Correcting Measurement Error Bias in Conjoint Survey Experiments: Supplementary Appendices*

Katherine Clayton† Yusaku Horiuchi‡ Aaron R. Kaufman§
Gary King¶ Mayya Komisarchik‖

July 23, 2023

*Our paper and this accompanying Supplementary Appendix are available at GaryKing.org/conjointE.
†Department of Political Science, Stanford University. kpc14@stanford.edu, kpclayton.com
‡Department of Government, Dartmouth College. yusaku.horiuchi@dartmouth.edu, horiuchi.org
§Division of Social Sciences, New York University Abu Dhabi. aaronkaufman@nyu.edu, aaronrkaufman.com
¶Institute for Quantitative Social Science, Harvard University. king@harvard.edu, GaryKing.org
‖Department of Political Science, University of Rochester. mayya.komisarchik@rochester.edu, mayyakomisarchik.com

Contents

A1 Prior Conjoint Studies

We replicate only the eight conjoint studies listed in Table A1; the conjoint method has been used by many others in our discipline and beyond. Systematic reviews of conjoint applications in political science include De la Cuesta, Egami, and Imai (2022), which finds that 59 conjoint experiments were published in ten of the discipline's top journals from 2014 to 2019, and Ganter (2021), indicating that 61 conjoint experiments appeared in six of the discipline's top journals from 2014 to 2021. Likewise, Schwarz and Coppock (2022) analyze 67 candidate-related conjoint experiments that include a gender attribute, Eshima and Smith (2022) analyze 16 candidate conjoint experiments that include an age attribute, and Incerti (2020) finds 26 studies that study candidate corruption and vote choice.

Outside of political science, conjoint experiments are no less popular. In environmental science, Alriksson and Öberg (2008) record 84 studies evaluating preferences for environmental policy and Mamine, Minviel, et al. (2020) list 70 studies related to agri-environmental practices; in marketing, Bastounis et al. (2021) analyze 43 conjoint experiments manipulating sustainability labeling on food products.

Across all fields, a search for "conjoint analysis" in Google Scholar returns 143,000 articles (accessed July 13, 2023).

A2 Alternative Observation Mechanisms

Every statistical model makes assumptions about the data generation process, and on the basis of these assumptions makes claims about statistical properties. Although not fully observable, researchers should always be cognizant of the implicit assumptions they are making. Our key assumption involves the "observation mechanism": the process by which a respondent's true preference is translated into a survey response.

In Section 3.1, we use the "swapping error" observation mechanism, which generalizes the "error-free" observation mechanism assumed or implied in most prior literature. We prove that our correction is vital under this observation mechanism, but it is not so under every conceivable observation mechanism. In this section we discuss six potential

observation mechanisms; some we reject them as implausible or as violating observed patterns, and for the remainder, we show that our correction is still valid.

First, we consider the *error-free* observation mechanism, which is assumed throughout most prior conjoint literature. Under this optimistic mechanism, respondents make choices identical to their fixed preferences: $C_i(a) = \rho_i(a)$. Assuming "fixed preferences" does not necessarily imply that respondents come to a survey knowing which candidate they would prefer among every possible set of randomly generated attributes, most of which they have never before seen or contemplated and some of which may not exist in the world (Bansak and Jenke, 2023). Instead, under this mechanism, researchers implicitly assume that each person has a set of observed and unobserved characteristics that determine their preferences and, more importantly, that these determinants stay fixed for at least the duration of the survey. Accordingly, a stochastic component does not affect observed choices. Put differently, this observation mechanism implies that intra-respondent reliability (IRR) calculated from two survey questions asked only moments apart should be 100% (or equivalently, $\tau = 0$). Because this model implication is contradicted by the numerous analyses of real data in Section 4, the error-free observation mechanism should be rejected as a model of the survey response in conjoint analysis.

Instead, the *swapping error* observation mechanism in Section 3.1 assumes that individuals make choices that differ from their preferences due to swapping error: "Preferences may be stable, but subjects make stochastic mistakes; this is the approach implicit in typical empirical analysis" (Agranov and Ortoleva, 2022).

Nevertheless, many different models of both the survey process and of human decision-making also lead to imperfect IRR. Some such models include errors in seeing, reading, understanding, thinking, deciding, recording, communicating, etc. In some cases, respondents may even make an intentional decision to randomize choices. Regardless of the exact sources of the respondents' stochastic choices, measurement error bias corrections, such as the ones proposed in the paper, are required for valid inference.

For expository clarity, consider a third observation mechanism, an extreme (and extremely implausible) case. This is the *defier* observation mechanism, where people in-

tentionally answer surveys in ways that are opposite to their preferences; that is, $C_i(a) = 1-\rho_i(a)$. If this is how respondents answer surveys, it would, of course, flip signs from almost all empirical analyses in the existing literature, and applying our measurement error corrections would only make matters worse. In fact, the defier mechanism is observationally equivalent to the error-free mechanism. We can, however, reject this deterministic mechanism because it implies $\tau = 0$. However, an alternative *random defier* observation mechanism (i.e., substituting $1 - \rho_i(a)$ for $\rho_i(a)$ in Equation 5) cannot be rejected for this reason. Fortunately, although we all know some contrarians, many years of detailed qualitative research into the social and cognitive aspects of the survey interview is sufficient to ignore these extreme mechanisms (e.g. Schwarz, 2007).

Fourth, the *rational choice random utility* observation mechanism is a modification of the rational choice fixed preference probability mechanism, where the utility $U_{ij}(a) = f_{ij}(a) + \epsilon_{ij}$, for choice $j = \{0, 1\}$ and individual i, is decomposed into a systematic component (a function of the attribute values) and a stochastic component. The mechanism, which predicts that rational respondents will maximize their utility by choosing Candidate 0 whenever $U_{i0}(a) > U_{i1}(a)$ and Candidate 1 otherwise, allows for IRR less than 100%. Thus, if respondents act in this way — and, in addition, no swapping error exists — then our measurement error bias correction should be avoided. However, although the random utility model has a venerable history, "A fairly large body of experimental evidence. . . shows that subjects systematically make choices that violate properties required by expected utility" (Keller, 1992). This is consistent with direct evidence from conjoint studies (Jenke et al., 2021). Violations of this mechanism are most prevalent when stakes are low, and choices are difficult, which is usually the case for social science or marketing conjoint applications, and especially so with randomly assigned attributes. Moreover, our evidence from open-ended questions, where we ask respondents to recount how they came to their decision, indicates that typical conjoint tasks are usually viewed by respondents as involving subtle differences between profiles for respondents. For these reasons, we reject this observation mechanism, though we acknowledge that in some contexts it may be appropriate.

Fifth, under the *probability matching* observation mechanism, each respondent has a preference intensity $\pi_i(a) \in (0, 1)$ for Candidate 0 (compared to Candidate 1). These intensities then determine respondents' fixed preferences: $\rho_i(a) = \mathbf{1}(\pi_i > 0.5)$. The surprise from a rational choice perspective is that people, under this mechanism, do not set their choices equal to their preferences; they instead make choices with probability equal to their intensities: $C_i \sim \text{Bernoulli}[\pi_i(a)]$. For example, if respondents think they will prefer Candidate 0 60% of the time (or with 60% intensity), they vote for that candidate only 60% of the time. This mechanism may be surprising, but "probability matching has been observed in thousands of geographically diverse human subjects over several decades, as well as in other animal species, including ants, bees, fish, pigeons, and primates (citations omitted). In virtually any setting where a [human or nonhuman] animal is able to make a choice between A versus B in a randomized experiment, we observe probability matching" (Lo, Marlowe, and Zhang, 2021). As it turns out, this mechanism is almost the same as the swapping mechanism, which we can see by computing IRR as $1 - 2\pi_i(a)[1 - \pi_i(a)]$. Our measurement error bias corrections can then be used directly with an estimate of $\pi_i(a)$ or via our preferred technique of computing the difference between two questions in the same survey.

Finally, the *radical empiricism* observation mechanism holds that interpretations based on (true unobserved) preferences are unnecessary (James, 1975), with the possible exception sometimes of potential outcomes not yet observed. Instead, this mechanism consists solely of observed choices without anything unobservable postulated as leading to them. This is a coherent view, but adherents must come to terms with the fact that a conjoint survey conducted (say) Wednesday morning could give very different answers than the same survey conducted Thursday morning (as we know from our extensive evidence on IRR).[1] Nevertheless, the radical empiricist mechanism claims that whatever we happen

[1] Real-world elections have this property, where the law formally declares that whoever gets more votes wins office; claims that true population preferences differ from votes (i.e., the observed choices) would be ignored in the law. The same is true of U.S. census data as the basis for federal allocation and redistricting decisions. Any arguments about under- or over-counts after the data are promulgated every decade are ignored. Although conjoint surveys are intended to approximate real-world decision making, no similar law declares that observed choices in academic surveys must be taken so seriously. No law, court, or other authority has decreed that the choices reported in a survey on Wednesday morning are more meaningful than the other ones that would have been observed the next day.

to observe is everything we need to care about when studying real-world behavior and outcomes. What this means is that all inferences are *conditional* on whatever haphazard events occur on the day the survey is conducted. For this approach, these "Wednesday morning" events are not treated as confounders and instead define a new quantity to be estimated. By definition, this approach's validity cannot be rejected from evidence. Of course, social scientists usually insist that inferences are conditional on only justifiable conditions. This commonly shared view among social scientists is sufficient to reject this observation mechanism.

An infinite range of other observation mechanisms is, of course, possible, such as mixtures, combinations, or modifications of those listed here or others. Studying whether mechanisms other than swapping error or its special cases or equivalents may be consistent with empirical evidence could help improve any bias correction, and so would be a valuable topic for future research (see Louviere et al., 2002; Starmer, 2000).

A3 Open Ended Surveys

We conducted an additional survey to study the mechanisms that drive respondents to select the same or a different profile when two profiles are repeated just moments apart. Drawing inspiration from Arias and Blair's (2022) study on preferences for different types of migrants, we recruited 134 respondents from Lucid Theorem to participate in a conjoint experiment where they were asked to choose between hypothetical migrants who could be allowed to stay in their state or sent back to their region of origin (see "Open-ended comments" study in Table A11). Respondents completed six total conjoint tasks, with the first and sixth tasks repeated, and the order of the left and right profiles in the repeated task flipped. After making their final choice, we asked respondents to reflect more deeply on their answer with a follow-up question: "In a few sentences, please tell us how you thought about this question and how you reasoned to get to your answer. Did anything cause you to have second thoughts?"

Responses to this follow-up question were varied but generally quite thoughtful. Many respondents mentioned the attribute or attribute(s) they thought were most important for

their choice. Some respondents remarked that they thought conjoint-style questions were interesting, engaging, or difficult. Notably, none indicated that they realized that they were being shown an identical pair of profiles. More research to understand the mechanism that produces swapping error in conjoint studies would be valuable, but the available evidence seems to indicate that respondents do not notice being presented with the same pair of profiles, so it seems impossible for them to be intentionally choosing a different profile in the repeated question.

A4 Standard Errors

We now show how to compute standard errors for $\tilde{\rho}(a)$ and $\tilde{\theta}(a, a')$ in three ways — (1) analytical derivation for speed, (2) bootstrapping for convenience, and (3) simulation for familiarity. As we show after describing the methods, all three give approximately the same empirical estimates.

Our preferred method is based on an *analytical derivation*, which we give below. This method is the fastest, but it involves some technical mathematics. *Bootstrapping* is the simplest approach, but the slowest computationally; indeed, it is about 790 times slower than the analytical approach. To use this approach, draw a sample of respondents (not respondent-tasks) with replacement and calculate $\tilde{\rho}(a)$ and $\tilde{\theta}(a, a')$ as in Equation 11. Repeat this a large number of times and, for estimates of the standard errors, take the standard deviation across simulated datasets.

Our third and final method uses *simulation*. It is much faster than bootstrapping but about 60% slower than the analytical method. It is based on a Clarify-like approach more familiar to political scientists (King, Tomz, and Wittenberg, 2000). To estimate the standard error, repeatedly simulate $\rho(a)$ and τ from a bivariate normal (given estimates of parameter values from our analytical derivation below), plug them into Equation 11, and compute the standard deviation across simulations.

We now turn to our analytical approach, the main complication of which is taking the variance of a ratio (for either the marginal mean or AMCE, in Equation 11). This is not straightforward because the variance is a linear operator, but the ratio, of course, is not.

We thus take the first-order Taylor expansion (a linear approximation to the ratio). We write this approximation generically first and afterward apply it to our problem. For two correlated random variables R and S, we approximate a ratio R/S as

$$V(R/S) \approx \frac{E(R)^2}{E(S)^2}\left(\frac{V(R)}{E(R)^2} - 2\frac{\text{Cov}(R,S)}{E(R)E(S)} + \frac{V(S)}{E(S)^2}\right). \tag{1}$$

We now apply the approximation in Equation 1 to the AMCE, $\tilde{\theta}(a,a') = \hat{\theta}(a,a')/[1-2\hat{\tau}]$ from Equation 11. We first compute the moments: $E[\hat{\theta}(a,a')] = \theta(1-2\tau)$, $E(1-2\hat{\tau}) = 1-2\tau$, $V[\hat{\theta}(a,a')] \equiv \sigma^2_\theta$, and $V(1-2\hat{\tau}) = 4\sigma^2_\tau$, where $V(\hat{\tau}) \equiv \sigma^2_\tau$.

When $\hat{\tau}$ is estimated from the IRR via Equation 8, we approximate its variance by computing the first derivative, $\partial\hat{\tau}/\partial\text{IRR} = -\frac{1}{2}(2\cdot\widehat{\text{IRR}}-1)^{-1/2}$ evaluated at the point estimate, and then use the variance of its Taylor series approximation: $V(\hat{\tau}) \approx V(\widehat{\text{IRR}})/4(2\cdot\widehat{\text{IRR}}-1)$.

We will also need the covariance, $\text{Cov}[\hat{\theta}(a,a'),\hat{\tau}] = \text{Cov}[\hat{\rho}(a),\hat{\tau}] - \text{Cov}[\hat{\rho}(a'),\hat{\tau}]$, where, letting $d_i = \mathbf{1}(C_{i1} = C_{iT})$ equal 1 for agreement and 0 disagreement on the same item asked twice, $\text{IRR} = \sum_i d_i/n$. Thus,

$$\begin{aligned}
\phi_a \equiv \text{Cov}(\hat{\rho}(a),\hat{\tau}) &= \text{Cov}\left(\hat{\rho}(a), -\frac{1}{2}(2\cdot\widehat{\text{IRR}}-1)^{-1/2}\cdot\text{IRR}\right) \\
&= -\frac{1}{2}(2\cdot\widehat{\text{IRR}}-1)^{-1/2}\cdot\text{Cov}(\hat{\rho}(a),\text{IRR})
\end{aligned}$$

where

$$\begin{aligned}
\text{Cov}(\hat{\rho}(a),\text{IRR}) &= \text{Cov}\left(\frac{1}{n_a}\sum_{it|\ell=a} C_{it}, \frac{1}{n}\sum_i d_i\right) \\
&= \frac{1}{n_a}\sum_{it|\ell=a}\frac{1}{n}\sum_i \text{Cov}(C_{it},d_i) \\
&= \frac{1}{n_a n}\sum_{it|\ell=a}\text{Cov}(C_{it},d_i) \\
&= \frac{1}{n}\text{Cov}(C_{it},d_i),
\end{aligned}$$

using the assumptions that respondents are independent of each other and covariances are constant within the treated and within the control groups, and where n_a is the number of observations in the treated group.

We then compute the variance by applying Equation 1:

$$V[\tilde{\theta}(a,a')] \approx \theta(a,a')^2 \left(\frac{\sigma_{\hat{\theta}^2}}{\theta(a,a')^2(1-2\tau)^2} - \frac{2 \cdot \text{Cov}(\hat{\theta}(a,a'), -2\hat{\tau})}{\theta(a,a')(1-2\tau)^2} + \frac{4\sigma_{\hat{\tau}}^2}{(1-2\tau)^2} \right)$$
$$= \frac{\theta(a,a')^2}{(1-2\tau)^2} \left(\frac{\sigma_{\hat{\theta}}^2}{\theta(a,a')^2} + \frac{4(\phi_a - \phi_{a'})}{\theta(a,a')} + 4\sigma_{\hat{\tau}}^2 \right).$$

We then give our analytical (squared) standard error for the AMCE by replacing parameters with their point estimates:

$$\hat{V}[\tilde{\theta}(a,a')] = \frac{\tilde{\theta}(a,a')^2}{(1-2\hat{\tau})^2} \left(\frac{\hat{\sigma}_{\hat{\theta}}^2}{\tilde{\theta}(a,a')^2} + 4\frac{\hat{\phi}_a - \hat{\phi}_{a'}}{\tilde{\theta}(a,a')} + 4\hat{\sigma}_{\hat{\tau}}^2 \right).$$

We now apply the same logic to compute the standard error of the marginal mean, $\tilde{\rho}(a) = [\hat{\rho}(a) - \hat{\tau}]/(1-2\hat{\tau})$. We again collect the moments: $E(\hat{\rho}(a)) = \rho(a)(1-2\tau)+\tau$, $E(\hat{\rho}(a) - \hat{\tau}) = \rho(a)(1-2\tau)$, $E(1-2\hat{\tau}) = 1-2\tau$, $V(\hat{\rho}(a) - \hat{\tau}) = \sigma_{\hat{\rho}}^2 + \sigma_{\hat{\tau}}^2 - 2\phi_a$, $V(1-2\hat{\tau}) = 4\sigma_{\hat{\tau}}^2$, and $\text{Cov}(\hat{\rho}(a) - \hat{\tau}, 1-2\hat{\tau}) = 2(\phi_{\hat{\tau}}^2 - \phi_a)$.

We compute the variance of the marginal mean by applying Equation 1:

$$V(\tilde{\rho}(a)) \approx \frac{\rho(a)^2}{(1-2\tau)^2} \left(\frac{\sigma_{\hat{\rho}}^2 + \sigma_{\hat{\tau}}^2 - 2\phi_a}{\rho(a)^2} + 4\frac{\phi_a - \sigma_{\hat{\tau}}^2}{\rho(a)} + 4\sigma_{\hat{\tau}}^2 \right),$$

and, by replacing parameters with their point estimates, give the (squared) standard error of the marginal mean:

$$\hat{V}(\tilde{\rho}(a)) = \frac{\tilde{\rho}(a)^2}{(1-2\hat{\tau})^2} \left(\frac{\hat{\sigma}_{\hat{\rho}}^2 + \hat{\sigma}_{\hat{\tau}}^2 - 2\hat{\phi}_a}{\tilde{\rho}(a)^2} + 4\frac{\hat{\phi}_a - \hat{\sigma}_{\hat{\tau}}^2}{\tilde{\rho}(a)} + 4\hat{\sigma}_{\hat{\tau}}^2 \right).$$

Finally, we conduct a Monte Carlo experiment to show how the different methods perform. As an illustration, we set $\rho(a) = 0.35$, $\rho(a') = 0.65$, $\tau = 0.25$, and $n = 1,000$. We generate 3,000 datasets, using 1,000 draws for both the bootstrapping and simulation methods. Figure A1 gives our results for the AMCE (left panel) and marginal mean (right panel), with the true standard error (the standard deviation across the 3,000 estimates of $\tilde{\theta}(a,a')$ and $\tilde{\rho}(a)$) given in vertical dashed lines. We then compute standard errors from each of the 3,000 datasets with each of the three methods and present them in different colored histograms in the Figure. As is apparent, the three histograms are almost exactly the same for all three methods, and all are centered at the true value. This suggests that users can easily choose among the methods based on their preference for speed (analytical), convenience (bootstrapping), or familiarity (simulation).

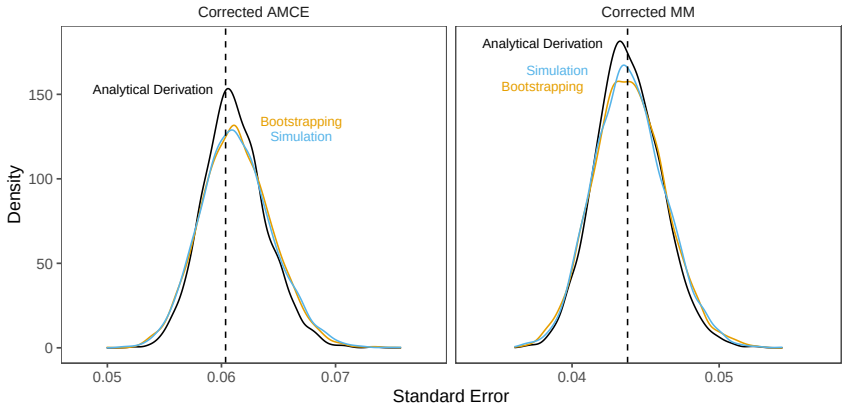

Figure A1: Standard Errors. Histograms from a Monte Carlo experiment of 3,000 standard error estimates (for AMCE in the left panel and MM in the right panel) from bootstrapping (gold), simulation (blue), and our analytical derivation (black). The true standard error is portrayed as a vertical dashed line in each figure.

A5 Study Selection

In this section we provide details on how we selected studies to replicate; it supplements the information in Section 4.

Our first studies that investigate intra-respondent reliability (IRR) in conjoint analysis via replications did not randomize the attributes and levels shown to respondents in replicated conjoint tables and instead focused on developing more controlled experiments. We searched for conjoint studies in political science and other social science domains that included (1) an example or screenshot of a conjoint table presented to respondents and (2) information on the introductory prompt and outcome question wording for the study. We restricted our search to studies that showed a pair of conjoint tables in a tabular format (excluding vignette-based designs or single profiles) and a forced-choice binary outcome question, the most commonly used conjoint design. We conducted this initial search in late 2018 using Google Scholar (starting with articles that cited Hainmueller, Hopkins, and Yamamoto (2014) or Hainmueller and Hopkins (2015)) and found 12 studies that had been published at that time that met our criteria: screenshot available, introductory prompt and outcome question wording available, paired tabular format, and forced binary choice outcome. The list of studies included: Atkeson and Hamel, 2020, Blackman, 2018, Bernauer and Gampfer, 2015, Hainmueller and Hopkins, 2015, Hankinson, 2018, Kertzer,

Renshon, Yarhi-Milo, et al., 2019, Ono and Burden, 2019, Mummolo, 2016, Mummolo and Nall, 2017, Leeper and Robison, 2020, Sances, 2018, and Schachter, 2016. We created standardized versions of the 12 example screenshots and conducted studies that asked respondents to make choices among all 12 at Time 1, asked them to make the same choices one week later, and calculated IRR between waves for each study (see the 12 replications v1.1, v1.2, and v2 in Table A11 for more details on these studies).

As we continued our analyses, we moved to fully replicate existing conjoint studies by randomizing all of the attributes and levels for a given study across respondents (see Section 4 in the main text). To select studies for these replications, we began with our initial list of 12 conjoint experiments but omitted studies with design choices that diverged from the standard fully randomized conjoint experiment, such as by including weighted probabilities for random assignment (Leeper and Robison, 2020), displaying randomly selected subsets of attributes across respondents (Kertzer, Renshon, Yarhi-Milo, et al., 2019), surveying non-representative samples (Sances, 2018), or incorporating complex cross-attribute constraints (Schachter, 2016).[2]

In October 2021, we then went back to Google Scholar and searched for more studies that cite Hainmueller, Hopkins, and Yamamoto (2014) or Hainmueller and Hopkins (2015) and that presented the conjoint in a tabular format and included a paired design and a forced-choice binary outcome variable. We also gave preference to studies that were published in top political science or general science journals (*American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *PNAS*, *Science*, and *Nature*), so we replaced two studies with those on similar topics published in this list of journals (Bernauer and Gampfer, 2015; Teele, Kalla, and Rosenbluth, 2018). We ultimately used Mummolo, 2016 for another replication study (see Section 4.4.2 in the main text) given its small number of potential attribute-level combinations, so we omitted it from this set of replications. This process resulted in a set of eight studies that reflect a variety of substantive topics (e.g., choices between housing developments, climate agreements, political candidates, immigrants, etc.): Arias and Blair (2022), Bechtel and

[2]Hainmueller and Hopkins, 2015 is an exception—this study does include cross-attribute constraints, but we felt that it was important to include it given its prominence in the conjoint literature. We implemented these cross-attribute constraints in our replication.

Scheve (2013), Blackman (2018), Hainmueller and Hopkins (2015), Hankinson (2018), Mummolo and Nall (2017), Teele, Kalla, and Rosenbluth (2018), and Ono and Burden (2019).

Table A1: Conjoint studies we replicate

Authors	Year	Title	Journal	Topic	Sample and provider	Respondents	Tasks	Attributes
Hainmueller & Hopkins	2014	The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants	AJPS	Immigrants	U.S. voters; KN	1407	5	9
Hankinson	2018	When Do Renters Behave Like Homeowners? High Rent, Price Anxiety, and NIMBYism	APSR	Housing	U.S. adults; GfK	3019	1	7
Teele, Kalla, & Rosenbluth	2018	The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics	APSR	Candidate gender	US local officials and voters; GfK	5088	3	6
Bechtel & Scheve	2013	Mass support for global climate agreements depends on institutional design	PNAS	Climate policy	French, German, U.K., U.S. adults; YouGov	4500	4	6
Ono & Burden	2018	The Contingent Effects of Candidate Sex on Voter Choice	Political Behavior	Candidate gender	U.S. adults; SSI	1583	10	13
Blackman	2018	Religion and Foreign Aid	Religion and Politics	Foreign aid	U.S. adults; SSI, Qualtrics	2810	3	7
Mummolo & Nall	2017	Why Partisans Do Not Sort: The Constraints on Partisan Segregation	JOP	Residential preferences	U.S. partisans; SSI	4800	5	7
Arias & Blair	2022	Changing Tides: Public Attitudes on Climate Migration	JOP	Migrants	German & U.S. adults; Dynata	2160	9	7

A6 Attentiveness and Intra-Respondent Reliability

In Section 3.1 we discuss the *origins* of IRR. In this section, we discuss a common hypothesis for *variation* in IRR: respondent attentiveness. Attentiveness varies widely across online samples; respondents who do not consistently pay close attention to conjoint tasks may be less likely to select the same profile at two different points in time than respondents who pay greater attention. Below we can evaluate this possibility by comparing IRR among attentive and inattentive respondents in our sample.

We used a variety of different sample providers and response quality indicators in all of the studies reported in Table A11.[3] In some studies, we included an attention check (or multiple attention checks) prior to the conjoint task and screened out respondents who failed. In other studies, we did not screen out respondents who failed the attention check but conducted our primary analyses among those who passed.[4] In studies conducted on respondents from DLABSS (see `dlabss.harvard.edu`), we do not include a pre-task attention check. DLABSS is a volunteer sample, and there is evidence that its subject pool is particularly attentive (Strange et al., 2019).Finally, in most of our studies, we included a post-treatment response quality check that asks respondents how often they provide humorous or insincere responses to survey questions.

In Table A2, we examine IRR by respondent attentiveness to investigate whether low IRR is unique to inattentive respondents. We find that IRR is higher for more attentive and/or more sincere samples but still around 75-80%, which is about halfway between random chance and perfect correspondence between choices at two points in time.

For more information on each of these studies, see Table A11. The wording of the attention check questions we employed is included below:

- Instructive attention check (v1): Next, we will provide you with several pieces of information about hypothetical students applying for admission to a university. Please

[3] Our first small pilots, "12 replications v1.1 and v1.2" in Table A11, are omitted.

[4] This approach is recommended by Prolific, which does not require researchers to pay respondents who fail the attention check, but allows them to complete the survey. Lucid, by contrast, recommends screening out respondents as soon as they fail the attention check. Employing attention checks on Lucid was particularly important given evidence that the quality of responses on Lucid has declined for a time during the COVID-19 pandemic (e.g., Peyton, Huber, and Coppock, 2022; Ternovski and Orr, 2022), and that this problem can be mitigated when attention checks are deployed.

indicate which of the two individuals you would personally prefer to be admitted as an undergraduate student at a university. But we would actually like to know if people are paying attention to the questions. Please ignore the second sentence on this screen and the question given on the next screen. Do not select either option and simply click "Next." *Candidate choice table presented.* (Applicant 1 / Applicant 2)

- Instructive attention check (v2): Please choose "somewhat agree" for this question. (Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree)

- True/false attention check: True or false? The letter "M" comes before the letter "L." (True / False / Neither)

- Checkbox attention check: We would like to get a sense of your consumption of political news. [paragraph break] To demonstrate that you've read this much, just go ahead and select both every day and never among the options below, no matter how often you watch political news. [paragraph break] Based on the text you read above, how often do you watch political news on TV? (Every day / Once a week / Once a month/Once a year/Never)

- Associational attention check: "Build" is most associated with... (Assemble / Commander / Find / Understand / Right)

- "Sincere" post-task quality check: We sometimes find people don't always take surveys seriously, instead providing humorous or insincere responses to questions. How often do you provide humorous or insincere responses to survey questions? (Never/Rarely/Some of the time/Most of the time/Always). *Note: Respondents who said that they "never" provide humorous or insincere responses to survey questions are coded as "sincere." All other respondents are coded as "insincere."*

Table A2: IRR by Attentiveness

Study	Study description	Sample provider	Type of IRR	Sub-sample	IRR	Sample N
1	12 replications v2	Turk Prime	Between Wave	All	0.76	205
1	12 replications v2	Turk Prime	Between Wave	Attentive only	0.81	129
1	12 replications v2	Turk Prime	Between Wave	Inattentive only	0.69	76
2	Consistency, complexity, divergence	Lucid Marketplace	Between Wave	All	0.69	474
2	Consistency, complexity, divergence	Lucid Marketplace	Between Wave	Insincere only	0.63	172
2	Consistency, complexity, divergence	Lucid Marketplace	Between Wave	Sincere only	0.72	302
3	Simplest case consistency	Lucid Marketplace	Between Wave	All	0.79	100
3	Simplest case consistency	Lucid Marketplace	Between Wave	Insincere only	0.75	30
3	Simplest case consistency	Lucid Marketplace	Between Wave	Sincere only	0.81	70
4	Consistency, policy only	Lucid Marketplace	Between Wave	All	0.75	594
4	Consistency, policy only	Lucid Marketplace	Between Wave	Insincere only	0.67	243
4	Consistency, policy only	Lucid Marketplace	Between Wave	Sincere only	0.79	351
5	Respondent characteristics vs. profile-pair combos v1	DLABSS	Between Wave	All	0.79	361
5	Respondent characteristics vs. profile-pair combos v1	DLABSS	Between Wave	Insincere only	0.77	80
5	Respondent characteristics vs. profile-pair combos v1	DLABSS	Between Wave	Sincere only	0.80	281
5	Respondent characteristics vs. profile-pair combos v1	DLABSS	Within Wave	All	0.88	885
5	Respondent characteristics vs. profile-pair combos v1	DLABSS	Within Wave	Insincere only	0.87	189
5	Respondent characteristics vs. profile-pair combos v1	DLABSS	Within Wave	Sincere only	0.88	696
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 1	All	0.80	503
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 1	Attentive only	0.81	478
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 1	Inattentive only	0.72	25
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 1	Insincere only	0.73	63
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 1	Sincere only	0.81	440
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 2	All	0.79	503
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 2	Attentive only	0.79	478
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 2	Inattentive only	0.84	25
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 2	Insincere only	0.68	63
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 2	Sincere only	0.80	440
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 3	All	0.80	503
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 3	Attentive only	0.80	478
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 3	Inattentive only	0.84	25
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 3	Insincere only	0.67	63
6	Is IRR worse in conjoints?	Prolific	Within Wave, Task 3	Sincere only	0.82	440
7	Respondent characteristics vs. profile-pair combos v2	Prolific	Between Wave	All	0.55	589
7	Respondent characteristics vs. profile-pair combos v2	Prolific	Between Wave	Insincere only	0.53	69
7	Respondent characteristics vs. profile-pair combos v2	Prolific	Between Wave	Sincere only	0.55	520
7	Respondent characteristics vs. profile-pair combos v2	Prolific	Within Wave	All	0.61	1027
7	Respondent characteristics vs. profile-pair combos v2	Prolific	Within Wave	Attentive only	0.61	896
7	Respondent characteristics vs. profile-pair combos v2	Prolific	Within Wave	Inattentive only	0.57	131
7	Respondent characteristics vs. profile-pair combos v2	Prolific	Within Wave	Insincere only	0.57	118
7	Respondent characteristics vs. profile-pair combos v2	Prolific	Within Wave	Sincere only	0.61	909
8	Do powerful attributes reduce error?	Prolific	Between Wave	All	0.67	97
8	Do powerful attributes reduce error?	Prolific	Between Wave	Insincere only	0.69	9
8	Do powerful attributes reduce error?	Prolific	Between Wave	Sincere only	0.67	88
8	Do powerful attributes reduce error?	Prolific	Within Wave	All	0.62	431
8	Do powerful attributes reduce error?	Prolific	Within Wave	Attentive only	0.61	288
8	Do powerful attributes reduce error?	Prolific	Within Wave	Inattentive only	0.66	143
8	Do powerful attributes reduce error?	Prolific	Within Wave	Insincere only	0.56	51
8	Do powerful attributes reduce error?	Prolific	Within Wave	Sincere only	0.63	380
9	8 replications v3	Lucid Theorem	Within Wave	All	0.72	3287
9	8 replications v3	Lucid Theorem	Within Wave	Insincere only	0.56	470
9	8 replications v3	Lucid Theorem	Within Wave	Sincere only	0.63	2817
10	Systematic ICR	Lucid Theorem	Within Wave	All	0.60	845
10	Systematic ICR	Lucid Theorem	Within Wave	Insincere only	0.56	290
10	Systematic ICR	Lucid Theorem	Within Wave	Sincere only	0.63	555
11	Open-ended comments	Lucid Theorem	Within Wave	All	0.81	134
11	Open-ended comments	Lucid Theorem	Within Wave	Insincere only	0.71	17
11	Open-ended comments	Lucid Theorem	Within Wave	Sincere only	0.83	117

A7 Profile Order Flipping for Repeated Conjoint Table

When researchers use our repeated-task approach to obtain an estimate of the IRR, we recommend flipping the order of the profiles that appear in the repeated task (i.e., the attribute-levels for a given profile would appear on the left in the first task and on the right in the repeated task, and vice versa). We recommend this to avoid three possible outcomes of including a repeated task: first, the possibility some respondents will select the same profile simply because they remember the pair of profiles that they saw in the first task and reflexively choose the same answer without carefully paying attention to the attributes and levels in the repeated task. Second, respondents have some inherent preference for profiles that appear on the left vs. right or for the profile labeled "A" or "B" or "1" or "2." Finally, we wish to avoid the possibility that respondents remember seeing the same task and then complain, thinking there was something wrong with the survey (a situation we ultimately never ran into).

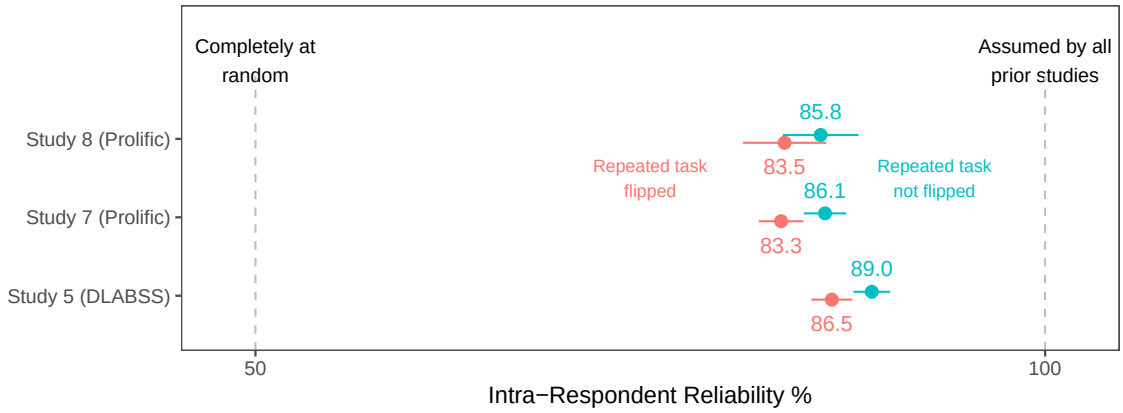

Figure A2: Relationship between IRR and whether repeated task profile order was flipped

Nevertheless, three of our studies allow us to examine whether our estimate of IRR varies by whether the order of profiles in the repeated task was flipped or not. In those studies, we randomly assigned whether the repeated task had the same profile order as the first task or if the order was reversed for each respondent. We then computed the average IRR for same-order profiles and for flipped-order profiles. The results are presented in Figure A2. As shown, IRR is slightly higher on average when the order of profiles in the repeated task is not flipped vs. flipped, but the differences across all three studies

are substantively very small and are unlikely to change the substantive meaning of our bias-corrected estimates.

A8 Traditional Survey Questions vs. Conjoint

This Suppemental Appendix elaborates upon the analyses from Section 4.3. We recruited 503 participants via Prolific to participate in a multi-format survey. We presented participants with a conjoint experiment and asked them to select between three pairs of candidates with randomly assigned policy positions, as well as a series of traditional multiple-choice survey questions on the same topic. Whether respondents saw the conjoint profiles or the traditional survey questions first was randomized, and these two survey modules were always separated by a series of unrelated questions. The attributes and levels for the conjoint experiment, as well as the wording of the question prompts for the traditional survey questions, are included below. Each level in the conjoint had identical wording to each answer choice in the traditional survey question.

- Attributes and levels:
  - Partisanship: Democrat / Republican
  - Position on abortion: By law, abortion should never be permitted. / The law should permit abortion only in case of rape, incest, or when the woman's life is in danger. / The law should permit abortion for reasons other than rape, incest, or danger to the woman's life, but only after the need for the abortion has been clearly established. / By law, a woman should always be able to obtain an abortion as a matter of personal choice.
  - Position on immigration: The number of immigrants from foreign countries who are permitted to come to the United States to live should be increased a lot. / The number of immigrants from foreign countries who are permitted to come to the United States to live should be increased a little. / The number of immigrants from foreign countries who are permitted to come to the United States to live should be left the same as it is now. / The number of immi-

grants from foreign countries who are permitted to come to the United States to live should be decreased a little. / The number of immigrants from foreign countries who are permitted to come to the United States to live should be decreased a lot.

  - Position on economy: We need a strong government to handle today's complex economic problems. / The free market can handle these problems without the government being involved.
  - Position on affirmative action: Preference in hiring and promotion of Black people is wrong because it gives Black people advantages they haven't earned. / Because of past discrimination, Black people should be given preference in hiring and promotion.

- Question prompts for standard questions with binary outcomes:
  - Vote: If you had to choose between them, would you vote for a Democrat or a Republican in a congressional election?
  - Position on economy: Which of the following two statements comes closer to your own opinion?
  - Position on affirmative action: Which of the following two statements comes closer to your own opinion?

Note that the analysis reported in Figure 3 in the main text focuses only on the traditional survey questions with binary outcomes (i.e., two levels in the conjoint or two answer choices in the traditional questions). We do not report intra-respondent reliability for standard-format survey questions with multiple options, as it is not directly comparable to conjoint-style outcome questions that present respondents with a forced choice between two alternatives.

A9 The Top-Down Approach

Here we expand on Section 4.4.1. We study whether intra-respondent reliability (IRR) can be predicted by the potentially high cognitive load generated by our hypotheses of

inconsistency, *complexity*, and *divergence*. We ran an unusually large number of studies to understand this question, in part because it is difficult to provide evidence for a negative and in part because we refined our hypotheses along the way. What follows is details about a sequence of studies we conducted (and corresponding tables or figures with results). All of the data and code necessary to reproduce our results for every study is available in our replication dataset.

We begin by recruiting 474 respondents through Lucid Marketplace to evaluate 15 conjoint tasks each, where five of the evaluation tasks had different levels of consistency, five varied in complexity, and five had different levels of divergence. We ask respondents to complete the same task again a week later (with tasks presented in random order), and we record the IRR. In the consistency conjoint tasks, we adapted the design used in Ono and Burden (2019) and asked respondents to evaluate and select one of two hypothetical candidates running for the U.S. House of Representatives. We varied the level of logical coherence across candidate partisanship and policy positions, such that the most consistent set of profiles presented to respondents might look like Table A3, whereas the least consistent would look the same, except that the party labels would be flipped so that they were inconsistent with the policy position attributes. Profiles in between would be scored from most to least consistent based on the number of available policy positions consistent with each candidate's party label.

Table A3: High Consistency Conjoint Profile

	Candidate A	Candidate B
Party	Democrat	Republican
Position on National Security	Wants to cut military budget and keep the U.S. out of war	Wants to maintain strong defense and increase U.S. influence
Position on Immigrants	Favors giving citizenship or guest worker status to undocumented immigrants	Opposes giving citizenship or guest worker status to undocumented immigrants
Position on Abortion	Abortion is a private matter (pro-choice)	Abortion is not a private matter (pro-life)
Position on Government Deficit	Wants to reduce the deficit through tax increase	Wants to reduce the deficit through spending cuts

Adapting a design from Kertzer, Renshon, Yarhi-Milo, et al. (2019), the complexity questions ask respondents to predict which country would be more likely to stand firm rather than concede in a territorial dispute between two hypothetical countries. The relevant attributes for each country were then described to respondents in increasingly complex terms with longer sentences, where the simplest presentation would look like Table A4 and the most complex would look like Table A5.

Table A4: Least Complex Conjoint: Territorial Dispute

	Country A	Country B
Interests in the Dispute	High stakes	Low stakes
Leader Gender	Woman	Man
Previous Behavior in International Disputes	Forceful	Peaceful
Current Behavior	No action	Issuing threats
Leader Background	Civilian	Ex-military
Military Capabilities	Powerful	Weak

We tested divergence by adapting a version of Hankinson (2018), in which respondents reviewed two proposed developments that might be built in their city or town. Table A6 depicts a sample of the most (and least) divergent housing developments respondents were asked to evaluate. Values in parentheses depict the *least* divergent profiles.

Figure A3 summarizes our results for consistency, complexity, and divergence in the three panels, respectively. Integers on the x-axes of each panel correspond to profile choice characteristics, with 1 referring to the least consistent, complex, or diverse profile and 5 indicating the most. The y-axis depicts the proportion of participating respondents who chose the same candidate, country, or housing development given the same profile choices a week after they saw the profiles for the first time (IRR). If one of these conjoint design attributes drove IRR, we would expect to see point estimates trending linearly: IRR would trend upward from left to right as profiles got more consistent and downward from left to right as profiles got more complex and less divergent. Clearly, the second and third panels reveal no upward linear trend. However, the results in the first panel do show a slight upward trend for consistency, but the effect is small and substantively trivial and cannot account for the vast majority of observed IRR: Average IRR among

Table A5: Most Complex Conjoint: Territorial Dispute

	Country A	Country B
Interests in the Dispute	Experts in foreign relations have described the country's stakes in the dispute as relatively high.	Experts in foreign relations have described the country's stakes in the dispute as relatively low.
Leader Gender	The leader of the country involved in the international dispute is a man.	The leader of the country involved in the international dispute is a woman.
Previous Behavior in International Disputes	The last time this country was involved in an international dispute, it initiated the crisis by issuing a public threat to use force against an adversary of the United States.	The last time this country was involved in an international dispute, it was challenged by an ally of the United States and ultimately mobilized troops in response to the challenge.
Current Behavior	In the current crisis, the country has yet to make any statements or carry out any actions.	In the current crisis, the country has made a public threat that they will use force if the other country does not back down.
Leader Background	The country's leader recently took office, and served in the military briefly before assuming power.	The country's leader has been in power for many years, and does not have experience in the military.
Military Capabilities	The country has a powerful military with a large number of troops that it is currently prepared to deploy.	The country has a not very powerful military with a small number of troops that it is currently prepared to deploy.

the most consistent profiles is just six percentage points greater than IRR among the least consistent, and the 95% confidence intervals overlap across the range of profiles.

We then pursued the small consistency finding by designing an even more extreme experiment that was intended to help us understand how far as we could take this result. We still find that consistency does not drive IRR noticeably, even in a clearly extreme case in which respondents are making a forced choice between two candidates based on just two attributes: party affiliation and one policy position (in our case, tax policy). To do this, we recruited 100 respondents via Lucid Marketplace to participate in an abbreviated form of

Table A6: Examples of Divergent (and Non-Divergent) Profiles in Housing Units

	Building A	Building B
How many units will the building have?	12 (10) units	96 (12) units
How many units will be available to rent?	6 (4) units	80 (6) units
What share of units will be affordable for low-income residents?	All (One quarter) of the units	None (Half) of the units
How far is the building from your home?	1/2 mile - 10 minute (1/4 mile - 5 minute) walk	2 miles - 40 minute (1/8 mile - 2 minute) walk
How tall will the building be?	3 (4) stories	12 (3) stories
How much will it cost to build the building?	$3 (7) million	$20 (6) million

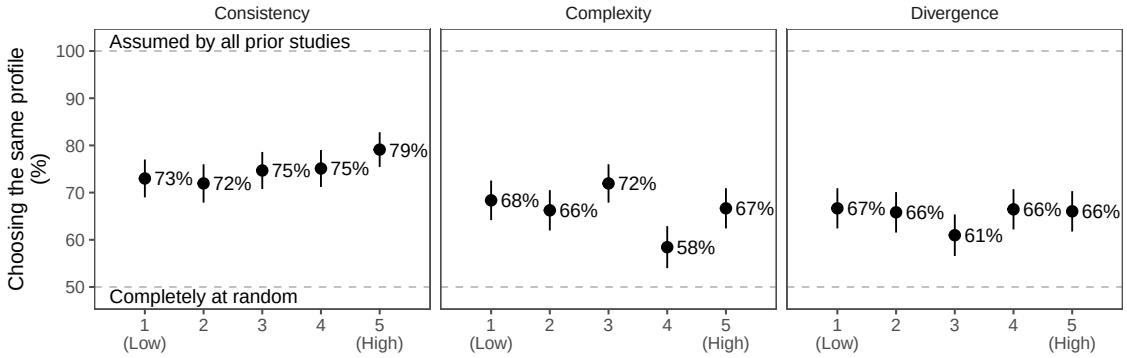

Figure A3: Relationship between IRR and profile consistency, complexity, or divergence.

the experiment depicted in Table A3. Respondents chose between a hypothetical Republican or Democrat who either "favors raising taxes on the wealthy" or "favors lowering taxes on everyone, including the wealthy." Respondents were randomly assigned to either consistent (Democrat, raise taxes; Republican, lower taxes) or inconsistent (Democrat, lower taxes; Republican, raise taxes) comparisons and subsequently asked to review the same match-ups one week later. This, of course, is not a substantively reasonable conjoint design as we would not normally see this type of variation in actual elections in the US, but we use it to pressure test the consistency idea.

Figure A4 summarizes these results. IRR was 78% for respondents evaluating inconsistent profiles and 80% for respondents evaluating consistent ones, with overlapping confidence intervals for both groups. Taken together, we conclude that these two studies

suggest that measurement error associated with observing choice in conjoint experiments is not related to the extent of consistency or coherence in the profiles respondents are being asked to evaluate.

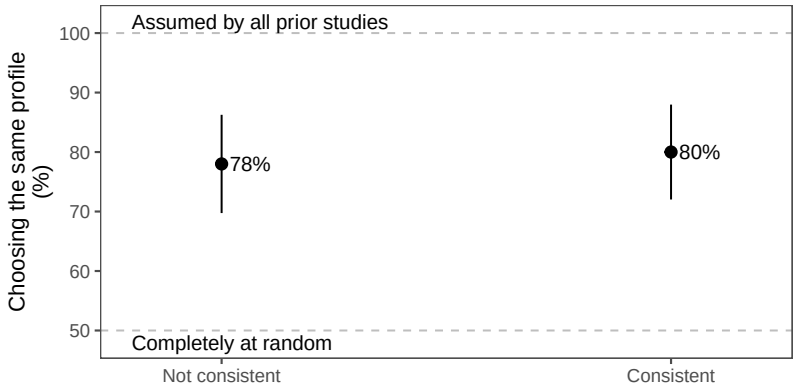

Figure A4: Relationship between IRR and profile consistency: A simple case with party affiliation and tax policy views.

We then go further and explore the possibility that consistency had such a small effect on IRR because candidate partisanship was so *dominant* an attribute that respondents used it to guide their selections, without regard to the policy positions associated with each candidate. This may well be a concern for researchers who study candidate choice in American politics, where party identification is a powerful heuristic for uninformed voters (Popkin, 1991; Rahn, 1993) and where high levels of antipathy towards members (and candidates) belonging to an out-party define the contemporary political landscape (Abramowitz and Webster, 2016). Indeed, several studies have demonstrated that respondents evaluate candidates differently when information on party affiliation is absent (Crowder-Meyer, Gadarian, and Trounstine, 2020; Kirkland and Coppock, 2018; McConnaughy et al., 2010).

This issue may be particularly significant for researchers studying domains in which any one attribute might dominate choice. For instance, in the context of high fuel prices, fuel efficiency might overwhelm drivers' choices of vehicles, even if individuals have genuine systematic preferences over other features. This limits the impact that logically inconsistent profiles might have on IRR but also limits researchers' ability to learn meaningful information about average preferences for, or causal impacts of, other features.

We thus tested the possibility that profile consistency might affect IRR considerably more without a partisanship attribute in a separate study. We recruited 599 participants for a two-wave study via Lucid Marketplace. In this version of the experiment, respondents chose between two candidates with different policy positions on health care, government spending priorities, affirmative action, and taxes. Respondents were not presented with either candidate's party affiliation. Respondents viewing the most consistent set of candidate profiles might have seen a conjoint table like Table A7, while the least consistent versions would have appeared to respondents following Table A8.

Table A7: High Consistency Nonpartisan Conjoint Profile

	Candidate A	Candidate B
Health Care	Supports government-funded health care system	Supports private health care system
Government Spending	Increase funding for renewable energy research	Increase funding for national security
Affirmative Action	College admissions decisions should take race into account	College admissions decisions should be based on merit only
Taxes	Raise taxes on the wealthy	Lower taxes on everyone, including the wealthy

Table A8: Low Consistency Nonpartisan Conjoint Profile

	Candidate A	Candidate B
Health Care	Supports government-funded health care system	Supports private health care system
Government Spending	Increase funding for national security	Increase funding for renewable energy research
Affirmative Action	College admissions decisions should be based on merit only	College admissions decisions should take race into account
Taxes	Raise taxes on the wealthy	Lower taxes on everyone, including the wealthy

The results show that logical consistency only seems to influence IRR in the most extreme and unrealistic cases. Figure A5 summarizes these results. There are four attributes (policy positions) and no listed party affiliations, so respondents are asked to choose between candidates in two separate tasks. In one task, both candidates have logically consistent profiles (all four policy positions are cohesively liberal or conservative).

In the other task, candidates have logically inconsistent profiles (exactly two policy positions are traditionally liberal, and the other two are traditionally conservative). The order in which candidates appeared to respondents and the specific policy positions that flip to produce the inconsistent profile shown to respondents were all randomly assigned. Respondents were asked to review the exact same profiles in a subsequent wave one week later. Figure A5 summarizes our results, broken out by three separate panels according to which attributes were flipped to generate inconsistent profiles (left: government spending and affirmative action, center: health care and affirmative action, right: health care and government spending).

In this study, the gaps between the least consistent and the fully consistent profiles are indeed larger than they are for the studies summarized in Figures A3 and A4. Overall, going from the inconsistent to fully consistent profiles across all policy issues increases IRR by 0.1 on a scale from 0 (no respondent agrees with herself a week later) to 1 (all respondents choose the same profiles in wave 2). This suggests that the party affiliation attribute may be a dominant heuristic that respondents use to simplify their choices when other attributes are inconsistent with party or each other, or are otherwise difficult to assess. However, note that this design is substantively unrealistic and extreme in that it includes unlikely bundles of policy positions. Given extremely high levels of partisan polarization (McCarty, Poole, and Rosenthal, 2007) in contemporary American politics, combined with the fact that candidates seeking election as challengers are likely to embrace the national party's ideology (Ansolabehere, Snyder, and Stewart, 2001), it is exceedingly rare to see candidates running on policy positions associated with an opposing party. Accordingly, most researchers who want to apply the conjoint design in an electoral context are unlikely to see much systematic error coming from profile inconsistency, especially if they constrain their randomization procedures to prevent the occurrence of extremely unlikely or impossible profiles.

We pushed this analysis to another extreme in yet another replication, asking whether a dominant attribute can systematically influence IRR in a conjoint with one overwhelmingly important attribute and a series of relatively inconsequential ones. We recruited

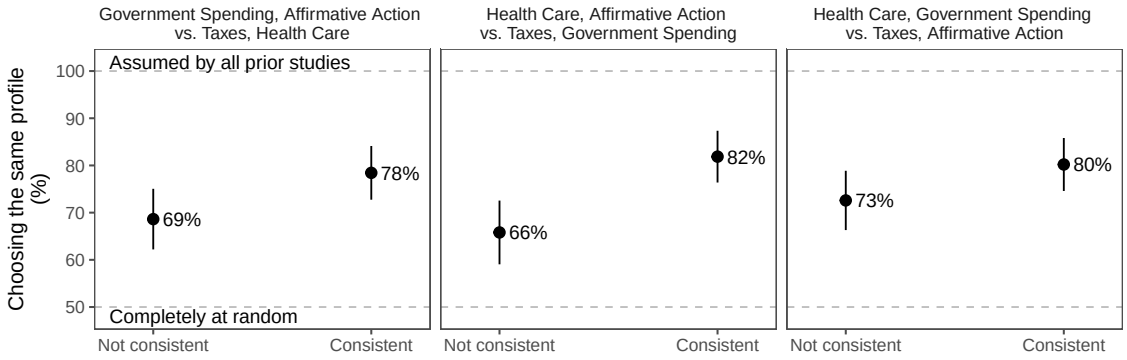

Figure A5: Relationship between IRR and profile consistency: No party labels. The leftmost panel represents the average % of respondents choosing the same profile twice when positions on government spending and affirmative action are flipped to generate inconsistent profiles; the central panel shows flipped health care and affirmative action positions; the rightmost panel shows health care and government spending positions flipped. Comparisons within each panel are to the same respondents evaluating the consistent profiles they were randomly assigned.

431 participants via Prolific and had them evaluate eight candidate choice tasks twice. The candidates that respondents could choose between were defined by their partisanship, age, race, gender, alma mater, and salient personal characteristics. Figure A6 summarizes the relationships between all possible pairs of candidate characteristics and IRR in this study. The baseline level of IRR was 88%. If a particular attribute drove IRR, we might expect within-respondent agreement to drop considerably when both candidates had the same levels of that attribute. If respondents rely most heavily on party heuristics to make decisions, for instance, the choices they make between pairs of Democrats might be the hardest and least consistent. Figure A6 shows the change from the baseline IRR when profile-pairs all possible pairs of combinations across attributes. An inability to discriminate between candidates along partisan lines does have a negative impact on IRR for respondents, though this is most pronounced within the same wave and almost disappears between waves. Otherwise, we find little evidence that having identical characteristics across other attributes moves IRR. In fact, most relationships between profiles with candidates with identical characteristics and IRR are positive (if not statistically distinguishable from zero). This suggests that it is possible to systematically affect IRR in conjoint experiments where respondents essentially load their decisions onto a single attribute, but this

approach is an unlikely one for researchers who utilize conjoint experiments precisely to learn how respondents make choices in the presence of a variety of important attributes.

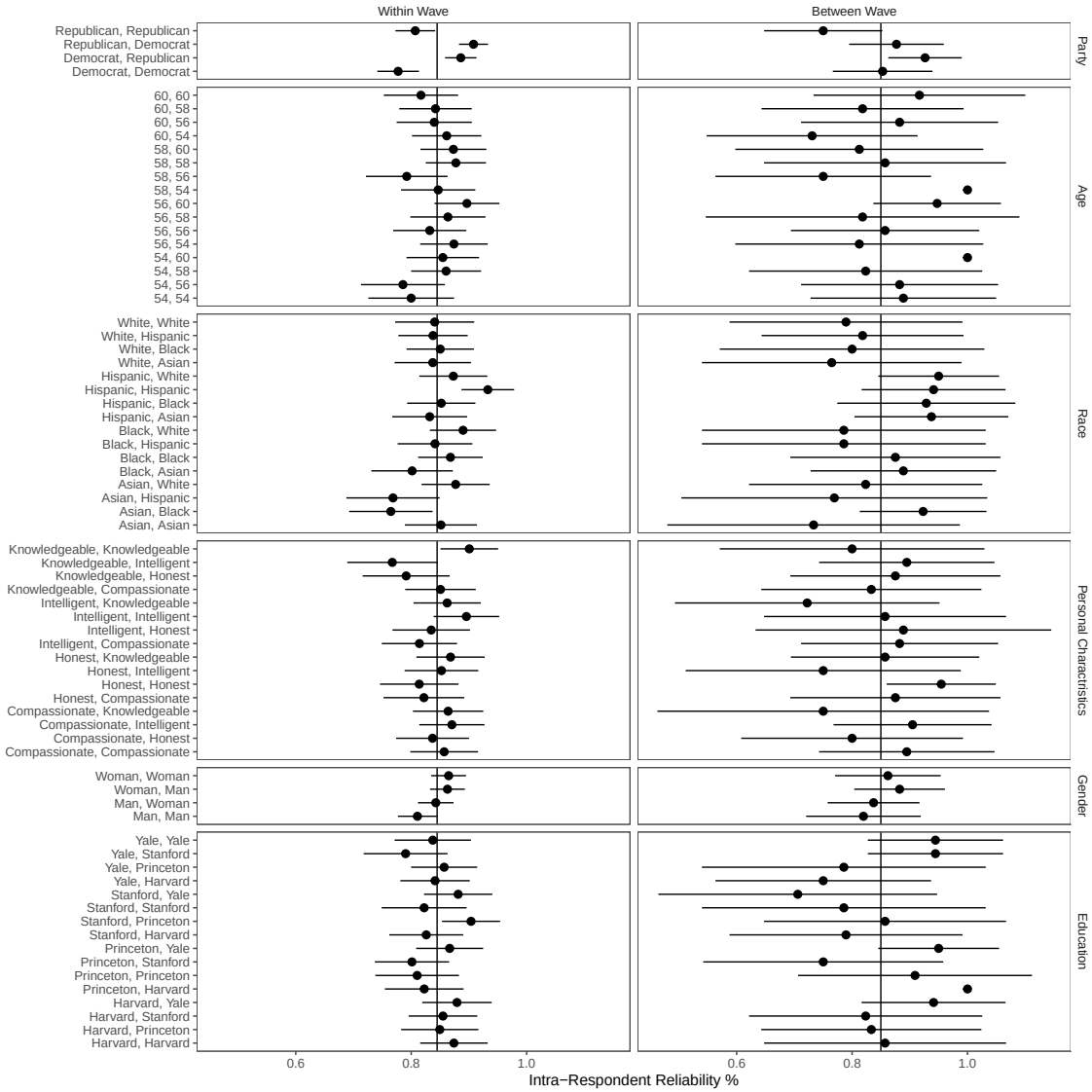

Figure A6: Relationship between IRR and profile-pair characteristics within wave (left) and between waves (right). Vertical lines represent the overall IRR for the replication (88%).

A10 The Bottom Up Approach

This section expands on the bottom-up approach of Section 4.4.2 with a separate set of data, with respondents from two sources. We recruited a sample of 335 respondents via the Harvard Digital Lab for the Social Sciences (DLABSS) and an additional sample of

611 respondents recruited via Prolific. As with the analyses in the text, we do not combine the two surveys in case they represent different populations with different sets of personal characteristics. We again adapted Hankinson (2018). Respondents were asked to choose between proposed housing developments with four possible attributes (distance from the respondent's home, the current land use to be replaced by the proposed development, the share of the units in the building that will be affordable to low-income residents, and total units in the proposed development) that had two to three possible levels each (see Table A6 for an example setup). Respondents saw eight pairs of conjoint evaluation tasks twice *within* the same experiment and then repeated the same tasks again two weeks later. We observed 621 different profile-pair combinations with these attributes and levels. This design allowed us to measure IRR across specific combinations of profile-pairs within respondents since each respondent evaluated a set of the same profile-pair combinations more than once, and it allowed us to do this both within and across waves of the same experiment.

Whereas the study described in Section 4.4.2 was fully nonparametric, we use robust least squares to analyze these two samples attempting to model IRR as a function of specific attribute combinations and personal characteristics. Our results suggest that both account for relatively little variation in IRR. Within the first wave of this study, the profile-pair combinations accounted for just 0.3.% of the variation in IRR, while respondent characteristics accounted for 8.4% of the variation in IRR. Across waves, profile-pair combinations accounted for 2.9% of the variation in IRR, while respondent characteristics accounted for 7.7%. Thus, the vast majority of variation in IRR, or 91.3% within wave 1 and 89.4% across waves, would seem to be attributable to random swapping error.

We enumerate the impact of each given possible pair of attribute level combinations that DLABSS respondents (Figure A7) and Prolific respondents (Figure A9) might have seen on IRR both within and between waves. These figures summarize estimates and confidence intervals from a robust OLS regression of a binary indicator for whether a specific respondent selected the same profile twice when faced with the same comparison on a series of factors representing possible combinations of attribute levels in the profiles that

might have appeared. In each case, the majority of possible attribute-level combinations that respondents might have seen appear to have no relationship to IRR. Figure A8 represents the correlations between estimated IRR for respondents in the DLABSS and Prolific studies, where points are specific combinations of attributes visible to respondents evaluating the same profile-pairs within wave (left) and between waves (right). Estimates for within-wave IRR across all possible attributes are tightly, and positively correlated across the two studies. Between-wave estimates of the relationships between particular attribute combinations and IRR across the two studies have more spread but are similarly positively correlated across attribute combinations.

We expand on this design using an additional replication, this time with a focus on limiting the number of possible profile-pair combinations in order to allow a sufficiently large number of respondents to evaluate each multiple times so as to provide enough power to assess the relationship between *every* possible profile-pair combination in an experiment and IRR. We replicated Mummolo (2016), as reported in the text. We recruited 2,641 participants to take part in the replication via Lucid Theorem. In our adaptation of this conjoint experiment, respondents chose between two articles with four possible headlines that could have come from three possible sources. A complete listing of possible attribute combinations appears in Table A9, which shows the estimates of the correlation between each profile-pair and IRR in both waves of the study along with standard errors and the numbers of respondents who evaluated each combination in each wave. Just 6% of the profile-pairs appear to have a significant relationship with IRR in Wave 1, and just 10.4% do in Wave 2, and just two of those profile-pair combinations have a significant relationship to IRR in both waves. This table provides the key to the left panel in Figure 4.

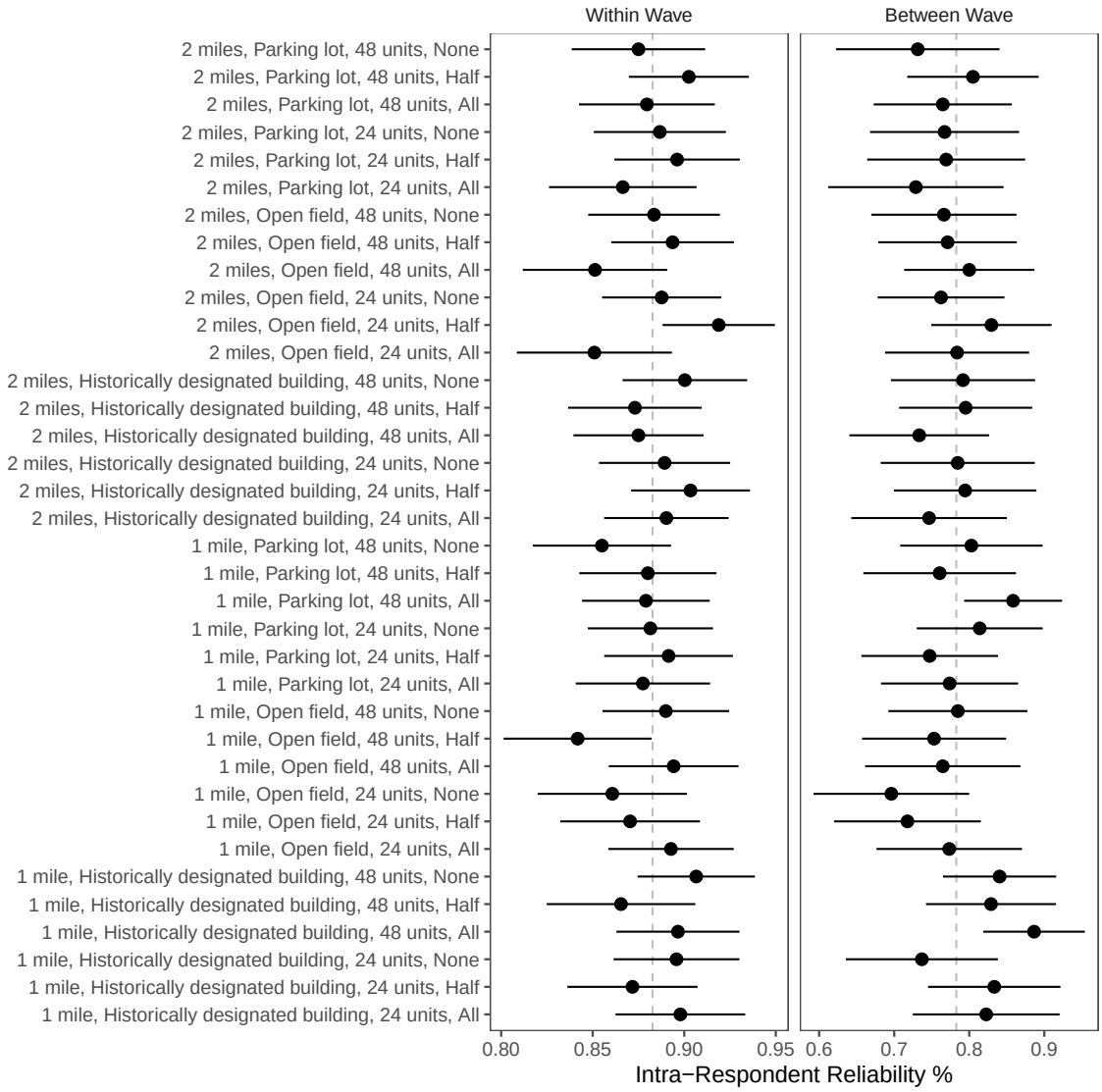

Figure A7: IRR for repeated tasks within-wave (left) and between wave (right) in a DLABSS replication of Hankinson (2018). Rows represent possible combinations of attributes visible to respondents in the study. Dashed vertical lines represent the overall mean IRR within each wave (left) and between waves (right).

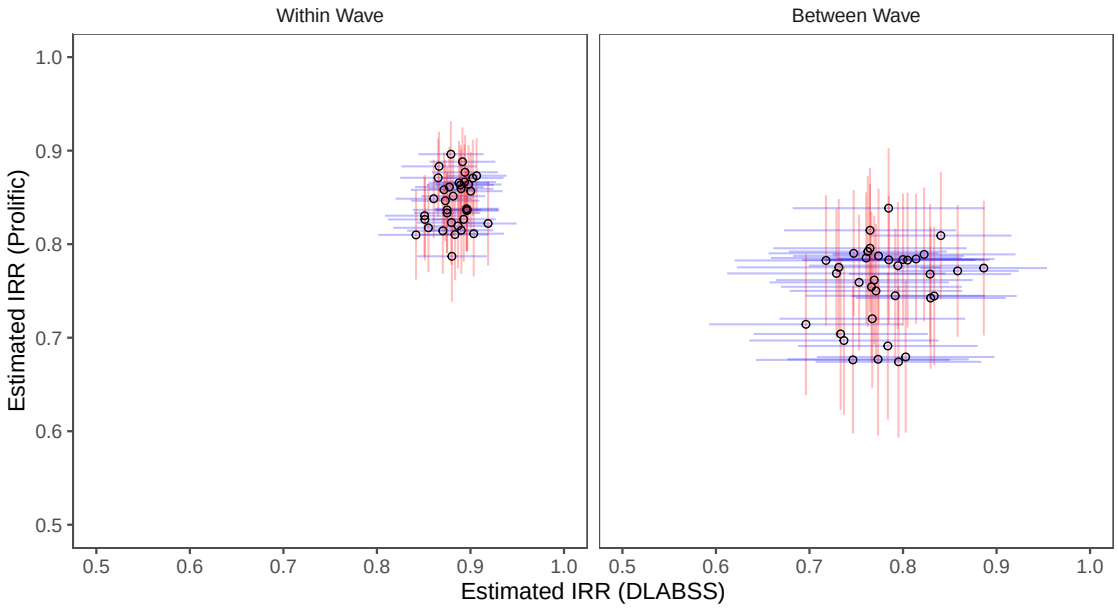

Figure A8: Correlations between estimated relationships between specific combinations of attributes in profile-pair comparisons evaluated by DLABSS and Prolific respondents within wave (left) and between waves (right). Blue segments represent confidence intervals associated with estimates from the DLABSS study, while red segments represent confidence intervals associated with estimates from the Prolific Study.

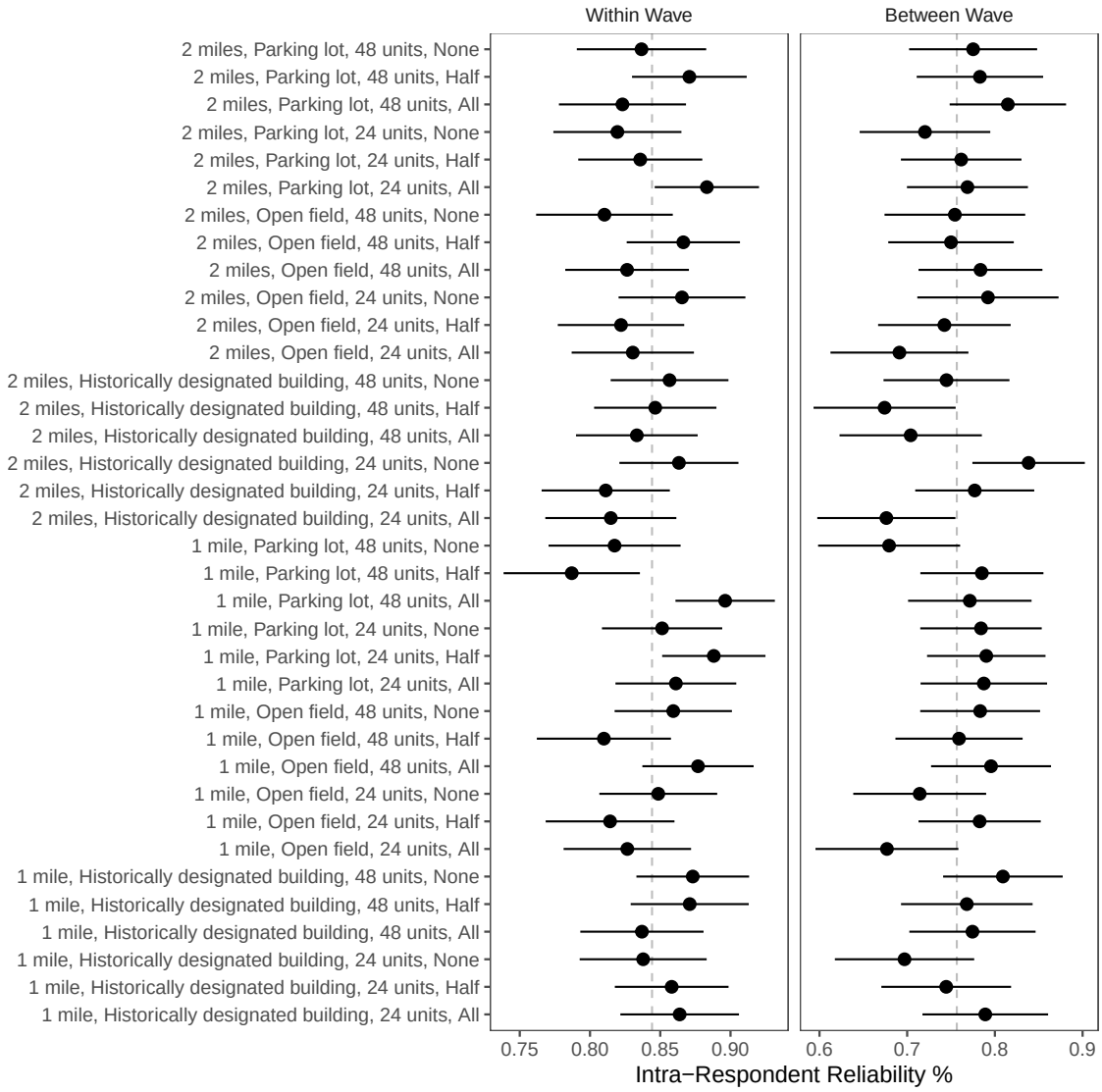

Figure A9: IRR for repeated tasks within-wave (left) and between wave (right) in a Prolific replication of Hankinson (2018). Rows represent possible combinations of attributes visible to respondents in the study. Dashed vertical lines represent the overall mean IRR within each wave (left) and between waves (right).

Table A9: All profile-pair Combinations from News Consumption Replication Experiment

	Headline 1	Headline 2	Source 1	Source 2	W1 Est.	W1 S.E.	W1 n	W2 Est.	W2 S.E.	W2 n
1	Celebrity dating fails	Celebrity dating fails	Fox News	MSNBC	0.73	0.06	63	0.69	0.05	94
2	Celebrity dating fails	Celebrity dating fails	Fox News	USA Today	0.81	0.05	57	0.78	0.04	106
3	Celebrity dating fails	Senate votes against bill that would ensure equal pay for women	Fox News	Fox News	0.75	0.05	63	0.85	0.03	112
4	Celebrity dating fails	Senate votes against bill that would ensure equal pay for women	Fox News	MSNBC	0.82	0.05	51	0.70	0.04	110
5	Celebrity dating fails	Senate votes against bill that would ensure equal pay for women	Fox News	USA Today	0.75	0.05	63	0.77	0.04	112
6	Celebrity dating fails	Smokers who quit may cut heart risk faster than had been thought, study finds	Fox News	Fox News	0.64	0.08	36	0.85	0.03	120
7	Celebrity dating fails	Smokers who quit may cut heart risk faster than had been thought, study finds	Fox News	MSNBC	0.78	0.05	63	0.78	0.04	112
8	Celebrity dating fails	Smokers who quit may cut heart risk faster than had been thought, study finds	Fox News	USA Today	0.70	0.06	63	0.72	0.05	94
9	Celebrity dating fails	Weight-loss tips that make a difference	Fox News	Fox News	0.65	0.06	65	0.72	0.05	81
10	Celebrity dating fails	Weight-loss tips that make a difference	Fox News	MSNBC	0.80	0.06	51	0.80	0.04	110
11	Celebrity dating fails	Weight-loss tips that make a difference	Fox News	USA Today	0.64	0.08	36	0.79	0.04	120
12	Celebrity dating fails	Celebrity dating fails	MSNBC	USA Today	0.77	0.05	65	0.65	0.05	81
13	Celebrity dating fails	Senate votes against bill that would ensure equal pay for women	MSNBC	MSNBC	0.78	0.05	63	0.74	0.04	94
14	Celebrity dating fails	Senate votes against bill that would ensure equal pay for women	MSNBC	USA Today	0.79	0.05	57	0.80	0.04	106
15	Celebrity dating fails	Smokers who quit may cut heart risk faster than had been thought, study finds	MSNBC	MSNBC	0.73	0.06	51	0.81	0.04	110
16	Celebrity dating fails	Smokers who quit may cut heart risk faster than had been thought, study finds	MSNBC	USA Today	0.84	0.05	57	0.85	0.03	106
17	Celebrity dating fails	Weight-loss tips that make a difference	MSNBC	MSNBC	0.76	0.06	59	0.79	0.04	107
18	Celebrity dating fails	Weight-loss tips that make a difference	MSNBC	USA Today	0.75	0.05	63	0.78	0.04	112
19	Celebrity dating fails	Senate votes against bill that would ensure equal pay for women	USA Today	USA Today	0.82	0.06	39	0.74	0.04	115
20	Celebrity dating fails	Smokers who quit may cut heart risk faster than had been thought, study finds	USA Today	USA Today	0.64	0.06	59	0.69	0.04	107
21	Celebrity dating fails	Weight-loss tips that make a difference	USA Today	USA Today	0.77	0.07	39	0.75	0.04	115
22	Senate votes against bill that would ensure equal pay for women	Senate votes against bill that would ensure equal pay for women	Fox News	MSNBC	0.86	0.05	57	0.83	0.04	106
23	Senate votes against bill that would ensure equal pay for women	Senate votes against bill that would ensure equal pay for women	Fox News	USA Today	0.70	0.06	63	0.65	0.05	112
24	Senate votes against bill that would ensure equal pay for women	Smokers who quit may cut heart risk faster than had been thought, study finds	Fox News	Fox News	0.78	0.05	59	0.67	0.05	107
25	Senate votes against bill that would ensure equal pay for women	Smokers who quit may cut heart risk faster than had been thought, study finds	Fox News	MSNBC	0.71	0.06	63	0.72	0.05	94
26	Senate votes against bill that would ensure equal pay for women	Smokers who quit may cut heart risk faster than had been thought, study finds	Fox News	USA Today	0.65	0.06	65	0.70	0.05	81
27	Senate votes against bill that would ensure equal pay for women	Weight-loss tips that make a difference	Fox News	Fox News	0.82	0.06	39	0.77	0.04	115
28	Senate votes against bill that would ensure equal pay for women	Weight-loss tips that make a difference	Fox News	MSNBC	0.69	0.06	59	0.69	0.04	107
29	Senate votes against bill that would ensure equal pay for women	Weight-loss tips that make a difference	Fox News	USA Today	0.89	0.04	57	0.75	0.04	106
30	Senate votes against bill that would ensure equal pay for women	Senate votes against bill that would ensure equal pay for women	MSNBC	USA Today	0.79	0.06	39	0.72	0.04	115
31	Senate votes against bill that would ensure equal pay for women	Smokers who quit may cut heart risk faster than had been thought, study finds	MSNBC	MSNBC	0.78	0.05	63	0.68	0.05	94
32	Senate votes against bill that would ensure equal pay for women	Smokers who quit may cut heart risk faster than had been thought, study finds	MSNBC	USA Today	0.67	0.07	51	0.79	0.04	110
33	Senate votes against bill that would ensure equal pay for women	Weight-loss tips that make a difference	MSNBC	MSNBC	0.68	0.06	59	0.67	0.05	107
34	Senate votes against bill that would ensure equal pay for women	Weight-loss tips that make a difference	MSNBC	USA Today	0.81	0.07	36	0.81	0.04	120
35	Senate votes against bill that would ensure equal pay for women	Smokers who quit may cut heart risk faster than had been thought, study finds	USA Today	USA Today	0.74	0.07	39	0.80	0.04	115
36	Senate votes against bill that would ensure equal pay for women	Weight-loss tips that make a difference	USA Today	USA Today	0.75	0.06	51	0.68	0.04	110
37	Smokers who quit may cut heart risk faster than had been thought, study finds	Smokers who quit may cut heart risk faster than had been thought, study finds	Fox News	MSNBC	0.75	0.06	59	0.70	0.04	107
38	Smokers who quit may cut heart risk faster than had been thought, study finds	Smokers who quit may cut heart risk faster than had been thought, study finds	Fox News	USA Today	0.69	0.06	65	0.68	0.05	81
39	Smokers who quit may cut heart risk faster than had been thought, study finds	Weight-loss tips that make a difference	Fox News	Fox News	0.78	0.05	65	0.72	0.05	81
40	Smokers who quit may cut heart risk faster than had been thought, study finds	Weight-loss tips that make a difference	Fox News	MSNBC	0.67	0.07	51	0.80	0.04	110
41	Smokers who quit may cut heart risk faster than had been thought, study finds	Weight-loss tips that make a difference	Fox News	USA Today	0.71	0.06	65	0.64	0.05	81
42	Smokers who quit may cut heart risk faster than had been thought, study finds	Smokers who quit may cut heart risk faster than had been thought, study finds	MSNBC	USA Today	0.75	0.07	36	0.70	0.04	120
43	Smokers who quit may cut heart risk faster than had been thought, study finds	Weight-loss tips that make a difference	MSNBC	MSNBC	0.73	0.06	63	0.70	0.04	112
44	Smokers who quit may cut heart risk faster than had been thought, study finds	Weight-loss tips that make a difference	MSNBC	USA Today	0.76	0.05	63	0.67	0.05	94
45	Smokers who quit may cut heart risk faster than had been thought, study finds	Weight-loss tips that make a difference	USA Today	USA Today	0.69	0.08	36	0.72	0.04	120
46	Weight-loss tips that make a difference	Weight-loss tips that make a difference	Fox News	MSNBC	0.81	0.05	57	0.77	0.04	106
47	Weight-loss tips that make a difference	Weight-loss tips that make a difference	Fox News	USA Today	0.69	0.08	36	0.69	0.04	120
48	Weight-loss tips that make a difference	Weight-loss tips that make a difference	MSNBC	USA Today	0.69	0.07	39	0.69	0.04	115

A11 Respondent Characteristics

Table A10 summarizes age, gender, race, and region of residence for the people who participated in our replication of Mummolo (2016). This table also provides a key to the right panel of Figure 4 in the main text, with numbers corresponding to the points in the plot.

Table A10: Respondent Characteristics and IRR in News Consumption Replication Experiment

	Respondent Characteristic	W1 Est.	W1 S.E.	W1 n	W2 Est.	W2 S.E.	W2 n
1	Age: 18-24	0.66	0.03	348	0.63	0.02	798
2	Age: 25-34	0.69	0.02	606	0.72	0.01	1302
3	Age: 35-44	0.70	0.02	726	0.73	0.01	1302
4	Age: 45-54	0.82	0.02	324	0.77	0.02	768
5	Age: 55+	0.86	0.01	594	0.87	0.01	900
6	Gender: Female	0.80	0.01	1272	0.76	0.01	3180
7	Gender: Male	0.69	0.01	1326	0.72	0.01	1890
8	Ethnicity: Hispanic	0.71	0.03	318	0.68	0.02	708
9	Ethnicity: Not Hispanic	0.75	0.01	2280	0.75	0.01	4362
10	Race: Black or African American	0.71	0.02	456	0.67	0.02	858
11	Race: Some other race	0.75	0.02	330	0.69	0.02	528
12	Race: White	0.75	0.01	1794	0.77	0.01	3660
13	Region: Midwest	0.75	0.02	444	0.73	0.01	930
14	Region: Northeast	0.76	0.02	522	0.77	0.01	1002
15	Region: South	0.74	0.01	1020	0.75	0.01	2292
16	Region: West	0.72	0.02	612	0.71	0.02	840

A12 Additional Studies on Measurement Error in Conjoint Studies

In this Appendix, we describe every conjoint experiment we conducted in the process of preparing this manuscript, in chronological order, including the preliminary studies that do not appear in the main text. All the survey data generated by these studies are available in our replication dataset. Table A11 lists all these studies, and it includes links to every study's pre-registration document when available. We only preregistered the more recent studies after we understood the problem we were seeking to solve.

Table A11: Description of each study conducted in preparing this manuscript.

Descriptive name	Provider	Topic	Start date	End date	Pre-registration	IRR estimate	Respondents	Tasks	Respondent-Tasks
12 replications v1.1	MTurk	Variety	1/20/2019	3/6/2019	NA	Between waves	113	24	2,712
12 replications v1.2	DLABSS	Variety	5/18/2019	7/5/2019	NA	Between waves	42	24	1,008
12 replications v2	MTurk	Variety	6/10/2019	7/3/2019	NA	Between waves	205	24	4,920
Consistency, complexity, and divergence	Lucid Marketplace	Candidates	5/11/2020	5/21/2020	NA	Between waves	474	30	14,220
Simplest case consistency	Lucid Marketplace	Candidates	5/22/2020	6/1/2020	NA	Between waves	100	4	400
Consistency, policy only	Lucid Marketplace	Candidates	6/6/2020	6/16/2020	NA	Between waves	594	4	2,376
Respondent characteristics vs profile-pair combos v1	DLABSS	Housing	9/23/2020	12/19/2020	NA	Between waves	335	32	10,720
Is IRR worse in conjoints?	Prolific	Policies	3/15/2021	3/23/2021	NA	Between waves	503	6	3,018
Respondent characteristics vs profile-pair combos v2	Prolific	Housing	6/30/2021	7/17/2021	https://osf.io/xgubq	Both	611	32	19,552
Do powerful attributes reduce error?	Prolific	Candidates	8/9/2021	8/11/2021	https://osf.io/y2edx	Within wave	431	16	6,896
8 replications	Lucid Theorem	Variety	3/24/2022	6/13/2022	https://osf.io/hw8r7	Within wave	3,289	12	39,468
Systematic IRR	Lucid Marketplace	Media sources	10/27/2022	11/2/2022	https://osf.io/f26am	Within wave	2,641	12	31,692
Open-ended comments	Lucid Theorem	Candidates	5/15/2023	5/17/2023	NA	Within wave	134	6	804
Total							**9,472**	**226**	**137,786**

References

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

Agranov, Marina and Pietro Ortoleva (2022). "Revealed Preferences for Randomization: An Overview". In: *AEA Papers and Proceedings*. Vol. 112. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 426–430.

Alriksson, Stina and Tomas Öberg (2008). "Conjoint analysis for environmental evaluation". In: *Environmental Science and Pollution Research* 15.3, pp. 244–257.

Ansolabehere, Stephen, James M. Snyder, and Charles Stewart (2001). "Candidate Positioning in U.S. House Elections". In: *American Journal of Political Science* 45.1, pp. 136–159.

Arias, Sabrina B. and Christopher W. Blair (2022). "Changing Tides: Public Attitudes on Climate Migration". In: *The Journal of Politics* 84.1, pp. 560–567.

Atkeson, Lonna Rae and Brian T Hamel (2020). "Fit for the job: Candidate qualifications and vote choice in low information elections". In: *Political Behavior* 42.1, pp. 59–82.

Bansak, Kirk and Libby Jenke (2023). "Odd Profiles in Conjoint Experimental Designs: Effects on Survey-Taking Attention and Behavior". In.

Bastounis, Anastasios, John Buckell, Jamie Hartmann-Boyce, Brian Cook, Sarah King, Christina Potter, Filippo Bianchi, Mike Rayner, and Susan A Jebb (2021). "The impact of environmental sustainability labels on willingness-to-pay for foods: a systematic review and meta-analysis of discrete choice experiments". In: *Nutrients* 13.8, p. 2677.

Bechtel, Michael M. and Kenneth F. Scheve (2013). "Mass support for global climate agreements depends on institutional design". In: *Proceedings of the National Academy of Sciences* 110.34, pp. 13763–13768.

Bernauer, Thomas and Robert Gampfer (2015). "How robust is public support for unilateral climate policy?" In: *Environmental Science & Policy* 54, pp. 316–330.

Blackman, Alexandra Domike (2018). "Religion and foreign aid". In: *Politics and Religion* 11.3, pp. 522–552.

Crowder-Meyer, Melody, Shana Kushner Gadarian, and Jessica Trounstine (2020). "Voting Can Be Hard, Information Helps". In: *Urban Affairs Review* 56.1, pp. 124–153.

De la Cuesta, Brandon, Naoki Egami, and Kosuke Imai (2022). "Improving the external validity of conjoint analysis: The essential role of profile distribution". In: *Political Analysis* 30.1, pp. 19–45.

Eshima, Shusei and Daniel M. Smith (2022). "Just a Number? Voter Evaluations of Age in Candidate-Choice Experiments". In: *The Journal of Politics* 84.3, pp. 1856–1861.

Ganter, Flavien (2021). "Identification of Preferences in Forced-Choice Conjoint Experiments: Reassessing the Quantity of Interest". Forthcoming, *Political Analysis*.

Hainmueller, Jens and Daniel J. Hopkins (2015). "The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants". In: *American Journal of Political Science* 59.3, pp. 529–548.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto (2014). "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments". In: *Political Analysis* 22.1, pp. 1–30.

Hankinson, Michael (2018). "When do renters behave like homeowners? High rent, price anxiety, and NIMBYism". In: *American Political Science Review* 112.3, pp. 473–493.
Incerti, Trevor (2020). "Corruption information and vote share: A meta-analysis and lessons for experimental design". In: *American Political Science Review* 114.3, pp. 761–774.
James, William (1975). *The Meaning of Truth*. Cambridge, MA: Harvard University Press.
Jenke, Libby, Kirk Bansak, Jens Hainmueller, and Dominik Hangartner (2021). "Using Eye-Tracking to Understand Decision-Making in Conjoint Experiments". In: *Political Analysis* 29.1, pp. 75–101.
Keller, L Robin (1992). "Properties of utility theories and related empirical phenomena". In: *Utility theories: Measurements and applications*, pp. 3–23.
Kertzer, Joshua D., Jonathan Renshon, Keren Yarhi-Milo, et al. (2019). "How Do Observers Assess Resolve?" In: *British Journal of Political Science* 51.1, pp. 308–330.
King, Gary, Michael Tomz, and Jason Wittenberg (Apr. 2000). "Making the Most of Statistical Analyses: Improving Interpretation and Presentation". In: *American Journal of Political Science* 44.2, pp. 341–355. URL: bit.ly/makemost.
Kirkland, Patricia A. and Alexander Coppock (2018). "Candidate Choice Without Party Labels: New Insights from Conjoint Survey Experiments". In: *Political Behavior* 40, pp. 571–591.
Leeper, Thomas J. and Joshua Robison (2020). "More important, but for what exactly? The insignificant role of subjective issue importance in vote decisions". In: *Political Behavior* 42.1, pp. 239–259.
Lo, Andrew W, Katherine P Marlowe, and Ruixun Zhang (2021). "To maximize or randomize? An experimental study of probability matching in financial decision making". In: *Plos one* 16.8, e0252540.
Louviere, Jordan, Deborah Street, Richard Carson, Andrew Ainslie, JR DeShazo, Trudy Cameron, David Hensher, Robert Kohn, and Tony Marley (2002). "Dissecting the random component of utility". In: *Marketing letters* 13, pp. 177–193.
Mamine, Fateh, Jean Joseph Minviel, et al. (2020). "Contract design for adoption of agrienvironmental practices: a meta-analysis of discrete choice experiments". In: *Ecological Economics* 176, p. 106721.
McCarty, Nolan, Keith T. Poole, and Howard Rosenthal (2007). *Polarized America The Dance of Ideology and Unequal Riches*. Cambridge: MIT Press.
McConnaughy, Corrine M., Ismail K. White, David Leal, and Jason Casellas (2010). "A Latino on the Ballot: Explaining Co-Ethnic Voting among Latinos and White Americans". In: *The Journal of Politics* 72.4, pp. 1199–1211.
Mummolo, Jonathan (2016). "News from the other side: How topic relevance limits the prevalence of partisan selective exposure". In: *The Journal of Politics* 78.3, pp. 763–773.
Mummolo, Jonathan and Clayton Nall (2017). "Why partisans do not sort: The constraints on political segregation". In: *The Journal of Politics* 79.1, pp. 45–59.
Ono, Yoshikuni and Barry C. Burden (2019). "The contingent effects of candidate sex on voter choice". In: *Political Behavior* 41.3, pp. 583–607.
Peyton, Kyle, Gregory A. Huber, and Alexander Coppock (2022). "The Generalizability of Online Experiments Conducted During the COVID-19 Pandemic". In: *Journal of Experimental Political Science* 9.3, pp. 379–394.

Popkin, Samuel L. (1991). *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. Chicago: University of Chicago Press.

Rahn, Wendy M. (1993). "The Role of Partisan Stereotypes in Information Processing about Political Candidates". In: *American Journal of Political Science* 37.2, pp. 472–496.

Sances, Michael W. (2018). "Ideology and vote choice in US mayoral elections: Evidence from Facebook surveys". In: *Political Behavior* 40.3, pp. 737–762.

Schachter, Ariela (2016). "From "different" to "similar" an experimental approach to understanding assimilation". In: *American Sociological Review* 81.5, pp. 981–1013.

Schwarz, Norbert (2007). "Cognitive aspects of survey methodology". In: *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 21.2, pp. 277–287.

Schwarz, Susanne and Alexander Coppock (2022). "What Have We Learned about Gender from Candidate Choice Experiments? A Meta-Analysis of Sixty-Seven Factorial Survey Experiments". In: *The Journal of Politics* 84.2, pp. 655–668.

Starmer, Chris (2000). "Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk". In: *Journal of economic literature* 38.2, pp. 332–382.

Strange, Austin M, Ryan D Enos, Mark Hill, and Amy Lakeman (2019). "Online volunteer laboratories for human subjects research". In: *PloS one* 14.8, e0221676.

Teele, Dawn Langan, Joshua Kalla, and Frances Rosenbluth (2018). "The ties that double bind: Social roles and women's underrepresentation in politics". In: *American Political Science Review* 112.3, pp. 525–541.

Ternovski, John and Lillia Orr (2022). "A Note on Increases in Inattentive Online Survey-Takers Since 2020". In: *Journal of Quantitative Description: Digital Media* 2, pp. 1–35.