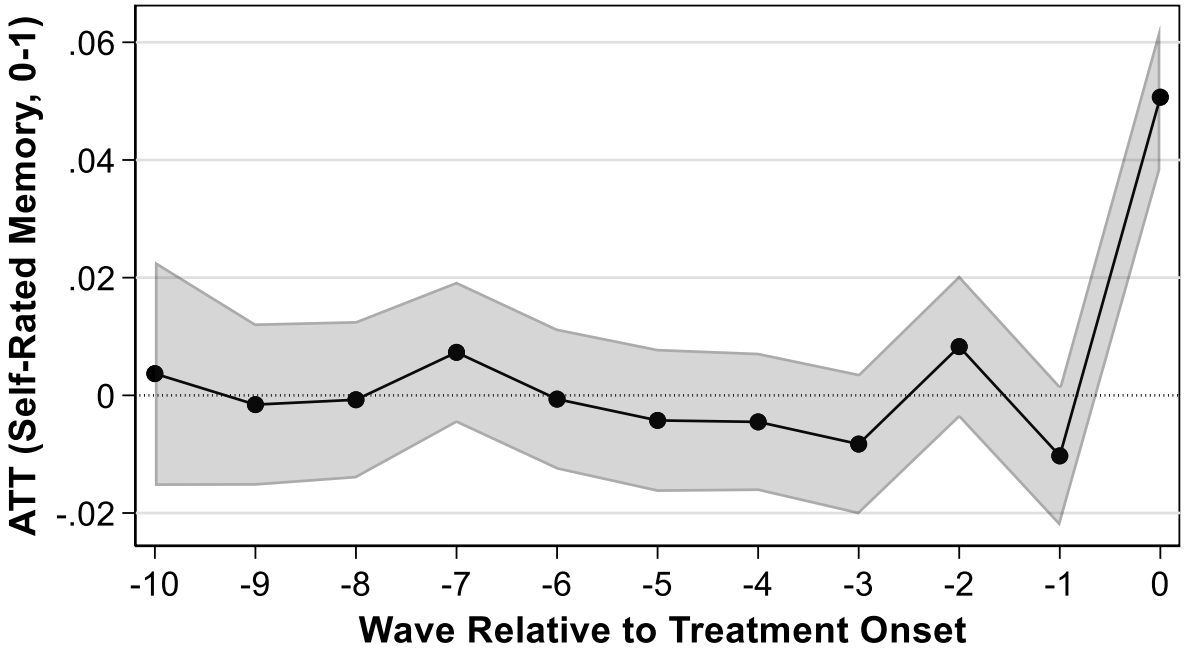
Note: Points are ATTs of taking online survey instead of live interview on average word recalls out of ten with 95 percent confidence intervals (bootstrapped standard errors). Treatment occurs in the 0th period. Periods -10 to -1 are placebo tests using pre-treatment waves. Biennial waves (1998-2020). Source: Health and Retirement Study.

Appendix D: Showing Self-Rated Memory is Mode Sensitive using Difference-in-Differences

In Figure D1, I plot the difference-in-differences (ATTs) of switching to the online mode from live interviews (in-person or phone) in the HRS on self-rated memory with de Chaisemartin and D'Haultfœuille's (2020) estimator. Self-rated memory is measured on a five-point scale that I rescaled from 0 (poor) to 1 (excellent). The ATT of primary interest is in the 0^{th} period, which corresponds to the 14^{th} or 15^{th} waves depending on when respondents switched online. The ATTs in waves preceding the 0^{th} period are pre-trends placebo tests. None of the placebo tests emerges as statistically significant, indicating no violations of parallel pre-trends.

The ATT of having completed the HRS in waves 14 or 15 online rather than through live interviews is a 0.05 increase in self-rated memory on a 0 to 1 scale ($p<0.001$). Self-rated memory is mode sensitive. Thus, when Runge et al. (2015) examined measurement effects on immediate word recalls using a cross-sectional design that controlled for self-rated memory, they very likely introduced post-treatment bias. Potentially, however, differences in self-rated memory by modes in Runge et al. (2015) also include selection effects; the omission of self-rated memory would therefore potentially introduce omitted variable bias. The DD avoids this trade-off between post-treatment bias and omitted variable bias when estimating online mode measurement effects on immediate word recalls, making it a better approach for recovering causal measurement effects.

Figure D1—Effect of Switching to Online Mode on Self-Rated Memory



Note: Points are ATTs of switching from live interviews to online surveys on self-rated memory on a 0 (poor) to 1 (excellent) scale with 95 percent confidence intervals (bootstrapped standard errors). Treatment occurs in the 0^{th} period. Periods -10 to -1 are placebo tests. Biennial waves (1998-2020). Source: Health and Retirement Study.