Correcting Measurement Error Bias in Conjoint Survey Experiments*

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

March 28, 2023

Abstract

Conjoint survey designs are spreading across the social sciences due to their unusual capacity to identify many causal effects from a single randomized experiment. Unfortunately, because the nature of the conjoint design violates aspects of best practices in questionnaire construction, conjoint experiments generate substantial measurement error-induced bias. By replicating both data collection and analysis of eight prominent conjoint studies, all of which closely reproduce published results, we show that about half of all observed variation in this most common type of conjoint experiment is effectively random noise. We then discover a common empirical pattern in how measurement error appears in conjoint studies and use it to derive an easy-to-use statistical method to correct the bias.

*This paper and its accompanying Supplementary Appendix are available at GaryKing.org/conjointE. We thank Sabrina Arias, Michael Bechtel, Barry Burden, Naoki Egami, Toby Heinrich, Dan Hopkins, Kosuke Imai, Josh Kalla, Bob Kubinec, Kasey Rhee, Dawn Teele, Dustin Tingley, Joonseok Yang, and the participants at the 2020 Meeting of the Society for Political Methodology, the 2021 Joint Quantitative Political Science Conference for Asia and Australasia, the 2021 Midwest Political Science Association Conference, a seminar at Seoul National University in January 2023, and a method reading group at American University in March 2023 for helpful suggestions.

†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

1 Introduction

The advantage of random treatment assignment in survey experiments is that modeling and ignorability assumptions are unnecessary. However, running multiple survey experiments can be expensive: if n survey respondents generate a causal estimate with acceptable uncertainty levels, $n \cdot m$ are usually needed to estimate m causal effects. One way to reduce this cost is to design and administer a *conjoint* experiment, which enables researchers in certain situations to estimate m causal effects with only n survey respondents (see Green and Srinivasan, 1978; Shamir and Shamir, 1995; Hainmueller, Hopkins, and Yamamoto, 2014). Conjoint analyses have been used in nearly 100,000 articles across academia and marketing (according to Google Scholar), and they are rapidly increasing in popularity in political science and across the social sciences (see Supplementary Appendix A1).

We analyze the most commonly used conjoint design, which presents each of n respondents with a choice between two "profiles" (i.e., candidates, products, etc.), each with randomly assigned values (or "levels") for a set of k "attributes." (Researchers also often ask each respondent to complete several randomly assigned conjoint "tasks" to increase statistical power further.) Modern conjoint estimators, which use no modeling assumptions, are unbiased for a specific type of causal effect that we clarify below.

Prior research shows that conjoint designs have strong external validity (Auerbach and Thachil, 2018; Hainmueller, Hangartner, and Yamamoto, 2015) and low social desirability bias (Horiuchi, Markovich, and Yamamoto, 2022), and that cognitive burdens do not increase much as attributes k (and tasks t) grow large (Bansak et al., 2018; Bansak et al., 2021a; Jenke et al., 2021). Recent methodological advances improve the conjoint estimand (De la Cuesta, Egami, and Imai, 2022; Ganter, 2021; Leeper, Hobolt, and Tilley, 2020; Zhirkov, 2022), clarify the interpretability of conjoint results (Abramson, Koçak, and Magazinnik, 2022), and address multiple testing issues (Goplerud, Imai, and Pashley, 2022; Liu and Shiraito, 2022).

Survey research best practices advise posing simple, concise, concrete, and easy-to-understand questions (Payne, 2014). Although the text of conjoint outcome questions

could not be simpler (e.g., in a candidate-choice conjoint experiment, "Do you prefer Candidate A or B?"), the content of the question is about hypothetical people, places, policies, or things described by long tables of attributes, often with complex descriptions, and sometimes in contradictory or confusing combinations. As attributes are randomly assigned, some profile sets are uncommon, illogical, or unintuitive. Although the worst of these (such as impossible combinations) are often excluded, many studies avoid imposing too many cross-attribute constraints to keep statistical analyses simple (see Bansak et al., 2021b). Instead of being able to follow best survey practices by eliciting information about individual attributes separately, the conjoint design asks the respondent (rather than the researcher) to resolve conflicts. The contrast with best practices for traditional surveys is striking because "[o]ne of the first things a researcher learns in questionnaire construction is to avoid double-barreled questions, that is, questions in which opinions about two objects are joined together so that respondents must answer two questions with one answer" (Bradburn, Sudman, and Wansink, 2004, p.142). Of course, choices between paired entities in the real world (e.g., in elections) are often multi-barreled, and so asking about attributes separately is not an option in conjoint any more than you can vote for only some of a political candidate's policy preferences and characteristics.[1]

The statistical consequence of these inherent complexities of conjoint analyses is *measurement error* (McCullough and Best, 1979), a well-known methodological problem that can potentially bias causal inferences in any direction by any amount. Unfortunately, measurement error and its consequences have been ignored in nearly all conjoint applications. Yet, as we demonstrate in this paper, even highly attentive survey respondents produce data with substantial measurement error, which we quantify via intra-respondent unreliability: when faced with two identical conjoint tasks just moments apart, respondents select the same profile only about 75% of the time. Because flipping coins produces 50%

[1] Bansak et al. (2021b) make clear the conjoint design trade-off between the importance of adding "relevant" attributes and the likelihood of inducing measurement error bias: "[I]ncluding too few attributes will make it difficult to interpret the substantive meaning of AMCEs [the causal effects], since respondents might associate an attribute with another that is omitted from the design" (Bansak et al., 2021b, p. 25; see also Dafoe, Zhang, and Caughey, 2018). But "including too many attributes might increase the cognitive burden of the tasks excessively, inducing respondents to satisfice" (Bansak et al., 2021b, p. 25; see also Krosnick, 1999).

agreement, 75% means that roughly half of all observed variation in binary choice conjoint experiments is measurement error before even considering conceptual, sampling, or misspecification errors. (These results are consistent with results found in other fields. See Bryan et al. 2000; Mørkbak and Olsen 2015; Skjoldborg, Lauridsen, and Junker 2009).

In this paper, we field a dozen surveys (with a total of more than 9,000 respondents and over 130,000 respondent-tasks) to replicate from scratch the data collection and analyses of eight major published conjoint studies in political science and estimate the levels and types of measurement error in each. We discover an empirical pattern in how conjoint studies generate measurement error across these analyses and a sequence of other auxiliary studies. Namely, we find that most measurement error is not systematically correlated with the information contained in conjoint tasks. We then use this pattern to develop a simple statistical correction for the resulting biases. As we explain, everything necessary to correct the bias in an application can be estimated via a slight modification of the standard conjoint design or via a separate survey run afterward. In many situations, correcting the bias will make results stronger, but not always; either way, the correction is easy to apply. Therefore, researchers should not merely make assumptions without the correction. We conclude with recommendations for conducting conjoint studies and offer open-source software (or a few lines of code to include in any other software) to implement them.

2 How Measurement Error Induces Bias

We begin by clarifying the standard conjoint design setup and then studying the type and consequences of measurement error. Below, we use mnemonic notation wherever convenient, which we highlight by underlining a character in a word corresponding to a symbol's meaning. We also use Greek letters for unknown (or chosen values of) quantities and Roman letters for observed quantities.

2.1 Data

Without loss of generality, consider a simple conjoint experiment with each respondent making a series of choices, each between two candidate profiles. Formally, we give each individual i ($i = 1, \ldots, N$) a task t ($t = 1, \ldots, T$) of choosing between Candidates 1 and 2 and record their choice as C_{it}, with values 1 if Candidate 1 is chosen and 0 if Candidate 2 is chosen. The choice C_{it} is made for each respondent-task, which is the unit of analysis (enabling us to structure the data as a rectangular dataset with $N \times T$ rows). Nothing in our analysis depends on the existence of more than one task per respondent, although most applications allow at least 3 to 5 tasks.

Each task is a randomly assigned vector of attribute values A_{it} describing the two hypothetical candidate profiles presented to a respondent. The contents of A_{it} for any one respondent and task is a vector of attributes, each of which has two levels corresponding to the two candidates. For example, if one element of the attribute vector is "incumbency status," which for any one candidate is 0 for a nonincumbent and 1 for an incumbent, the element of A_{it} takes on one of four possible values (aka "attribute levels") for the two candidates: (0, 0), (0, 1), (1, 0), and (1, 1). All the other elements of the attribute vector would thus also have pairs of values.[2]

More generally, we partition the attribute vector of pairs as $A_{it} = \{A_{it,\ell}, A_{it,-\ell}\}$ and define a_ℓ as the pair of values for one "attribute of interest" and $a_{-\ell}$ as a vector of the pairs for all the other attributes. Because the levels for all attributes are randomly assigned, post-treatment bias is not an issue. Therefore, different attributes can take turns as the attribute of interest. Finally, we represent the two chosen values for the attribute of interest for Candidates 1 and 2 as $A_{it,\ell} = \alpha \equiv \{\alpha_1, \alpha_2\}$.

The vector of pairs of values A_{it} represents attributes researchers offer to respondents. It also includes two additional attributes implicitly offered every time, the *names* of the

[2]Some researchers structure the data differently by defining the unit of analysis as the respondent-task-profile choice, with a respondent's choice represented twice: $C_{it_1} = 1$ if the respondent chooses Candidate 1 and 0 otherwise and $C_{it_2} = 1$ if the same respondent for the same task chooses Candidate 2 and 0 otherwise, with $C_{it_2} = 1 - C_{it_1}$. The advantage of this alternative data structure is that each element of A_{t_b} (with $t = 1, \ldots, T$ and binary option $j = 1, 2$) corresponds to a single candidate. It is easier to code (e.g., incumbency status is merely 1 or 0) and, under some conditions, does not change point estimates. However, due to the artificially induced dependence, researchers will sometimes need to use special procedures to calculate standard errors and more complicated statistical modeling to estimate interactions or other quantities.

candidates in each task (e.g., "Candidate 1" and "Candidate 2") and the *order* in which they are presented (e.g., "left" and "right" columns). Conjoint experiments that keep the names constant over i and t do not explicitly code this attribute. When researchers assume that candidate choice is not a function of presentation order—so that swapping the candidates presented in the left and right columns has no effect—the order attribute is also not explicitly coded. Although these implicit features are not used in most applications, they could be useful for some purposes, such as studies on party-label or ballot-order effects.

Finally, we also describe each respondent by a vector of exogenous personal characteristics P_i, such as demographics, socioeconomic status, or political or other views. Although the content of A_{it} is controlled by the investigator and randomly assigned, P_i is observed and cannot be randomized. Researchers commonly use P_i to define subgroups (Leeper, Hobolt, and Tilley, 2020), within which all our methods can be repeated.

2.2 Quantities of Interest

We assume that each respondent has a true, unobserved *preference* between the two candidates on each task, $\rho_{it} \in \{0, 1\}$. Importantly, it may be different from the respondent's observed choice due to measurement error, as formalized below. We then define two primary quantities of interest, which can be calculated for the entire sample or within subgroups defined by personal characteristics, P_i.

First is the *average preference* or "marginal mean" for a candidate that has specific values of the attribute of interest for the two candidates. For example, we may be interested in the average preference for Candidate 1 in tasks with $A_{it,\ell} = \alpha \equiv \{\alpha_1, \alpha_2\}$, which is:[3]

$$\rho(\alpha) = \underset{A_{it,\ell}=\alpha}{\text{mean}} \left(\rho_{it}\right) \tag{1}$$

For example, consider the percentage of respondents choosing a candidate if the candidate's incumbency status is 1 (an incumbent). Other descriptive quantities can be similarly

[3]To simplify the notation in the text, we formally define a mean function: for set S with cardinality $\#S$, the mean over i of a function $g(i)$ as $\text{mean}_{i \in S}[g(i)] = \frac{1}{\#S} \sum_{i=1}^{\#S} g(i)$. When the set S is unambiguous, we omit it and write $\text{mean}_i[g(i)]$.

defined.[4]

The second quantity of interest is the *average marginal component effect* commonly known (and herein referred to) as AMCE (Hainmueller, Hopkins, and Yamamoto, 2014). The AMCE is a treatment effect—changing the values of $A_{it,\ell}$ from α to the hypothetical values α', $\rho_{it}(\alpha) - \rho_{it}(\alpha')$—averaged over respondents:

$$\theta = \underset{i \in A_{it,\ell}=\alpha}{\text{mean}} \; \rho(\alpha) - \rho(\alpha') \tag{2}$$

In the example we used above, the AMCE is the effect of a candidate as an incumbent (as compared to a nonincumbent) on respondents' probability of choosing the candidate averaged over respondents. Other causal quantities can be defined in similar ways, such as differences among subgroups (see Section 6.2).

2.3 Randomization

A *randomized conjoint experiment* assigns the values of A_{it} randomly with respect to i and t. In contrast, the values of P_i are fixed characteristics of respondents and cannot be assigned by the investigator. As such, the power of randomization can be used to identify causal effects of elements of A but not of P.

Given this setting, studies of the effects of P, even in a conjoint experiment, should be regarded as an observational study, requiring careful specification and ignorability assumptions. Researchers can also use P to define exogenous strata within which the effects of A can be estimated with the benefits of randomization (aka subgroup or heterogeneous treatment effects).

2.4 Observation Mechanism

The particular type of measurement error in conjoint studies is what we call *swapping error*, where some of the respondents' reported answers to a binary question represent

[4]By averaging over i, we also average over all other (randomly assigned) attribute sets for attributes other than ℓ, and all individuals i. The crucial point is that the population over which this average is being taken may not be the one of interest (such as the population of all U.S. adult citizens in a traditional survey seeking the percent who approve of the president). Instead, this average is defined by the attributes the investigator happens to include in the conjoint experiment. Therefore, if researchers add an extra attribute to estimate an additional marginal mean, they also change the marginal mean for all other attributes. De la Cuesta, Egami, and Imai (2022) show how to avoid this issue by changing Equation 1 to a weighted mean.

their (true) preference ρ, but other answers are swapped with the wrong ones $1-\rho$. (This can occur with any binary outcome variable, even if not from a conjoint.) To formalize this idea, define the respondent's reported *choice* between the candidates as:

$$C_{it} = \begin{cases} \rho_{it} & \text{w.p. } 1-\tau_{it} \\ 1-\rho_{it} & \text{w.p. } \tau_{it}, \end{cases} \tag{3}$$

where τ_{it} is the probability of a swapping error, i.e., when the respondent's choice does not reflect their true preference ("w.p." is a standard mathematical notation for "with probability"). Almost all prior conjoint research assumes $\tau_{it} = 0$, for all i and t, which we show below is not justified.

We assume that we can obtain an unbiased and consistent estimate of $1-\tau_{it}$ by calculating *intra-respondent reliability* (IRR), usually by asking the same question twice at the start and end of a series of conjoint tasks or different surveys administered at different times. We also make the reasonable assumption that some information exists in the data, so that $\tau_{it} \in [0, 0.5)$.[5]

We assume the absence of measurement error in the attributes A, which the investigator chooses. For most of our results, we do not use or need to make assumptions about the presence or absence of measurement error in the personal characteristics P, which are elicited from the respondent via traditional survey questions.

2.5 Consequences of Ignoring Measurement Error

As introductory regression textbooks commonly show, random, mean-zero measurement error added to a continuous outcome variable in a linear regression causes no coefficient bias. This consequence is easy to see: it is equivalent to a regression with no measurement error in the outcome variable but a higher residual error. However, swapping error in the binary outcome variable used in conjoint analyses cannot be mean zero as it swaps some zeros with ones and ones with zeros. Thus, there is no sense in which swapping error comes out in the wash; it cannot be ignored. The bias induced by swapping error also

[5] Our methods make no assumptions about the exact mechanism that produces $\tau > 0$, although future research to understand the process may suggest new ways of avoiding or correcting for the bias. For example, do respondents choose randomly, by convenience, or in some other way when they are exactly indifferent? Could measurement error be reduced by better studies of or corrections for inattention? Could we reduce τ by posing the conjoint question or presenting the attribute list differently?

cannot be corrected by general purpose methods for correcting measurement error bias which assume the error is mean zero (e.g., Blackwell, Honaker, and King, 2017).[6]

Formally, the standard estimators of $\rho(\alpha)$ and θ,

$$\hat{\rho}(\alpha) = \underset{A_{it,\ell}=\alpha}{\text{mean}}\left(C_{it}\right), \qquad \hat{\theta} = \hat{\rho}(\alpha) - \hat{\rho}(\alpha'),$$

are unbiased if $\tau_{it} = 0$ for all i and t. However, they are biased in the presence of non-zero swapping error:

$$\begin{aligned} E[\hat{\rho}(\alpha)] &= \underset{A_{it,\ell}=\alpha}{\text{mean}}\left(E[C_{it}]\right) \\ &= \underset{A_{it,\ell}=\{\alpha_1,\alpha_2\}}{\text{mean}}\left[\rho_{it}(1-\tau_{it}) + (1-\rho_{it})\tau_{it}\right] \\ &= \hat{\rho}(\alpha) - 2 \cdot \underset{A_{it,\ell}=\alpha}{\text{mean}}\left(\rho_{it}\tau_{it}\right) + \tau \\ &\neq \rho(\alpha) \end{aligned} \tag{4}$$

and thus

$$\begin{aligned} E(\hat{\theta}) &= E[\hat{\rho}(\alpha)] - E[\hat{\rho}(\alpha')] \\ &= \hat{\theta} + 2\left[\underset{A_{it,\ell}=\alpha}{\text{mean}}\left(\rho_{it}\tau_{it}\right) - \underset{A_{it,\ell}=\alpha'}{\text{mean}}\left(\rho_{it}\tau_{it}\right)\right] \\ &\neq \theta. \end{aligned} \tag{5}$$

When estimating the marginal mean (or average preference) or AMCE by subgroups (defined by P_i), all of our results hold within each subset.

3 Correcting Measurement Error Bias

Our quantities of interest are functions of the unobserved preferences ρ_{it}. But our estimators are functions of the observed choices C_{it}, which differ from ρ_{it} because of the swapping error probability τ_{it}. As we demonstrate below, correcting measurement error bias is straightforward if we have an estimate of τ_{it}. In principle, however, τ_{it} may vary over individuals i and tasks t, which would seem to require that researchers obtain $n \times T$ estimates of the probability of swapping error. If we estimated this swapping error

[6] Measurement error in alternative conjoint designs with ranking or rating also induce bias that cannot be ignored.

probability by intra-respondent reliability, estimating each of the $n \times T$ swapping error probabilities would require a reasonably sized sample (at least, say, a hundred observations) with each respondent asked the same question twice. This approach will typically be infeasible given research budget constraints.

We solve this problem in two steps. First, as described in Section 4, we provide extensive empirical evidence that τ_{it} does not vary systematically with different combinations of attributes. That is, within a conjoint survey, $\tau_{it} \approx \tau$, a finding that drastically simplifies the estimation of swapping error probability down to a single parameter. We find that τ can vary somewhat across applications (and respondent characteristics). So it needs to be estimated for every conjoint survey. If researchers are interested in subgroup comparisons, they also need to estimate τ for each subgroup. Second, as described in Section 5, we offer several easy ways of estimating τ in any situation that involves collecting a small amount of additional data by altering the survey design or from existing conjoint data with no new data collection at all.

Furthermore, although researchers need to use or estimate only a single parameter for each bias correction, they do not need to assume $\tau_{it} = \tau$. Instead, as we now demonstrate, we only need to make a less restrictive assumption that swapping error probabilities are linearly unrelated to respondent preferences: $\text{Cov}(\rho_{it}, \tau_{it}) = 0$, which implies that $\text{mean}_{A_{it,\ell}=\alpha}(\rho_{it}\tau_{it}) = \rho(\alpha)\tau$.

Under this relaxed assumption, we simplify the bias expressions in Equations 4 and 5, respectively, as

$$E\left[\hat{\rho}(\alpha) \;\middle|\; \text{Cov}(\rho_{it}, \tau_{it}) = 0\right] = \rho(\alpha) \cdot (1 - 2\tau) + \tau \tag{6}$$

$$E\left[\hat{\theta} \;\middle|\; \text{Cov}(\rho_{it}, \tau_{it}) = 0\right] = \theta(1 - 2\tau) \tag{7}$$

With these results, we define alternative estimators for MM and the AMCE as,

$$\tilde{\rho}(\alpha) = \frac{\hat{\rho}(\alpha) - \tau}{1 - 2\tau}, \qquad \tilde{\theta} = \frac{\hat{\theta}}{1 - 2\tau}, \tag{8}$$

which are unbiased if τ is known, $E[\tilde{\rho}(\alpha)] = \rho(\alpha)$ and $E(\tilde{\theta}) = \theta$. They are consistent and approximately unbiased with a consistent estimate of τ (Section 6 also shows that they are also approximately unbiased with smaller mean square error). Finally, unlike logit, probit,

regression, and other fully parametric approaches, these estimators require no modeling assumptions at all.

This approach can also be used for interactions by redefining the value of an attribute $A_{it,\ell} \equiv \alpha$ to indicate more than a single element of the attribute vector.

In some applications, researchers are interested in comparing $\tilde{\rho}(\alpha)$ or $\tilde{\theta}$ among subgroups of respondents, such as Democrats and Republicans. In most of these studies, researchers will want to estimate τ within the chosen subgroups, effectively repeating all our procedures and recommendations within each.

Equations 8 show that the bias correction will always increase the absolute value of the AMCE. Similarly, the bias correction for MM will always increase its absolute distance from 0.5; that is, if $\hat{\theta} < 0.5$, the corrected estimate will be smaller than the biased estimate, but if $\hat{\theta} > 0.5$, the corrected estimate will be larger. This can be seen by solving for the difference between the corrected and uncorrected estimates as

$$\tilde{\rho} - \hat{\rho} = \frac{\tau}{1 - 2\tau}(2\hat{\rho} - 1),$$

and recalling that $\tau_{it} \in [0, 0.5)$. Subgroup differences of either MM or AMCE can increase, decrease, or flip the signs of the estimates.

Computing standard errors for $\tilde{\rho}(\alpha)$ and $\tilde{\theta}$ requires an extra step because of the uncertainty in $\hat{\tau}$. We show how to do this in Appendix A in three different ways that optimize for speed, convenience, or familiarity. Researchers who use the open source software we make available with this paper will have the advantage of all three.

4 Patterns in Conjoint-Induced Measurement Error

We now narrow down the necessary statistical assumptions for our measurement error corrections by (1) replicating the data collection and analysis of eight prior published conjoint studies; (2) estimating the average intra-respondent reliability within each study; (3) revealing the lower reliability of conjoint questions compared to traditional survey questions; (4) describing the lack of evidence for systematic variation in reliability across attribute combinations within studies; and (5) showing how reliability varies over the personal characteristics used for subgroup estimation.

4.1 Eight Replications

First, we replicate existing studies. To keep our analyses as close to the literature as possible, we choose eight published political science conjoint studies, with a preference for those in major journals (see Appendix A2 for details on our study selection procedures). Furthermore, we chose substantively diverse topics, including choices between housing developments, climate agreements, political candidates, immigrants, etc. They include Arias and Blair (2022), Bechtel and Scheve (2013), Blackman (2018), Hainmueller and Hopkins (2015), Hankinson (2018), Mummolo and Nall (2017), Teele, Kalla, and Rosenbluth (2018), and Ono and Burden (2019). We then conduct a series of survey experiments using U.S. samples with quotas with respect to age, gender, race, ethnicity, and region (from Lucid Marketplace; see Coppock and McClellan 2019). Although only Bechtel and Scheve (2013) report using attention checks, we give conservative results on intra-respondent reliability by dropping respondents who failed an attention check administered prior to our conjoint task (see Appendix A3 for details on the attention checks used in our studies and a comparison of IRR by respondent attentiveness; we find little evidence that respondent inattentiveness explains low IRR in conjoint studies).

Most replication studies begin with the data and methods from a published article and try to replicate (or "reproduce") its tables and figures (King, 1995). We instead begin at an earlier point in the replication process: For each of the eight studies, we collect new survey responses following each article's experimental design and rerun the same statistical analysis. We do this for all the average marginal causal effects (AMCEs) computed in any of the eight studies, 170 estimates in total. Figure 1 presents a scatterplot of these AMCEs from the original studies plotted horizontally and our replication of each AMCE in our new data plotted vertically. The AMCEs from each study, along with a regression line fit to its points, are color coded (see the figure legend).

Despite the expected sampling error and any systematic error due to differences in the sample time frame, details of implementation, and sample characteristics, the results in Figure 1 reveal a surprisingly close correspondence between the originally published estimates and the estimates based on our replications of these studies. This can be seen in the

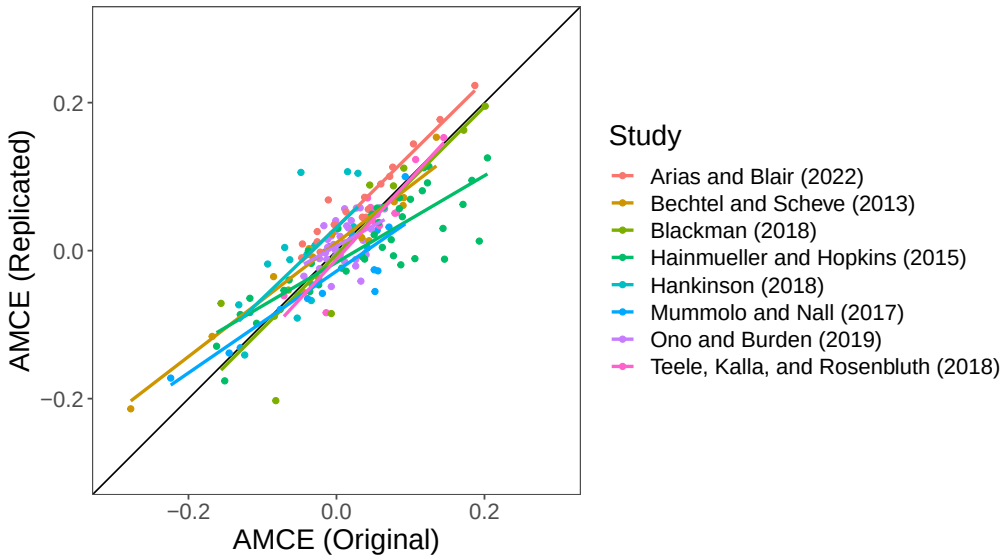

Figure 1: Eight Replications: Scatterplot of average marginal causal effect point estimates from the original studies (plotted horizontally) and our replications in new data (plotted vertically), color-coded along with a regression line fit to all estimates from each study.

eight (different colored) regression lines, all fairly close to the (black) 45-degree line, and the point estimates color-coded are tightly scattered around each corresponding regression line. Indeed, the median correlation for the estimates in a study between the published and our replicated results is a remarkable 0.9. Given the number of replication failures across scientific fields in recent years (Gilbert et al., 2016; Open Science Collaboration, 2015), it is comforting to see the uniformly high level of transparency and scientific rigor achieved in the literature on conjoint-based political science experiments displayed in Figure 1.

4.2 Estimates of Average Reliability

Second, we estimate the average intra-respondent reliability $(1 - \tau_{it})$ for each of the eight studies. We do this by assigning a different random subset of two of the eight original studies to each of the 3,289 respondents. We standardize the number of tasks per respondent across our eight replications to five (which is the mean, median, and mode of the studies) and then add a sixth conjoint question that repeats the first (randomly selected) question at the end of the task list with the profile order switched.[7] That is, just a few moments

[7]Supplementary Appendix A9 reports on three additional surveys we conducted to study the effect of flipping vs. not flipping the profiles in the repeated task. We found that in two of the three surveys keeping the profile the same generated a slightly smaller estimate of τ, small enough so that the difference for our

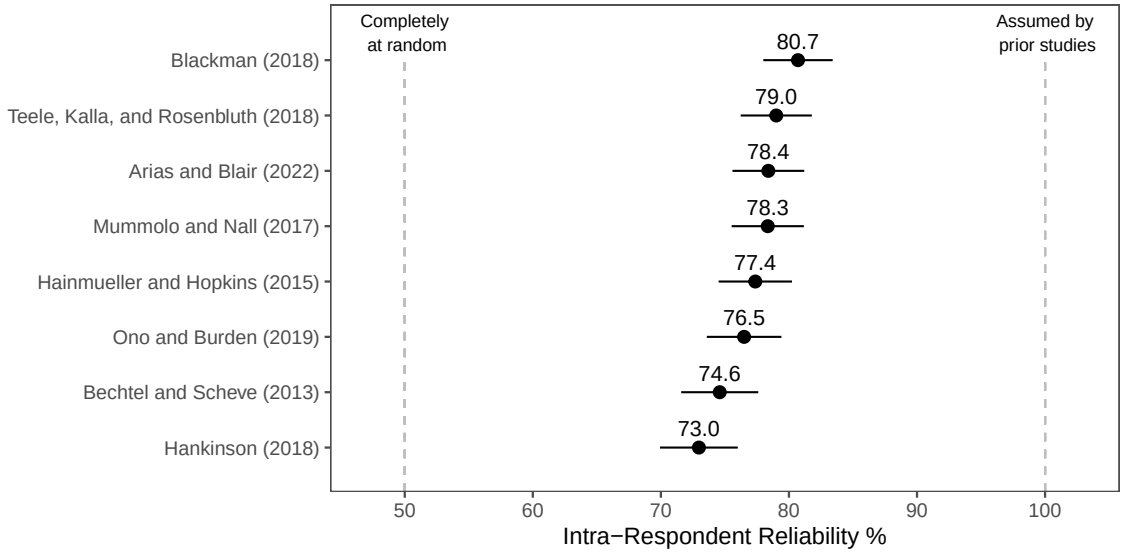

Figure 2: Intra-Respondent Reliability of Eight Prior Studies. Point estimates appear as dots with 95% confidence intervals as a horizontal line.

after a respondent chooses between two candidates, we ask this same person essentially the same question a second time and see whether their answer is the same. Then, our estimate of intra-respondent reliability is the average percent agreement between these first and last (repeated) questions.[8] Results appear in Figure 2. The horizontal axis in Figure 2 indicates intra-respondent reliability, ranging between respondents flipping coins (50%, at the left) and no measurement error at all as is assumed by most conjoint applications (100%, at the right). Our point estimates appear as dots, with 95% confidence intervals as lines. Intra-respondent reliability for each study is approximately halfway between flipping coins and no measurement error with an average of 77% (and a range of point estimates from 73.0–80.7%). Although almost all conjoint studies assume the absence of measurement error, this figure indicates that approximately half of the variation in these studies is based solely on measurement error.

bias corrections would rarely be substantively meaningful.

[8]We also repeated the entire experiment twice for each individual to increase efficiency, resulting in $12 = (5 + 1) \times 2$ conjoint tasks per respondent. The results focusing only on the first of these two experiments, which was as close as possible to what the original articles used, are substantively similar.

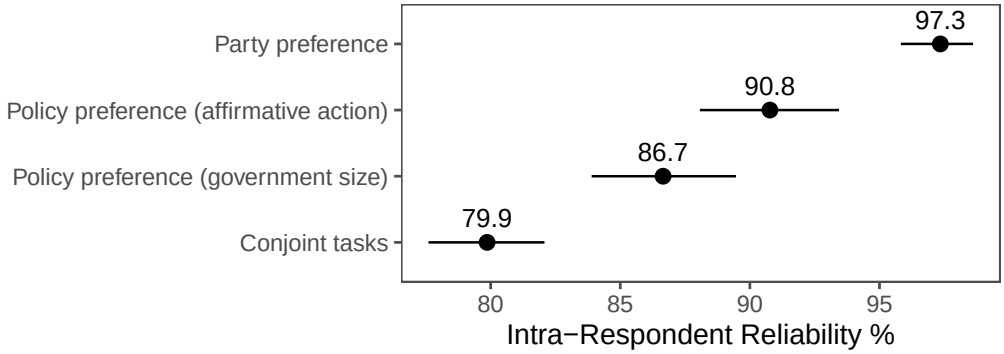

Figure 3: Intra-Respondent Reliability of Traditional Surveys vs. Conjoint

4.3 Reliability Comparisons with Traditional Survey Questions

Third, we provide evidence that, as might be expected, measurement error is higher in conjoint questions than in traditional multiple-choice survey questions with similar content. To do this, we designed and administered a survey with both a candidate-choice conjoint experiment and a series of traditional questions tapping attitudes toward each of the candidate's attributes (i.e., various policy positions and partisanship). In the conjoint experiment, for a given attribute (e.g., "Position on economy"), each level (e.g., "We need a strong government to handle today's complex economic problems" or "The free market can handle these problems without the government being involved") corresponds to one of the multiple answer choices in a traditional survey question (e.g., "Which of the following two statements comes closer to your own opinion?"). We then repeated this survey about one week later and calculated intra-respondent reliability for the conjoint tasks vs. the traditional survey questions. (See Supplementary Appendix A4 for details.)

The results appear in Figure 3. The horizontal axis is again intra-respondent reliability. While the conjoint experiment is 79.9% (at the bottom), all three survey questions have, as would be expected, substantially higher reliability, ranging from 86.7% to 97.4%. Also as expected, all three of the survey reliability estimates are higher than all eight of the original conjoint studies in Figure 2. These results suggest that the source of most of the lower reliability in conjoint experiments is inherent in the more complex design, rather than in details of how the design is implemented.

4.4 No Reliability Variation by Attributes

Fourth, we present evidence that the reliability of conjoint survey questions does not vary systematically within individual studies as a function of the pairs of attribute combinations (i.e., the information contained in conjoint tables). We do this from both a top-down theoretical approach, which we describe here and give empirical evidence in Supplementary Appendix A5, and a bottom-up empirical approach that we present next (and another analogous bottom up approach with details in Supplementary Appendix A6).

4.4.1 Top-down approach

We apply the literature on survey best practices to conjoint studies and develop three theories of how reliability may be reduced as a function of the content of the profile pairs presented to respondents. We developed three plausible hypotheses, all of which failed in empirical tests. First, *inconsistency* is the level of disagreement across attributes within a profile when interpreted on its most prominent dimension. For instance, do Democratic candidate profiles have a coherent set of liberal policy positions? If profiles are inconsistent, we hypothesize that some respondents may become confused, increasing cognitive demands and thus intra-respondent unreliability. Second, *complexity* refers to survey question-wording: how many words appear to respondents in the conjoint table describing each candidate's attributes? How many attributes of each profile are presented to respondents? How complicated is the language used to describe attribute levels? The hypothesis here is that complex conjoint tables may confuse respondents and increase error. Finally, *divergence* refers to the level of dissimilarity between the attribute levels of both profiles. Attribute levels with small absolute differences may encourage respondents to assess options essentially at random, reducing intra-respondent agreement. In a candidate experiment, "moderate Democrat" versus "moderate Republican" is less divergent than "extreme Democrat" versus "extreme Republican"; in a housing development experiment, "3 units versus 5 units" is less divergent than "3 units vs 50 units." Adding to these small divergences are the common situation of some attributes having identical levels for the two candidate profiles, making it more difficult to find those that differ. We hypoth-

esize that respondents will have an easier time choosing between the extreme candidates and the large absolute difference in units than they do between the moderate candidates and the small absolute difference in units, even though if a respondent's preference is for Democrats or bigger developments, both choices should be equally straightforward.

As Supplementary Appendix A5 shows, through numerous survey experiments, we were unable to find evidence that inconsistency, complexity, or divergence account for any of the variation in intra-respondent reliability in realistic conjoint studies. On the theory that the exception proves the rule, we were only able to identify some slight evidence, and just for consistency, in extreme cases that are well outside the bounds of what researchers would likely choose or respondents would see in the real world. We also went further and studied the consequence of attribute sets with a single substantive dominant attribute and failed to find an explanation for reliability there either.

Thus, we find that, in ordinary conjoint experiments, with the types of attributes and levels used in social science applications and with variation one would see in reality, intra-respondent reliability rarely varies in substantial ways as a function of the content of conjoint tasks. An advantage of conjoint analysis is the ability of researchers to offer complicated information to respondents, which then makes measurement error inevitable. However, in part because of this complexity, the degree and type of measurement error are also unrelated to the content of the profile pairs.

4.4.2 Bottom-up approach

As a second approach, we conduct an experiment where we present respondents with a series of six hypothetical media articles (taken from Mummolo 2016). Each profile pair has two attributes (source and headline); the first has three possible levels and the second has four, with both randomly assigned. We exclude ties (i.e., identical profiles on the left and right), leading to a total of 48 possible combinations of profiles. To measure intra-respondent reliability, we also present respondents with another six profile pairs identical to the first six (with the profile appearing on the left and right flipped). In addition to excluding ties, we avoid showing the same conjoint table twice in each set of six tasks for each respondent. As a result, each respondent sees six different profile-pair combina-

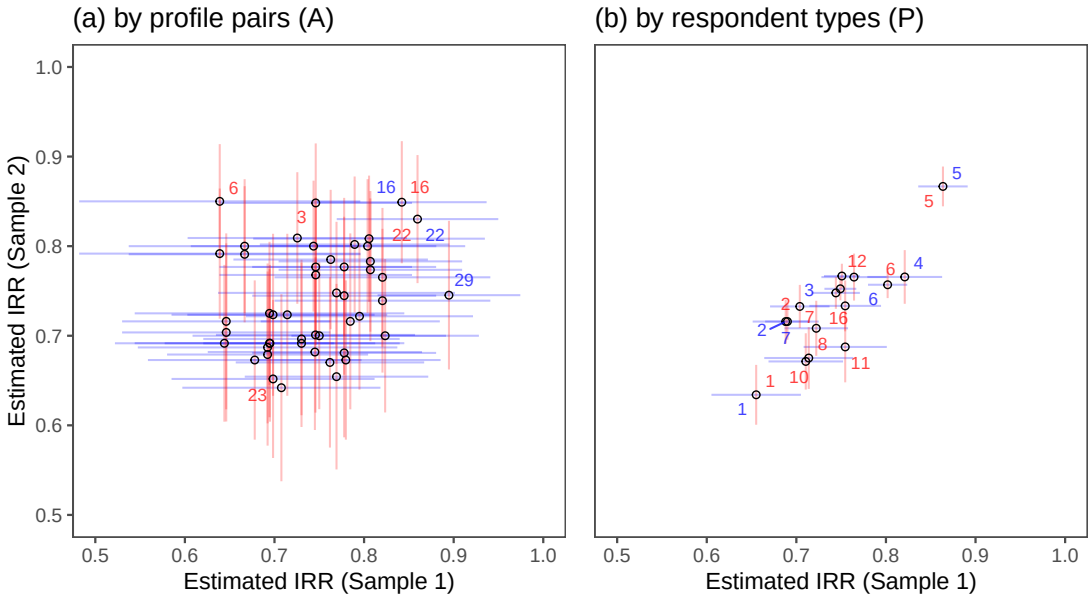

Figure 4: Variation in Intra-Respondent Reliability over (a) Attributes and (b) Personal Characteristics (with a key to the numbers appearing in Table A9 in Supplementary Appendix A7.)

tions.[9] We then have about 50 respondents take this survey for each of our 48 profile-pair combinations (Sample 1). We repeat the entire experiment with 100 responses for each combination (Sample 2). With these samples, we have two reliability estimates for each combination. All other design details of this experiment appear in Supplementary Appendix A6.

These estimates of intra-respondent reliability from our experiment appear in Figure 4, Panel (a). Each point represents one profile-pair combination, with intra-respondent reliability estimated from Sample 1 plotted horizontally and Sample 2 plotted vertically. The mean in each sample is about the same as for our eight replications of published articles (about 75%; see Figure 2). We include 80% confidence intervals (rather than 95% to reduce graphical clutter) in blue for Sample 1 and red for Sample 2. Points that differ from the mean (for each sample) at the 95% level are given a numeric code (blue for Sample 1 and red for Sample 2) so they can be linked back to the specific profile-pair combination (listed in Supplementary Appendix A6).

[9]We fix the first task in each set so that the last task in the first set and the first task in the second set are always different.

Panel (a) of Figure 4 does not reveal any systematic, predictable differences in reliability as a function of the attribute pairs. For example, if reliability estimates differ from the mean only due to sampling error, we expect to see, on average, 2.4 of these 48 points labeled as "significant" at the 95% level. In fact, we see two in Sample 1 and five in Sample 2. Even these are questionable, given that the two samples disagree on the significance of all but two profile pairs (numbers 16 and 22, which appear in both red and blue). Even via post hoc interpretations, we have not been able to ascertain any coherent theory that might account for the specific content of the profiles that turned out to be significant here (see Supplementary Appendix A6). Finally, if the reliability estimates in both samples differed only by random chance, we would expect the samples to correlate at no more than chance levels, which is just what we find: the empirical correlation of the points in the graph is 0.23 with an (insignificant) p-value for a difference from zero of 0.112.

In conclusion, all the evidence on this question seems to point in the same direction: If predictable differences in reliability exist as a function of the profile pairs with randomly assigned attributes, they are unlikely to be large enough to matter substantively. We provide more detail on the experimental design for the replication study in this section, along with the others cited throughout this study, in Section A8.

4.5 Reliability Variation by Personal Characteristics

Finally, we use the same methodology from Panel (a) of Figure 4 to demonstrate that reliability varies systematically over certain characteristics of respondents (P). The results of this analysis appear in Figure 4, Panel (b). As can be clearly seen from all the points labeled with numbers (the key for which appears in Supplementary Appendix A7), most of the effects differ significantly from the mean. The high correlation between the two samples (i.e., 0.85) confirms that the association between respondent types and intra-respondent reliability is indeed systematic. Although these effects vary over studies, we often find (as reported in Supplementary Appendix A7) that younger, minority and male respondents tend to have lower levels of reliability.

These results indicate that assuming constant reliability over attributes is usually justified. However, researchers should use separate reliability estimates for descriptive and

causal analyses that are analyzed within subgroups defined by personal characteristics.

5 Estimating the Degree of Measurement Error

We propose here four methods of estimating the intra-respondent reliability $(1 - \tau)$. Two are for new conjoint projects that work via simple adjustments to the survey design (Section 5.1), while the other two are for analyses of existing conjoint datasets where data new data collection is infeasible (Section 5.2).

5.1 Estimation via New Survey Data

Conjoint studies still in the design stage can be easily modified to estimate τ using one of the following two procedures. The first, which we recommend for most researchers and use in Section 4.2, is to estimate only the average value of τ by adding one extra task at the end of a conjoint survey that repeats the first task but with the order of profiles swapped between left and right. We find no evidence that respondents notice the repetition, which makes this a simple, inexpensive, and widely applicable approach to measuring swapping error.[10]

Estimating the average value of τ is useful for researchers willing to make the assumption we justify empirically in Section 4. Researchers who prefer not to make this assumption can instead choose a second, more detailed procedure, which involves estimating τ_{it} for some or all combinations of i and t. This procedure requires repeating the first procedure for every i and t with enough observations to get a reasonably sized confidence interval.

5.2 Estimation Without Additional Data

We now offer two methods of estimating intra-respondent reliability from a pre-existing conjoint survey without any new data collection or survey design changes. Avoiding new

[10] Although having more than one task is not necessary to apply our methods of bias correction, multiple tasks can increase efficiency without much cost. If a researcher prefers to have only a single task, then a few other survey questions should be used between the pair of repeated questions to estimate intra-respondent reliability. These additional questions ensure respondents do not recall they are being asked the identical question twice.

data collection obviously saves costs, but these methods may be especially useful for datasets where going back to the field may not even be informative because of changes in respondent opinions, choices, or reliabilities.

In most situations, we recommend using both of the following methods. The first approach is to choose a value for τ based on substantively similar studies for which intra-respondent reliability has already been estimated. The existing estimates the researchers can use include one or more of the eight articles we replicate (with values in Figure 2). Uncertain estimates from less similar studies can be studied via sensitivity testing by repeating the bias correction for a range of τ values.

The second approach involves estimating intra-respondent reliability directly from the original survey data. This approach may seem impossible because the survey design includes no repeated tasks. Although ordinary conjoint surveys typically include no task pairs with zero attribute-value differences, we show here that one can accurately extrapolate to this point from pairs of other tasks that differ by varying amounts.[11]

By example, Hankinson (2018), one of the studies we replicate, includes seven attributes for each of the two candidates, meaning that a pair of tasks can differ (for either of the candidates) in the values of 0, 1, 2, 3, 4, 5, 6, or 7 attributes. The unobserved proportion agreement in task pairs with 0 differences is the object of our inference. Because attribute values are assigned randomly and independently, more task pairs with 7 differences will exist than pairs with 1, for example. In fact, in this study, with 30,190 task pairs (i.e., 3,019 observations $\times$ 5 tasks $\times$ 4 $\div$ 2), we only observe pairs with differences of 3, 4, 5, 6, and 7.

In the top left panel of Figure 5, the horizontal axis is the number of attributes that differ within task pairs, and the vertical axis is the percent agreement in candidate choice. For each observed level of difference within pairs, we plot a black dot and confidence interval (although uncertainty is only large enough to see the intervals for the two left dots, representing 3 and 4 attribute-value differences). Next, in this same panel, we plot a weighted least squares regression line fit (in dotted red) to these five data points (at 3, . . . ,

[11] We develop this approach by adapting a method for estimating mortality rates from surveys of people about their siblings; see Gakidou and King 2006.

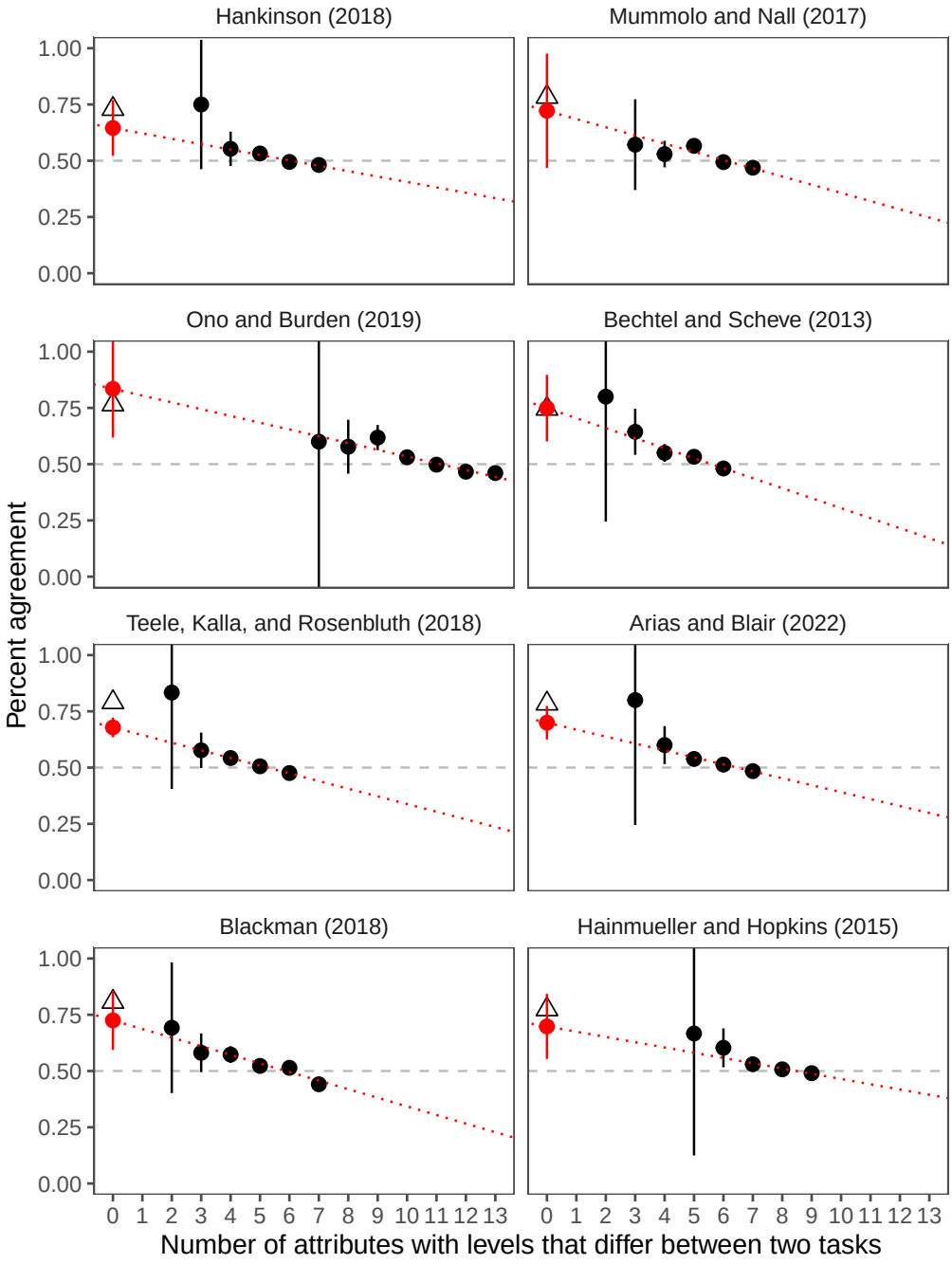

Figure 5: Estimating Intra-Respondent Reliability From Data Without Repeated Tasks. The red dotted line extrapolates the black dots, representing percent agreement conditional on the number of attribute-value differences within task pairs, to the 0 difference point (see the red dot and 95% confidence interval). The black triangle is out-of-sample validation based on a direct estimate with new data, repeated from Figure 2.

7) and use it to extrapolate to 0 on the horizontal axis, the object of our inference.[12]

[12]Weights are calculated from the standard errors of each of the points, which differ because they are based on different numbers of observations.

Our estimate for the percentage agreement when there is no attribute with levels that differ between pairs is the constant term in the regression. We plot this extrapolated estimate of the intra-respondent reliability along with a confidence interval in red in Figure 5. As always with extrapolation, the total uncertainty includes both sampling uncertainty (represented by the red vertical line) and model-based uncertainty, which is not represented but is indicated to some degree by how well the black dots fit the regression (King and Zeng, 2006).

Also in this top left panel is a triangle, which is our direct estimate of intra-respondent reliability based on the repeated task added to our replication study. This estimate serves as an out-of-sample validation for our extrapolated estimate. Remarkably, in this panel, the extrapolation estimate based on no new data (the red dot) and the direct estimate based on a new sample with the repeated task (the triangle) are quite close to each other. We then repeated this sample procedure for all eight of the studies we replicated. For all eight studies, our extrapolated estimate is close to the direct estimate (each of which is in a separate panel in Figure 5.2). This finding may not hold in all future datasets, but it is certainly a promising result.

We note that this procedure can be extended by combining the extrapolation estimate with the estimate from one or more of our replication studies which seem similar to the study being analyzed. We could also extend the procedure with a more fine-grained task pair difference metric, such as by recognizing that some levels are ordered or interval scaled.

6 Finite Sample Properties and Empirical Examples

Section 3 offers estimators for the marginal mean and average marginal component effect corrected for measurement error (see Equations 8). That section shows mathematically that the estimators are unbiased when τ is known and statistically consistent when τ is estimated. Section 4 shores up the key simplifying assumption in these estimators. To complement those analyses, we show first, via Monte Carlo simulation, that the estimators are approximately unbiased even when τ is estimated. Our estimators have slightly

larger standard errors due to the requirement of estimating τ (rather than assuming $\tau = 0$ as in previous studies). We thus also show that the mean square error (a proper combination of bias and variance) is lower for our new corrected estimator, which leads to the conclusion that our bias correction should normally be used. We then show the pattern across estimates from our replications of our corrections decreasing, increasing, and flipping signs of subgroup differences.

6.1 Simulation

We begin with a population of 100,000 individuals with known true preferences, the true marginal mean $\rho(\alpha)$, and AMCE θ. We then generate 1,000 datasets of size $n = 1,000$, each via simple random sampling (with replacement). Next, we add swapping error of sizes $\tau = \{0.1, 0.15, \ldots, 0.4\}$ by using the observation mechanism in Equation 3. Finally, in each simulated dataset, we compute the uncorrected estimates (used throughout the literature) and our alternative corrected estimate for both quantities of interest. Complete details and code necessary to replicate this simulation can be found in our replication package.

We give results for bias and root mean square error (RMSE) in Figure 6, with the marginal mean (MM) in the first column of panels and the AMCE in the second column. In the first row, we present the degree of bias for the uncorrected estimator (measured as deviation from the horizontal dashed line at zero) for each value of τ (the degree of measurement error, on the horizontal axis) and values of the two quantities of interest (in shades of grey, with values indicated in the figure legend). As anticipated by the mathematical results in Section 2, for both the marginal mean and the AMCE, bias increases quickly as measurement error increases, in different amounts depending on the size of the MM and AMCE.

The second row of panels in Figure 6 reveals that, for all combinations of values of τ and for both MM and AMCE, our new estimator is approximately unbiased, which can be seen by all the lines appearing at zero bias (on top of one another and on top of the horizontal dashed line indicating zero bias).

Finally, we compare the difference in RMSE for the uncorrected and corrected esti-

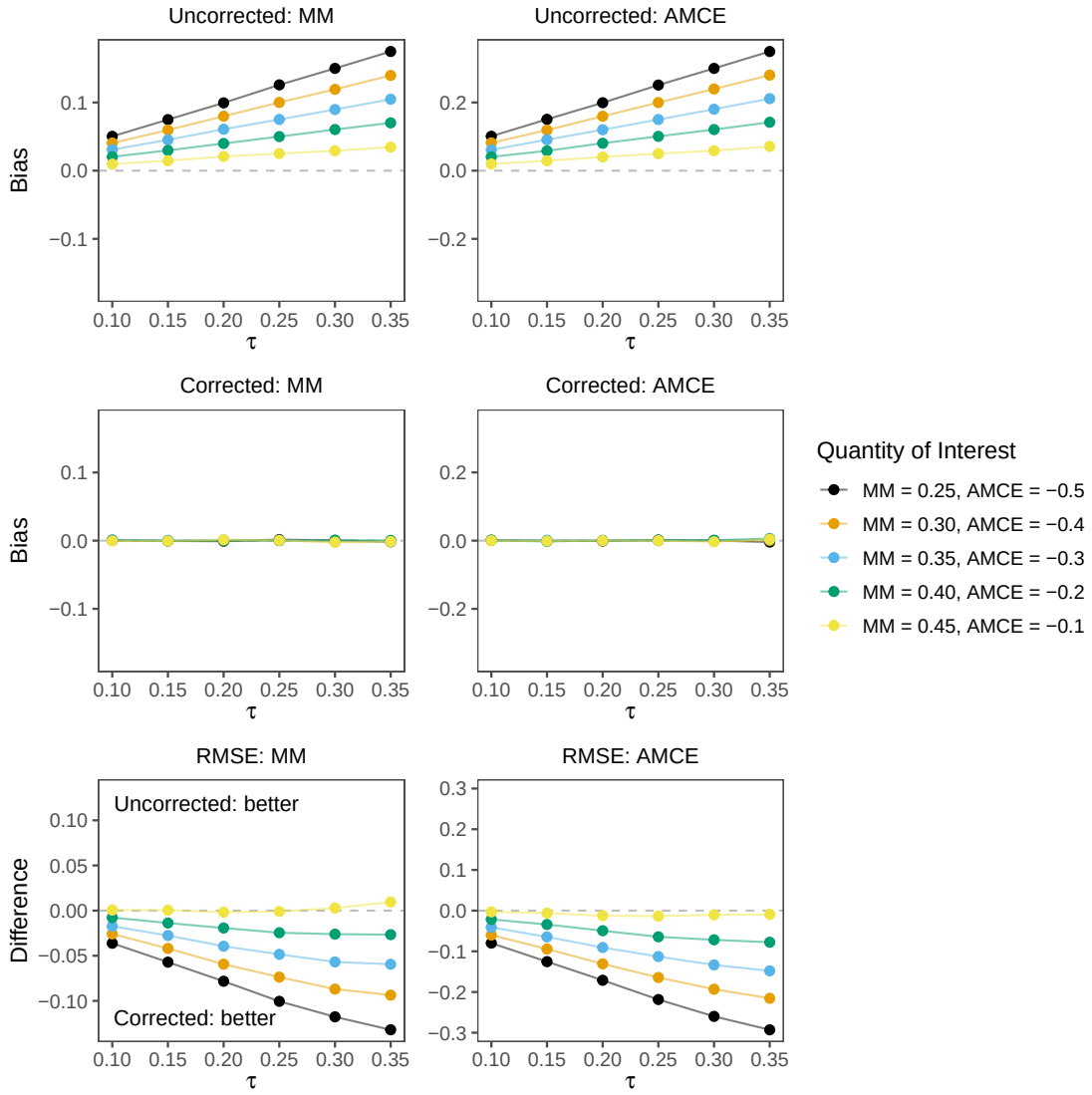

Figure 6: Reducing Bias and Mean Square Error in the marginal mean (MM) and AMCE in Conjoint Experiments

mators in the last row of panels, with the MM on the left and the AMCE on the right. In all cases, the RMCE is lower for our corrected estimator than the uncorrected approach. Every line for all simulations with different quantities of interest (indicated by shades of grey described in the figure legend) appears below the horizontal dashed line indicating no difference. Therefore, correcting bias is always recommended regardless of the degree of measurement error and the expected magnitude of MM and AMCE.

6.2 Empirical Examples

Equation 3 shows that the corrected estimator for the AMCE is always farther from zero than the uncorrected one, and for the MM is always farther from 0.5. However, for differences in MMs or AMCEs among subgroups of respondents (such as comparing AMCEs or MMs for men v. women, young v. old, or with v. without a college degree), the bias correction can increase, decrease, or flip the signs compared to the uncorrected estimate.[13]

Although the only way to ascertain the bias in a new or existing study is to estimate τ and apply our bias correction, we provide here some intuition for what might happen by studying a large number of empirical estimates from our eight replicated studies. To do this, we begin with all seven dichotomous variables used across any of the eight original studies we replicate and then add four additional variables we had available, including whether a respondent used a mobile device or a desktop computer, an end-of-survey question measure of attentiveness, and two variables based on time to complete the survey (above v. below the median, and the top v. bottom quartiles). With these variables, across the eight studies, we estimated uncorrected and corrected estimators for 1,870 AMCEs and 2,552 MMs.

Each uncorrected subgroup difference has an arbitrary sign, based on which subgroup comes first in the difference. We resolve this ambiguity in the present analysis by always subtracting the smaller estimate from the larger one, making all uncorrected estimates positive. These values are plotted on the horizontal axis in each panel of Figure 7 (which thus begins at zero on the left). The left panel gives AMCE estimates and the right panel plots MM estimates. The vertical axis in both panels is the bias-corrected subgroup difference, which of course can be positive or negative. We have also color coded (and separated by dashed red lines) the three resulting effects of the corrections. For both AMCE and MM, we find that the bias correction increases the subgroup difference effect for about 80% of the estimates, decreases it in about 10%, and switches the sign in about 10%. The size of the effect in each category has a wide range relative to the size of the original estimate.

If the next study to be conducted is like these eight, then we might expect that correct-

[13] On on subgroup (or "heterogeneous treatment") effects in conjoint studies, see Goplerud, Imai, and Pashley (2022), Leeper, Hobolt, and Tilley (2020), and Clayton, Ferwerda, and Horiuchi (2021).

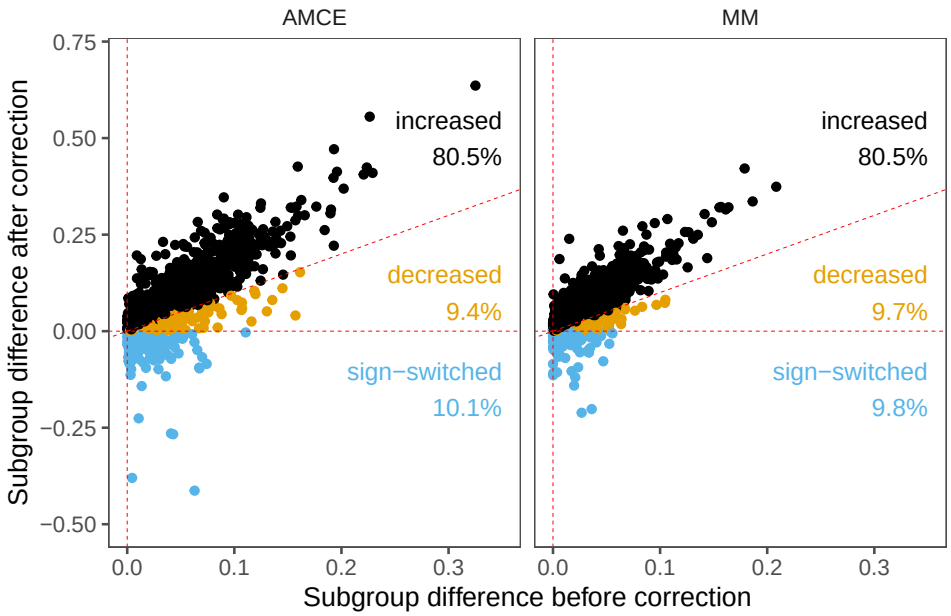

Figure 7: Consequences of Bias Correction in 8 Studies. The horizontal axis is the positive difference between the two subgroups, and the vertical axis is the corrected value for AMCE (left panel) and MM (right panel).

ing the bias will increase the subgroup difference most of the time. However, although this figure gives some sense of what may happen in real examples, the 4,422 estimates across the two panels do not represent a probability distribution from which any new study will be drawn from. The only way to know the direction of the bias, and therefore the effect of the correction, is to follow the advice in this study, estimate τ, and make the correction.

7 Best Practices For Conjoint Analyses

Based on the research presented here, we make some practical recommendations. Specifically, researchers planning conjoint survey analysis should consider four points, roughly in this order.

First, survey experiments can be greatly improved with conjoint designs, but only if appropriately corrected for measurement error. Although hypothetical conjoint experiments without measurement error are more efficient than one-shot randomized experiments, conjoint designs introduce more measurement error than traditional survey questions. This causes some of the apparent efficiency advantages to be lost and potentially

substantial bias to be induced. Unless applied research is to rely on sheer luck, the bias cannot be ignored. Particularly when researchers are interested in subgroup comparison, the bias may attenuate, exaggerate, or flip signs of the differences in MMs or AMCEs.

Second, measurement errors should be reduced in the design phase, where possible, by following best practices in standard survey design. Most important among these is the so-called "cognitive debriefing," where researchers administer a draft survey to a small sample of respondents and immediately go to the start of the survey and ask the same respondents how they understood each question. Researchers should repeat this procedure while continuously adjusting each question's wording, perhaps multiple times. Conjoint analyses are more complicated to understand than traditional survey questions, making this standard advice even more important. Even though we find no evidence for systematic variation in intra-respondent reliability as a function of the complexity, divergence, or inconsistency of the conjoint attributes (see Section 4), researchers should consider conjoint designs at least as carefully as well run traditional survey instruments.

Third, although ratings, rankings, binary choice, and other designs have been used in the broader conjoint literature, only binary choice has been widely used in the political science literature. Some conjoint studies, including three of the eight we replicate, immediately follow a binary choice question with rankings questions for each profile, often on the same page; these, of course, are not independent measures (and which our data analyses confirm). Thus, we recommend that researchers who wish to use conjoint designs other than binary choice should conduct measurement error studies, perhaps analogous to the strategy followed here, and validate their measures in other ways. There is much potential for future methodological research in this area, but it will be easier for researchers planning applied studies to stick for now with the binary choice conjoint design.

Finally, measurement error bias can easily be corrected, and mean squared error can be reduced, by estimating intra-respondent reliability and applying the simple correction methods to causal effect and marginal mean estimates (see Equation 8). We suggest (in Section 5) that researchers choose among four approaches to estimating intra-respondent reliability (ordered by the simplicity of application):

1. If your research topic is similar enough to one or more of the studies we replicate — in content and target population — use the corresponding estimate of intra-respondent reliability from Figure 2. Because the estimates in this figure (and others we estimate) do not vary much, choosing the wrong one may not be very consequential, but one should be clear about the implied assumption. Of course, this approach is only applicable in studies of subgroup effects if researchers have some prior information about subgroup differences in intra-respondent reliability.

2. You can estimate intra-respondent reliability from an existing conjoint without new data collection by extrapolating patterns in the existing data, as we show in Section 5.2. An estimate from this method can also be combined with the first option if one of the studies we replicated is similar to the one you are analyzing.

3. If you are in the planning stage of a conjoint study, we strongly recommend adding a repeat of the first task presented at the end (and with the two profiles with the order of the two columns switched). This enables researchers to estimate intra-respondent reliability by simply computing the percent agreement between the first and last questions and averaging over all respondents or the relevant subgroup. To use this direct and robust estimate to correct bias, the researcher would rely on the extensive empirical evidence we offer that intra-respondent reliability does not differ systematically over information contained in conjoint tables. This assumption, although far less restrictive than the assumption needed for the first approach, should still be noted.

4. Researchers may choose to estimate the level of intra-respondent reliability for every profile pair, as we did for Figure 4, Panel (a). This makes the assumptions from the first and second approaches unnecessary. The cost of this approach, however, is the requirement to collect a substantially larger number of observations.

Although we recommend that most researchers adopt the third strategy, these researchers can still check whether intra-respondent reliability varies over selected types of profile-pair combinations by grouping them in different ways.

8 Concluding Remarks

Through empirical, theoretical, and simulation-based evidence, we show that measurement error in conjoint designs can create substantial bias in estimates of descriptive and causal effects — on average, within subgroups, and for subgroup differences. Although we find that measurement error can lead to attenuation, exaggeration, or sign switches, we show that the error tends to have common empirical patterns for binary choice conjoint designs. We then use these patterns to develop easy-to-use methods to correct the resulting biases. These bias corrections will often make effects larger, but not in all situations; fortunately, our corrections are easy to apply and so researchers can use them and see for themselves.

Our approach applies only to the most common type of conjoint design, with a binary choice outcome variable. Future research should study the types of measurement error, consequent biases, and possible corrections in alternative conjoint designs, such as multiple choice outcomes, ratings, rankings, and others. The additional cognitive demands these alternative conjoint designs place on respondents suggest that they would lead to even higher levels, and more complicated forms, of measurement error than for binary choice outcomes. This may make corrections more difficult but, without this additional work, using these alternative conjoint designs would put a researcher's results and conclusions at considerable risk.

Appendix A Standard Errors

We now show how to compute standard errors for $\tilde{\rho}(\alpha)$ and $\tilde{\theta}$ in three ways — analytical derivation for speed, bootstrapping for convenience, and 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 $\tilde{\rho}(\alpha)$ and $\tilde{\theta}$ as in Equation 8. 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 $\rho(\alpha)$ and τ from a bivariate normal (given estimates of parameter values from our analytical derivation below), plug them into Equation 8, 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 8). 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 afterwards apply it to our problem. For two correlated random variables R and S, we approximate a ratio R/S as

$$V(R/S) \approx \frac{E(R)^2}{E(S)^2}\left(\frac{V(R)}{E(R)^2} - 2\frac{\text{Cov}(R,S)}{E(R)E(S)} + \frac{V(S)}{E(S)^2}\right). \tag{9}$$

We now apply the approximation in Equation 9 to the AMCE, $\tilde{\theta} = \hat{\theta}/(1-2\hat{\theta})$ from Equation 8. We first compute the moments: $E(\hat{\theta}) = \theta(1-2\tau)$, $E(1-2\hat{\tau}) = 1-2\tau$, $V(\hat{\theta}) \equiv \sigma_\theta^2$, and $V(1-2\hat{\tau}) = 4\sigma_\tau^2$, where $V(\hat{\tau}) \equiv \sigma_\tau^2$. We will also need the covariance, $\text{Cov}(\hat{\theta}, \hat{\tau}) = \text{Cov}(\hat{\rho}(\alpha), \hat{\tau}) - \text{Cov}(\hat{\rho}(\alpha'), \hat{\tau})$, where, letting $d_i = \mathbf{1}(C_{i1} \neq C_{iT})$ equal 1 for disagreement and 0 agreement on the same item asked twice,

$$\begin{aligned}
\phi_\alpha \equiv \text{Cov}(\hat{\rho}(\alpha), \hat{\tau}) &= \text{Cov}\left(\frac{1}{n_\alpha}\sum_{it|\ell=\alpha} C_{it}, \frac{1}{n}\sum_i d_i\right) \\
&= \frac{1}{n_\alpha}\sum_{it|\ell=\alpha}\frac{1}{n}\sum_i \text{Cov}(C_{it}, d_i) \\
&= \frac{1}{n_\alpha n}\sum_{it|\ell=\alpha}\text{Cov}(C_{it}, d_i) \\
&= \frac{1}{n}\text{Cov}(C_{it}, d_i),
\end{aligned}$$

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_α is the number of observations in the treated group.

We then compute the variance by applying Equation 9:

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

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

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

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

We compute the variance of the marginal mean by applying Equation 9:

$$V(\tilde{\rho}(\alpha)) \approx \frac{\rho(\alpha)^2}{(1-2\tau)^2} \left(\frac{\sigma^2_{\hat{\rho}} + \sigma^2_{\hat{\tau}} - 2\phi_\alpha}{\rho(\alpha)^2} + 4\frac{\phi_\alpha - \sigma^2_{\hat{\tau}}}{\rho(\alpha)} + 4\sigma^2_{\hat{\tau}} \right),$$

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

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

Finally, we conduct a Monte Carlo experiment to show how the different methods perform. As an illustration, we set $\rho(\alpha) = 0.35$, $\rho(\alpha') = 0.65$, $\tau = 0.25$, and $n = 1,000$. We generate 3,000 datasets, using 1,000 draws for both the bootstrapping and simulation methods. Figure 8 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 $\tilde{\theta}$ and $\tilde{\rho}(\alpha)$) 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 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).

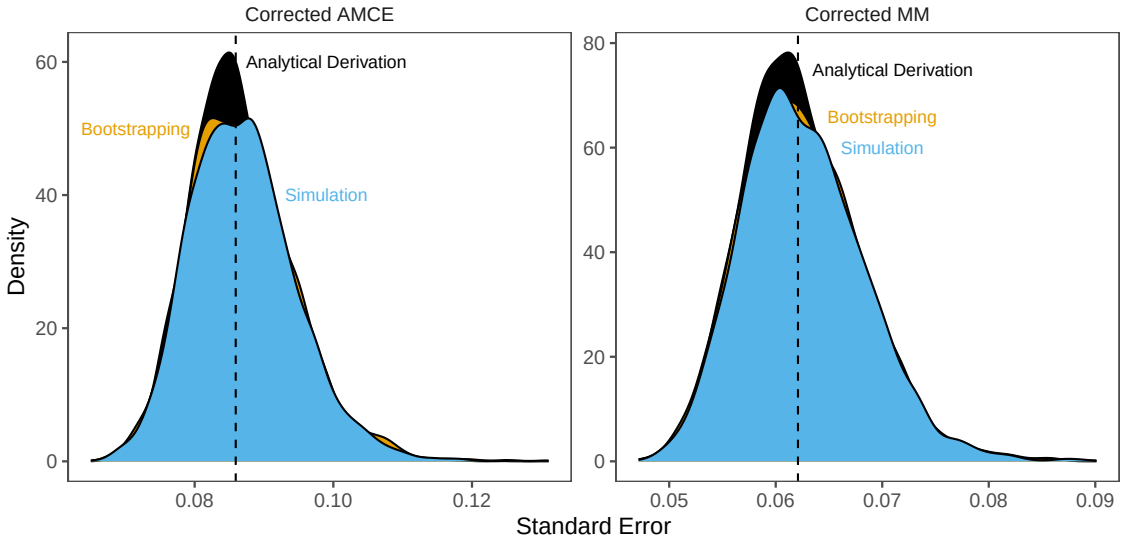

Figure 8: Standard Errors. Histograms from a Monte Carlo experiment of 3,000 standard error estimates (for AMCE in the left panel and MM in the right panel) from boostrapping (orange), simulation (blue), and our analytical derivation (black). The true standard error is portrayed as a vertical dashed line in each figure.

References

Abramson, Scott F., Korhan Koçak, and Asya Magazinnik (2022). "What do we learn about voter preferences from conjoint experiments?" In: *American Journal of Political Science* 66.4, pp. 1008–1020.

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.

Auerbach, Adam Michael and Tariq Thachil (2018). "How Clients Select Brokers: Competition and Choice in India's Slums". In: *American Political Science Review* 112.4, pp. 775–791.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto (2018). "The Number of Choice Tasks and Survey Satisficing in Conjoint Experiments". In: *Political Analysis* 26.1, pp. 112–119.

— (2021a). "Beyond the Breaking Point? Survey Satisficing in Conjoint Experiments". In: *Political Science Research and Methods* 9.1, pp. 53–71.

— (2021b). "Conjoint survey experiments". In: *Advances in Experimental Political Science*. Ed. by James N. Druckman and Donald P. Green. New York: Cambridge University Press, pp. 19–41.

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.

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

Blackwell, Matthew, James Honaker, and Gary King (2017). "A Unified Approach to Measurement Error and Missing Data: Overview". In: *Sociological Methods and Research* 46.3, pp. 303–341.

Bradburn, Norman M., Seymour Sudman, and Brian Wansink (2004). *Asking questions: The definitive guide to questionnaire design–for market research, political polls, and social and health questionnaires*. San Francisco: Jossey-Bass.

Bryan, Stirling, Lisa Gold, Rob Sheldon, and Martin Buxton (2000). "Preference measurement using conjoint methods: an empirical investigation of reliability". In: *Health Economics* 9.5, pp. 385–395.

Clayton, Katherine, Jeremy Ferwerda, and Yusaku Horiuchi (2021). "Exposure to Immigration and Admission Preferences: Evidence from France". In: *Political Behavior* 43.1, pp. 175–200.

Coppock, Alexander and Oliver A. McClellan (2019). "Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents". In: *Research & Politics* 6.1, pp. 1–14.

Dafoe, Allan, Baobao Zhang, and Devin Caughey (2018). "Information equivalence in survey experiments". In: *Political Analysis* 26.4, pp. 399–416.

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.

Gakidou, Emmanuela and Gary King (2006). "Death by survey: estimating adult mortality without selection bias from sibling survival data". In: *Demography* 43.3, pp. 569–585.

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

Gilbert, Daniel, Gary King, Stephen Pettigrew, and Timothy Wilson (2016). "Comment on 'Estimating the Reproducibility of Psychological Science'". In: *Science* 351.6277, 1037a–1038a. URL: https://j.mp/openrepl.

Goplerud, Max, Kosuke Imai, and Nicole E. Pashley (2022). "Estimating Heterogeneous Causal Effects of High-Dimensional Treatments: Application to Conjoint Analysis". Unpublished manuscript. Available at: https://arxiv.org/abs/2201.01357.

Green, Paul E. and Venkatachary Srinivasan (1978). "Conjoint Analysis in Consumer Research: Issues and Outlook". In: *Journal of Consumer Research* 5.2, pp. 103–123.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto (2015). "Validating Vignette and Conjoint Survey Experiments Against Real-World Behavior". In: *Proceedings of the National Academy of Sciences* 112.8, pp. 2395–2400.

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.

Horiuchi, Yusaku, Zachary Markovich, and Teppei Yamamoto (2022). "Does conjoint analysis mitigate social desirability bias?" In: *Political Analysis* 30.4, pp. 535–549.

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.

King, Gary (1995). "Replication, Replication". In: *PS: Political Science and Politics* 28.3, pp. 443–499.

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.

King, Gary and Langche Zeng (2006). "The Dangers of Extreme Counterfactuals". In: *Political Analysis* 14.2, pp. 131–159. URL: j.mp/dangerEC.

Krosnick, Jon A. (1999). "Maximizing questionnaire quality". In: *Measures of Political Attitudes*. Ed. by John P. Robinson, Phillip R. Shaver, and Lawrence S. Wrightsman. New York: Academic Press, pp. 37–58.

Leeper, Thomas J, Sara B. Hobolt, and James Tilley (2020). "Measuring subgroup preferences in conjoint experiments". In: *Political Analysis* 28.2, pp. 207–221.

Liu, Guoer and Yuki Shiraito (2022). "Multiple Hypothesis Testing in Conjoint Analysis". Forthcoming, *Political Analysis.*

McCullough, James and Roger Best (1979). "Conjoint measurement: temporal stability and structural reliability". In: *Journal of Marketing Research* 16.1, pp. 26–31.

Mørkbak, Morten Raun and Søren Bøye Olsen (2015). "A within-sample investigation of test–retest reliability in choice experiment surveys with real economic incentives". In: *Australian Journal of Agricultural and Resource Economics* 59.3, pp. 375–392.

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.

Open Science Collaboration (2015). "Estimating the reproducibility of psychological science". In: *Science* 349.6251, pp. 943–952.

Payne, Stanley Le Baron (2014). *The Art of Asking Questions*. Princeton: Princeton University Press.

Shamir, Michal and Jacob Shamir (1995). "Competing values in public opinion: A conjoint analysis". In: *Political Behavior* 17, pp. 107–133.

Skjoldborg, Ulla Slothuus, Jørgen Lauridsen, and Peter Junker (2009). "Reliability of the discrete choice experiment at the input and output level in patients with rheumatoid arthritis". In: *Value in Health* 12.1, pp. 153–158.

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.

Zhirkov, Kirill (2022). "Estimating and using individual marginal component effects from conjoint experiments". In: *Political Analysis* 30.2, pp. 236–249.

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

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

March 28, 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 Review of Existing Conjoint Studies

We replicate the eight conjoint studies in political science listed in Table A1, but the method has been used by many others in our discipline and beyond. Systematic reviews of conjoint applications in political science in separate literature reviews includeDe 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) analyzes 67 candidate-related conjoint experiments that includes a gender attribute, Eshima and Smith (2022) analyzes 16 candidate conjoint experiments that includes 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) records 84 studies evaluating preferences for environmental policy and Mamine, Minviel, et al. (2020) lists 70 studies related to agri-environmental practices; in marketing, Bastounis et al. (2021) analyzes 43 conjoint experiments manipulating sustainability labeling on food products.

Across all fields, a search for "conjoint analysis" in Google Scholar returns 98,300 articles (accessed 1/11/2022).

A2 Study Selection

This Appendix provides details of how we selected studies to replicate; it supplements 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 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 were transparent enough to show a pair of conjoint

tables in a tabular format 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, 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 A10 for more details on these studies).

As we continued our analyses, we moved to fully replicating 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).[1]

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 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*

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

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 used Mummolo, 2016 for another replication study (see 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. Ultimately, 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 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	13	10
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

A3 Attentiveness and Intra-Respondent Reliability

One hypothesis for what explains variation in IRR is respondent attentiveness. Attentiveness varies widely across online samples, and so one hypothesis is that respondents who were not paying close attention to the conjoint task may have been less likely to select the same profile at two different points in time than respondents who were paying attention. We can evaluate this possibility by comparing IRR among more and less attentive respondents in our samples.

We used a variety of different sample providers and response quality indicators in all of the studies reported in Table A10.[2] 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.[3] In studies conducted on respondents from DLABSS (see `dlabss.harvard.edu`), we do not include a pre-task attention check but assuming a generally high response quality is reasonable given that respondents are volunteers who do not receive compensation for their participation and likely share some connection to the research community at Harvard University. 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.

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

[2] Our first small pilot, "12 replications v1.2" in Table A10, is omitted.

[3] This is the method 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.

- 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" or "rarely" provide humorous or insincere responses to survey questions are coded as "sincere." Respondents who said that they "sometimes," "most of the time," or "always" provide humorous or insincere responses to survey questions are coded as "insincere."*

A4 Traditional Survey Questions vs. Conjoint

This appendix supplements information in Section 4.3 in the paper. 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, is 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 immigrants 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:
  - Partisanship: 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.

A5 The Top-Down Approach

We now 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 the hypotheses along the way. What follows is details about a sequence of studies we conducted (and corresponding tables or figures with results). All 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 5 of the evaluation tasks had different levels of consistency, 5 varied in complexity, and 5 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 A2, 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 A2: 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 A3 and the most complex would look like Table A4.

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 A5 depicts a sample of the most (and least) divergent housing developments respondents

Table A3: 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

Table A4: 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.

were asked to evaluate. Values in parentheses depict the *least* divergent profiles.

Figure A1 summarizes our results for consistency, complexity, and divergence in the

Table A5: 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

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 not account for the vast majority of observed IRR: Average IRR among the most consistent profiles is just 6% greater than IRR among the least consistent, and the 95% confidence intervals overlap across the range of profiles.

We then pursued the small consistency result by designing an even more extreme experiment, to see how far as we could take this result. We still find that consistency does not drive IRR noticeably, even in a clearly extreme case, where 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 the experiment depicted in

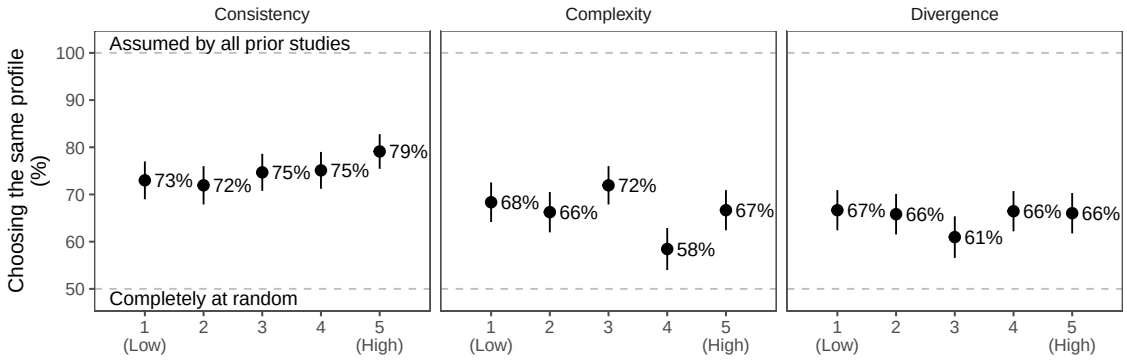

Figure A1: Relationship between IRR and profile consistency, complexity, or divergence.

Table A2. 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 A2 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.

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

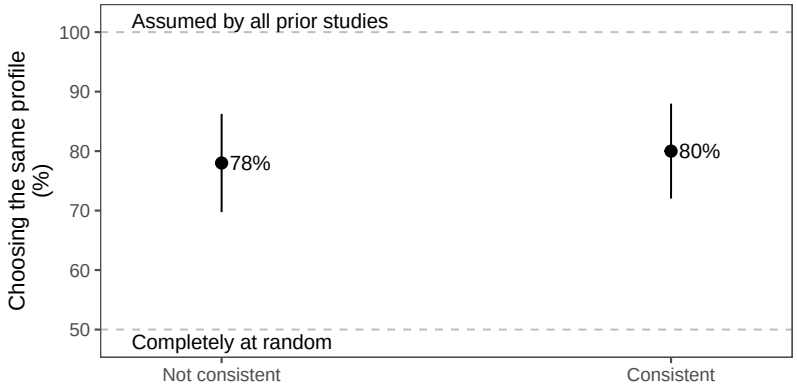

Figure A2: Relationship between IRR and profile consistency: simplest case with party affiliation and tax policy views.

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 overwhelmingly 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 into 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 A6, while the least consistent versions would have appeared to respondents following Table A7.

The results show that logical consistency only seems to influence IRR in the most extreme and unrealistic cases. Figure A3 summarizes these results. There are four at-

Table A6: 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 A7: 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

tributes (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, 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 are asked to review the exact same profiles in a subsequent wave one week later. Figure A3 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 A1 and A2. 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.

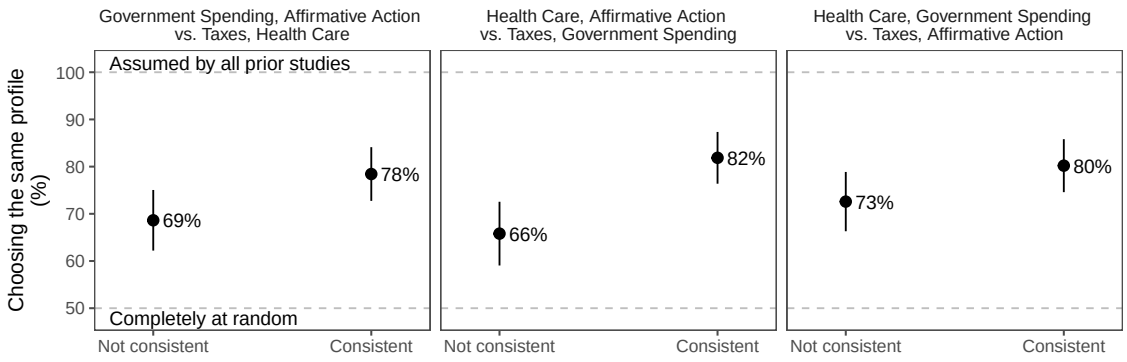

Figure A3: Relationship between IRR and profile consistency: nonpartisan. Leftmost panel represents average % of respondents choosing same profile twice when positions on government spending and affirmative action are flipped to generate inconsistent profiles; central panel shows flipped health care and affirmative action positions; 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.

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 431 participants via Prolific and had them evaluate eight candidate choice tasks twice. The candidates that respondents could choose between are defined by their partisanship, age, race, gender, alma mater, and salient personal characteristics. Figure A4 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 A4 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 conjoints precisely to learn how respondents make choices in the presence of a variety of important attributes.

A6 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. For this study, 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

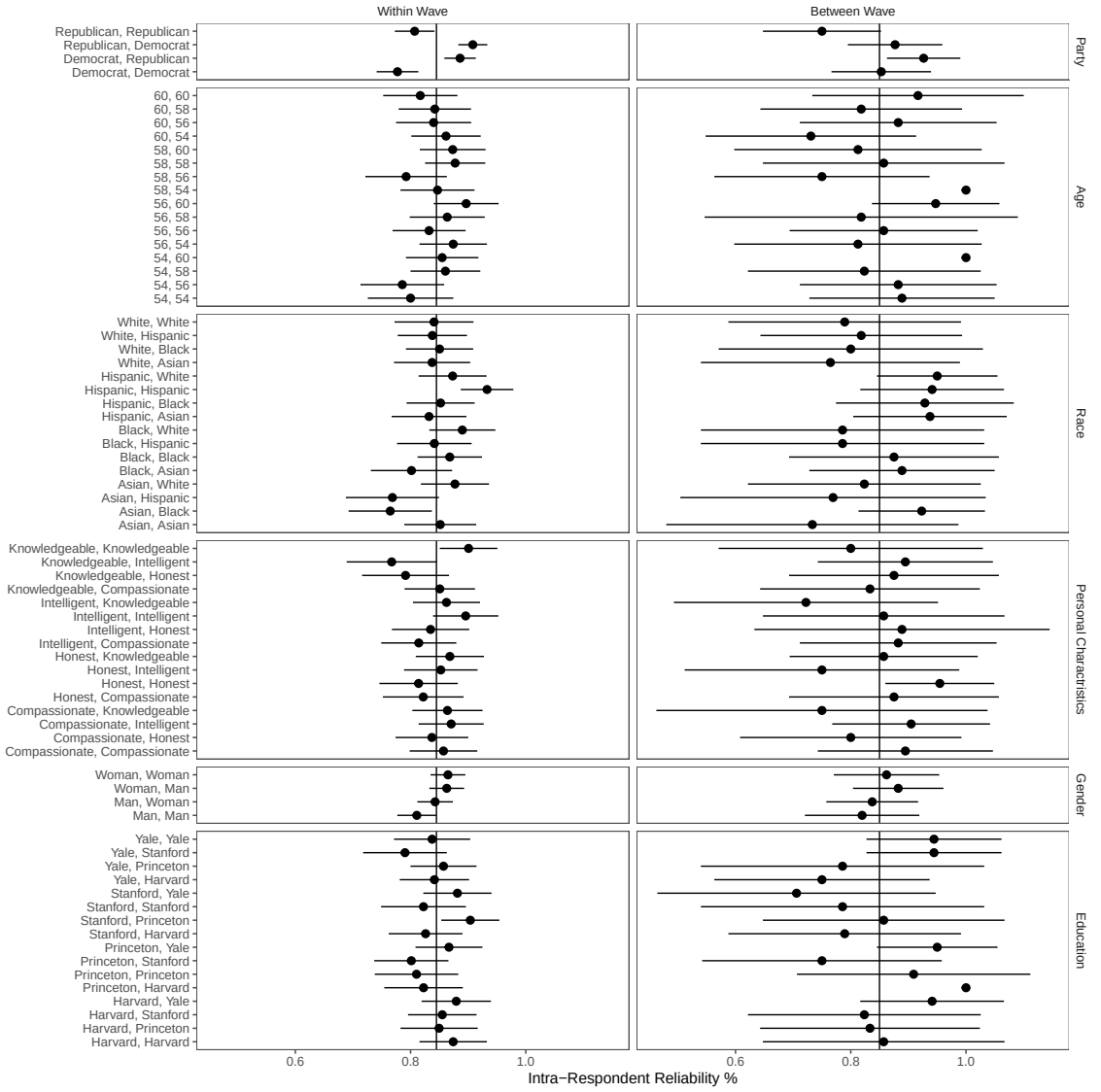

Figure A4: Relationship between IRR and profile pair characteristics within wave (left) and between waves (right). Vertical lines represent the overall IRR for the replication (88%).

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 A5 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 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 A5) and Prolific respondents (Figure A7) 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 A6 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, 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

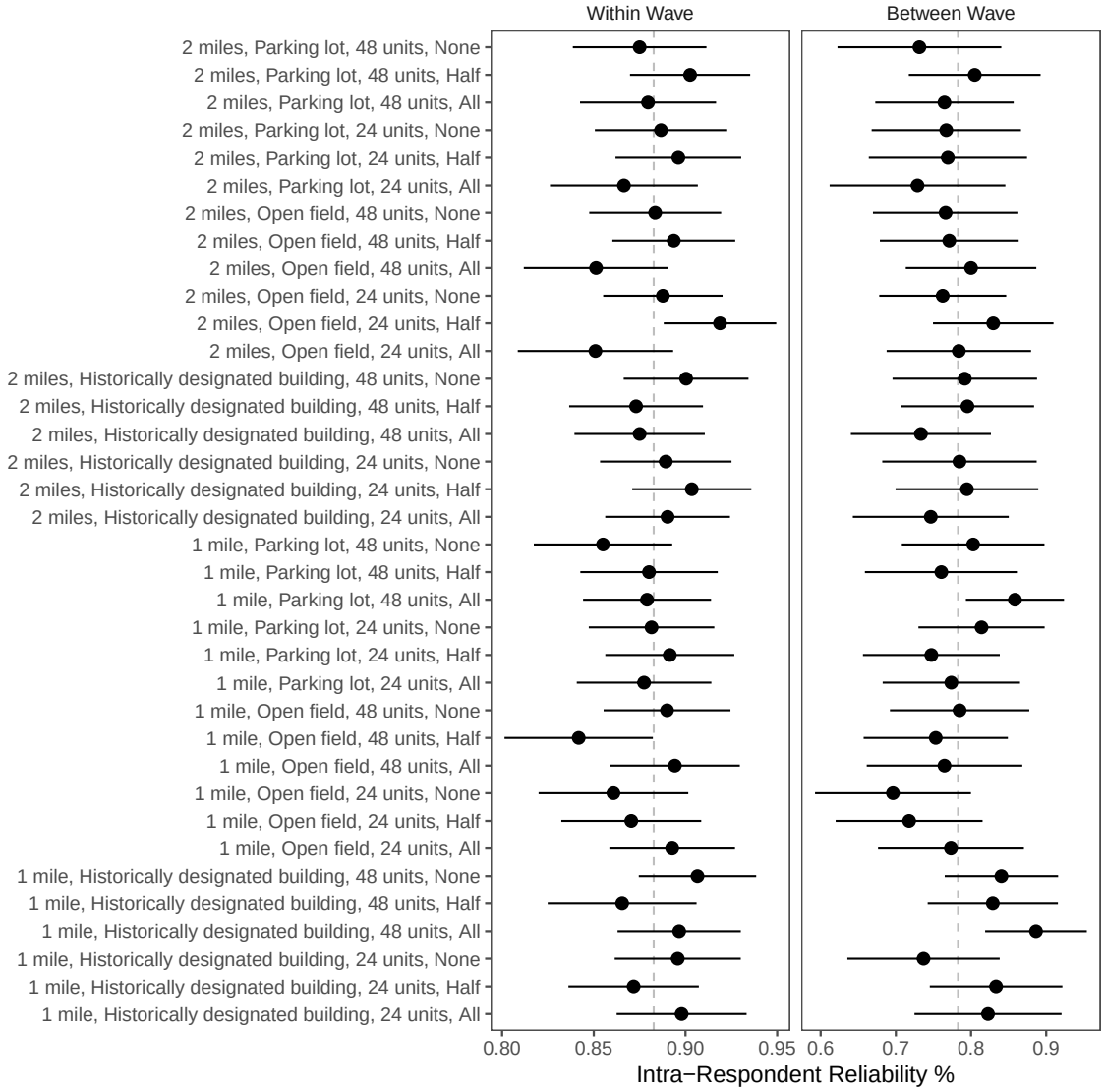

Figure A5: IRR for repeated tasks within-wave (left) and between wave (right) in a DLABSS replication of Hankinson (2018). Rows represent possible combinations of attributes visible to respondents in the study. Dashed vertical lines represent the overall mean IRR within each wave (left) and between waves (right).

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 A8, which shows the estimates of the correlation between each profile pair and IRR in both waves of the study along with standard errors

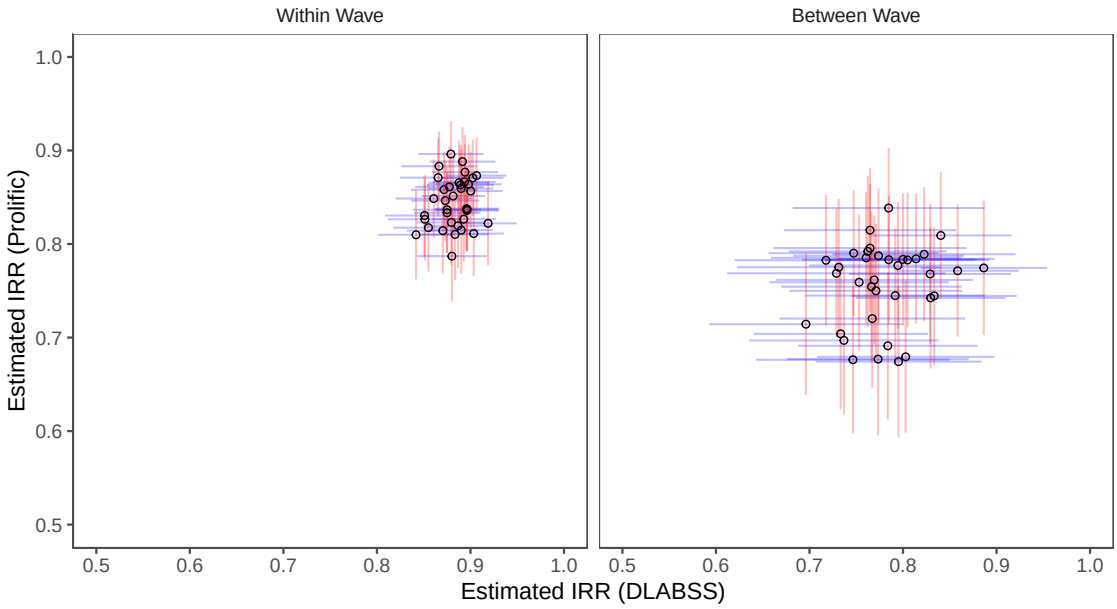

Figure A6: Correlations between estimated relationships between specific combinations of attributes in profile-pair comparisons evaluated by DLABSS and Prolific respondents within wave (left) and between waves (right). Blue segments represent confidence intervals associated with estimates from the DLABSS study, while red segments represent confidence intervals associated with estimates from the Prolific Study.

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.

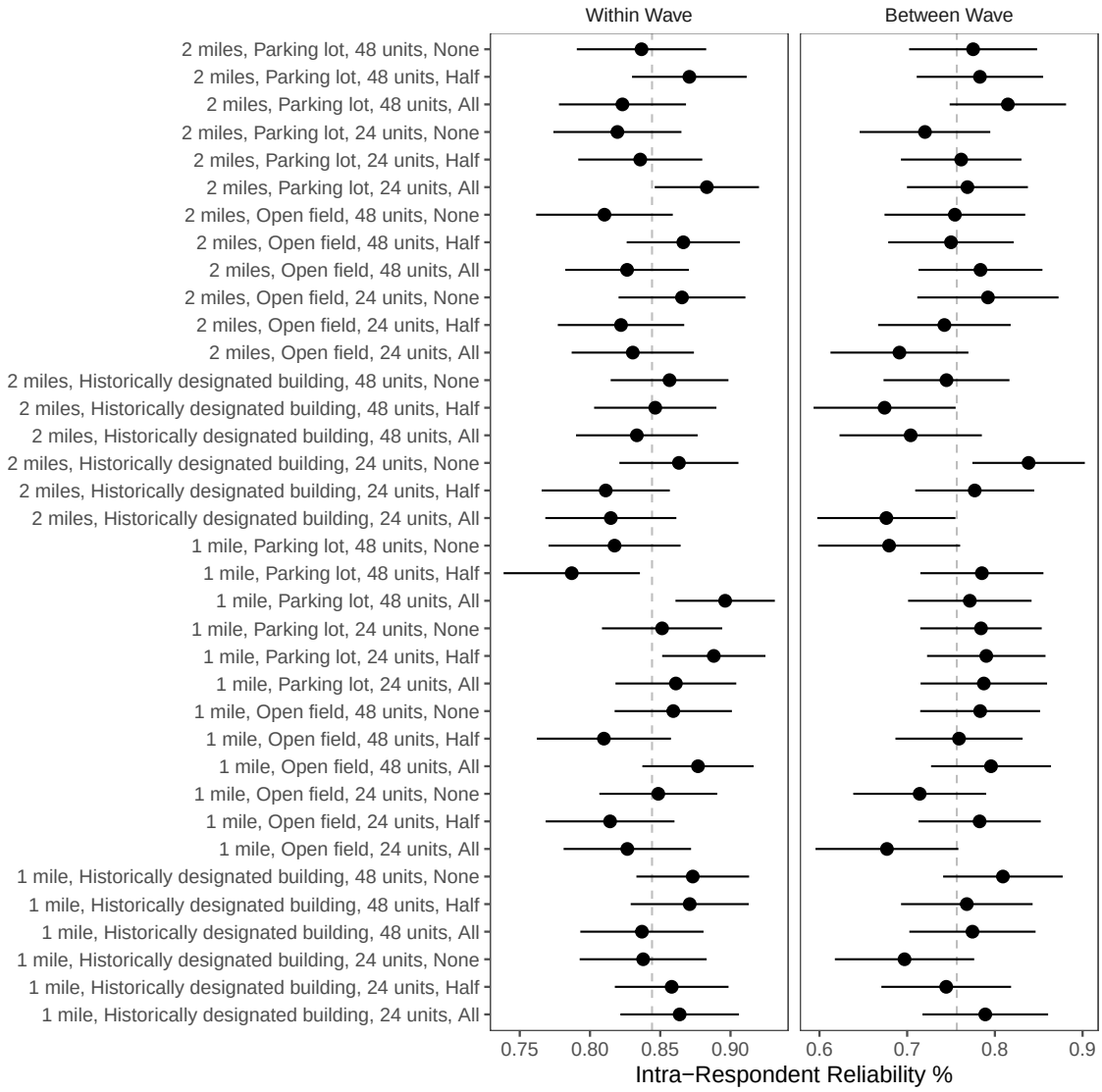

Figure A7: IRR for repeated tasks within-wave (left) and between wave (right) in a Prolific replication of Hankinson (2018). Rows represent possible combinations of attributes visible to respondents in the study. Dashed vertical lines represent the overall mean IRR within each wave (left) and between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

Table A8: 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

A7 Respondent Characteristics

Table A9 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 A9: 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

A8 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 A10 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 A10: 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, 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 Theorem	Media sources	10/27/2022	11/2/2022	https://osf.io/f26am	Within wave	2,641	12	31,692
Total							**9,338**	**220**	**136,982**

A9 Profile Order Flipping for Repeated Conjoint Table

When researchers use our repeated-task approach to obtain an estimate of the average IRR in their study, 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, that 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." And 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 never did run into).

Nevertheless, we conducted three studies to examine whether our estimate of IRR varies by whether the order of profiles in the repeated task was flipped or not. Specifically, 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 average IRR for same-order profiles and for flipped-order profiles. The results are presented in Figure A8. 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, unlikely to change the substantive meaning of our bias corrected estimates.

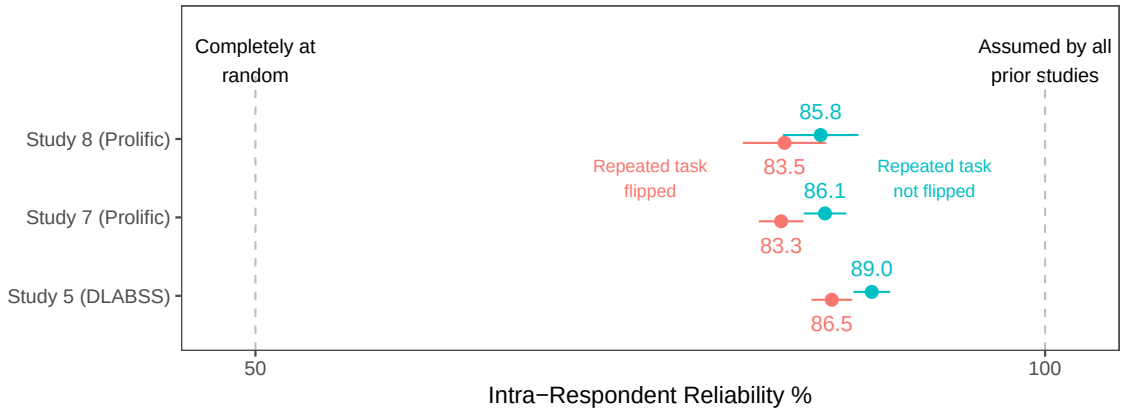

Figure A8: Relationship between IRR and whether repeated task profile order was flipped

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.

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.

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.

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.

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.

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.

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, 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.

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.