Correcting Measurement Error Bias in Conjoint Survey Experiments*

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

March 4, 2026

Keywords: Conjoint experiments, measurement error, survey bias

*This paper and its accompanying Supplementary Appendix are available at GaryKing.org/conjointE. We thank Sabrina Arias, Kirk Bansak, Michael Bechtel, Barry Burden, John Cho, Naoki Egami, Jens Hainmueller, Toby Heinrich, Dan Hopkins, Kosuke Imai, Josh Kalla, Bob Kubinec, Kasey Rhee, Anton Strezhnev, Dawn Teele, Dustin Tingley, Teppei Yamamoto, Joonseok Yang, and the participants at the 2020 and 2023 Meetings 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, a methods reading group at American University in March 2023, and the MENA and Asian Political Methodology in January 2024 for helpful suggestions.

†Department of Political Science, Stanford University. kpc14@stanford.edu, kpclayton.com

‡Department of Government and Program in Quantitative Social Science, Dartmouth College. yusaku.horiuchi@dartmouth.edu, horiuchi.org

§Division of Social Sciences, New York University Abu Dhabi. aaronkaufman@nyu.edu, aaronrkaufman.com

¶(Corresponding author) 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

Abstract

Conjoint survey designs are spreading across the social sciences due to their unusual capacity to estimate many causal effects from a single randomized experiment. Unfortunately, by their ability to mirror complicated real-world choices, these designs often generate substantial measurement error and thus bias. We replicate both the data collection and analysis from eight prominent conjoint studies, all of which closely reproduce published results, and reveal high levels of measurement error in all. We then discover a common empirical pattern in how measurement error appears in conjoint studies and, with it, introduce an easy-to-use statistical method to correct the bias. Along the way, we provide a much simpler and simultaneously more powerful approach to designing, organizing, understanding, and analyzing conjoint data analyses.

The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at https://doi.org/10.7910/DVN/NJ0X.

The Cornell Center for Social Sciences verified that the data and replication code submitted to the AJPS Dataverse replicates the numerical results reported in the main text of this article.

Word Count: 10,301

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, under certain assumptions, 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 designs are used in about 14,000 surveys a year. Allenby, Hardt, and Rossi, 2019. have been the subject of over 140,000 articles across academia and marketing (according to Google Scholar), and have rapidly increased in popularity in the social sciences (see Supplementary Appendix A, p. 1).

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" of the profiles. (Researchers also often ask each respondent to complete several randomly assigned conjoint questions, which we call "tasks," to increase statistical power further.) Modern conjoint estimators, which add no modeling assumptions, are unbiased for specific types of means and causal effects that we clarify below.[1]

In real-world choices about everything from voting, donating, protesting, and debating politics to shopping, parenting, and socializing, people compare a range of attributes between options and make their decisions. Furthermore, these choices typically involve complicated trade offs (indeed, conjoint analysis is often called "trade off analysis" in the marketing literature). If you prefer a candidate except for their policy position on international trade, a car except for its price, or a potential romantic partner except for their inability to stop talking about political science research during dinner, you have a difficult

[1]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 the number of attributes k (and tasks t) increase. Bansak, Hainmueller, et al., 2018; Bansak, Hainmueller, et al., 2021; Jenke et al., 2021. Recent methodological advances clarify the conjoint estimands. De la Cuesta, Egami, and Imai, 2022; Ganter, 2023; Leeper, Hobolt, and Tilley, 2020; Zhirkov, 2022. how to interpret estimates. Abramson, Koçak, and Magazinnik, 2022. and multiple testing issues. Goplerud, Imai, and Pashley, 2022; Liu and Shiraito, 2022.

decision to make. In contrast, traditional survey research best practices, which include asking simple, concrete questions about specific aspects of choices or attitudes, try to sidestep these trade offs. Payne, 2014. Indeed, "[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. Avoiding trade offs, however, is not possible with real-world choices, which usually makes conjoint designs more realistic and gives them all the advantages and disadvantages inherent in these real-world choices.

A statistical consequence of the inherent complexities in asking questions that reflect real-world choices is *measurement error*. McCullough and Best, 1979. Although this is a well-known methodological problem that can potentially bias causal inferences in any direction by any amount, measurement error and its consequences have been ignored in nearly all conjoint applications, potentially leading to rampant biases throughout the literature. As we demonstrate in this paper, even highly attentive survey respondents produce data with substantial measurement error, which we can see via estimates of intra-respondent reliability. When faced with two identical conjoint tasks just moments apart, respondents select the same profile only about 75% of the time, which is about halfway between flipping coins (50% agreement) and perfect reliability (100%) — results which are consistent with those in other fields. Bryan et al., 2000; Mørkbak and Olsen, 2015; Skjoldborg, Lauridsen, and Junker, 2009. As we show below, measurement error is notably worse in conjoint than traditional survey questions.

Based on thirteen surveys we fielded on five survey platforms (with a total of 9,472 respondents and 137,786 respondent-tasks), we 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. We then use this pattern to develop a simple statistical correction for the resulting biases. This method can be used not only by researchers at the design

stage but also by those analyzing preexisting data. As we explain, everything necessary to correct the bias in an application can be estimated via a slight modification of the standard conjoint design, a separate survey run afterward, or sometimes without new data collection at all. In many situations, correcting the bias will make results stronger; but sometimes, results will be weaker, or signs will flip. Either way, the correction is easy to apply.

We begin, in the next section, with a self-contained introduction to conjoint analysis that is intended to be far simpler, easier to understand, and simultaneously more general than common approaches in the literature. After we describe our modifications to correct for measurement error, we conclude with recommendations for conducting conjoint studies and offer easy-to-use open-source software.

The Conjoint Survey

We describe the data in the first subsection, quantities of interest in the next, and how to avoid historically relevant confusions and limitations in the third. Throughout we simplify the notation and concepts used in the literature, while still allowing maximum flexibility in the substantive questions to which these methods can be applied.

Data

For expository purposes, we begin with the special case of one task per respondent. We then extend our approach to any number of tasks. (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 quantities and Roman letters for observed quantities.)

Consider a simple conjoint experiment where we present individual i ($i = 1, \ldots, N$) the task of making a choice between two options (which we refer to as "candidates" to fix ideas) so that $C_i \in \{0, 1\}$ is the outcome variable. The explanatory variable values are randomly assigned from an investigator-chosen vector of attributes A_i, each element of which is a categorical variable describing the pair of profiles (candidates) together. (The

attributes and their possible values are chosen by the researcher and fixed; their values are randomized to respondents like in any experiment, but leading in this case to a choice for each respondent between two candidates with randomly selected attribute values.) Three types of profile-pair choice attributes organize the many types of substantive questions that can be addressed with this survey question:

1. *Independent Attributes* characterizing the two candidates separately and unconstrained within the pair, such as race (with categories, e.g., "Black, White," "Asian, Black," "White, White," etc.) or prior elective experience (e.g., "state legislator, not previously elected," "mayor, state legislator," etc.);

2. *Dependent Attributes* constrained across the two candidates in the pair, such as the probability of winning (with values that sum to 1, such as "0.4, 0.6", "0.75, 0.25", or "0.5, 0.5") or party membership in partisan elections (e.g., "Democrat, Republican" or "Republican, Democrat") and

3. *Pair-level Attributes* summarizing the pair of candidates together, such as the contest (e.g., "lower house" or "upper house" elections) or region in which the hypothetical election is taking place (e.g., "south," or "west").[2]

Then the unit of analysis (and the randomization of attribute values) is at the level of each respondent's choice between two candidate profiles. We thus structure the dataset in a familiar way with N rows, with columns coding the outcome variable as the respondent's choice and the explanatory variable as the investigator-chosen profile-pair attribute values. Given random selection of the respondents from a population, the N rows are independent and data analysis can be conducted without any specialized procedures for point or uncertainty estimation. For example, we can simply compute the average proportion of respondents choosing the Democrat in partisan races or the difference in this

[2]Most researchers fix and ignore the effect of which candidate appears on the left vs. right when presenting attributes to respondents, but this can be coded as an additional pair-level attribute. We could also include an attribute for the names of the choices (e.g., Candidate 1, 2; Democrat, Republican; Alice, Bob; or Tropicana, Minute Maid). Furthermore, although the order of attributes itself is not an "attribute" for each profile pair, we can explicitly code this order presented to each respondent, as may be useful for studies on party-label or ballot-order effects. e.g., Horiuchi, Kuriwaki, and Smith, Forthcoming. In short, many different types of substantive questions can be encoded in a conjoint design.

average for contests with an open seat and a contest with an incumbent running against a non-incumbent. Because profile-pair attributes are randomly assigned, modeling assumptions are rarely necessary. Uncertainty estimates such as standard errors can be computed with classical approaches without any specialized procedures.

Researchers usually increase efficiency by giving each respondent $T > 1$ tasks (with $T \approx 5$). Thus, for individual i ($i = 1, \ldots, N$) and task t ($t = 1, \ldots, T$), we denote choices as $C_{it} \in \{0, 1\}$ and the vector of candidate attribute pair values as A_{it}. For expository simplicity, we follow common practice by assuming independence across tasks and respondents after conditioning on attributes and personal characteristics (although allowing correlation across tasks within individuals may sometimes be useful). In this case, the unit of analysis of our simplified approach is the respondent-task, with the data structured as $N \times T$ rows, and each row still represents one choice. This data structure also requires no specialized uncertainty estimation and can be used to study all three types of attributes.

Finally, we also measure a vector of exogenous personal characteristics P_i for each respondent, such as demographics, socioeconomic status, or political or other views. Although the content of A_{it} is controlled by the investigator, P_i is observed and cannot be randomized but is useful for defining subgroups, within which all our methods can be easily applied without modification.

Quantities of Interest

We assume each individual i has a *preference* for one of the two candidates $\rho_i(a) \in \{0, 1\}$ for each possible vector of attribute values a. Even without measurement error, preferences are "potential outcomes" and thus only observable for attribute values actually asked of respondents; that is, $\rho_i(A_{it})$ is observed, but preferences $\rho_i(a)$ for all $a \neq A_{it}$ are unobservable. We also partition the attributes as $a = \{a_\ell, a_{-\ell}\}$, where a_ℓ is the scalar value of the "attribute of interest" and $a_{-\ell}$ is a vector of the remaining values. (Because random assignment makes post-treatment bias irrelevant, we can compute different quantities of interest from the same survey by merely redefining which one is the attribute of interest, ℓ, and applying the same methods of calculation repeatedly for different attributes.)

Most commonly used quantities of interest are linear combinations of what we call the choice-level *marginal mean* (MM), given a researcher-chosen set of attribute values. MM is a simple average of preferences over individuals and all possible values of $a_{-\ell}$ for each individual, with the selected value of the attribute of interest a_ℓ held constant:[3]

$$\rho(a_\ell) = \underset{i,a_{-\ell}}{\text{mean}} \Big[\rho_i(a_\ell, a_{-\ell})\Big]. \tag{1}$$

Averaging over the possible values of the remaining attributes (instead of holding them constant as in traditional survey analysis; King, Tomz, and Wittenberg 2000) is unusual but is a common practice in conjoint designs because it makes estimation easier.[4]

Although also used as a building block for all other quantities, the choice-level MM is of substantive interest in and of itself. For example, one MM of interest is the proportion of respondents who prefer an incumbent candidate when running against a nonincumbent (on average over different situations), which we can create by defining a_ℓ as the pair "incumbent, nonincumbent."

The key quantity computed from a combination of MMs is the choice-level *average marginal component effect* (AMCE), which is the change in the average preference when altering the value of the attribute of interest from a_ℓ to an alternative value a'_ℓ, averaged over all possible values of all other attributes, $a_{-\ell}$. This is the simple difference between two choice-level MMs:

$$\theta(a_\ell, a'_\ell) = \rho(a_\ell) - \rho(a'_\ell). \tag{2}$$

In the incumbency example we used above, the AMCE is the causal effect of having an incumbent candidate (running against a non-incumbent) in an electoral contest on the respondents' probability of choosing the incumbent minus the probability of the same party's candidate in an open-seat electoral contest. The choices are averaged over respondents and tasks.

[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 $\text{mean}_i[g(i)]$.

[4] We define all quantities of interest as analogies to the commonly used sample (as opposed to the population) average or average treatment effect. Imbens, 2004; Imai, King, and Stuart, 2008. See Abramson, Kocak, et al. 2023. Equation 1 can also be changed to a weighted mean to reflect different reference populations. De la Cuesta, Egami, and Imai, 2022; Ganter, 2023.

Avoiding (Historically Relevant) Data Confusions and Limitations

In marketing, where conjoint was first widely used, scholars often ask respondents to rate each of the two candidate profiles of a conjoint separately, rather than to choose between the two. The two separate ratings resulted in the sensible use of the profile rather than the choice as the unit of analysis. This meant, for example, that a survey with n respondents, each answering only one task, would yield a dataset with $2n$ observations.

Unfortunately, this profile-level perspective was then adopted by some social scientists using binary choice conjoints. This led to suboptimal "best practice" recommendations where researchers first duplicate and stack up their data in this way and induce unnecessary dependence (for example, in a partisan electoral contest the choice of the Democrat ($C = 1$) is one unit and the repetitive "not the Republican" choice ($1 - C$) is another). Whereas the two ratings are separate pieces of information, a choice is only one. In other words, researchers first created a statistical problem for no reason and then had to turn around and correct the problem they originally created. Correcting it (usually with clustered standard errors or extra modeling) addresses the problem, and so no harm done. However, the busy work, the strange data structure, the limitations on substantive questions that can be asked, the complicated corrections, and the unnecessary difficulty in learning can all be avoided altogether with the simpler choice-level strategy described in previous two subsections above.

Even more consequential is that profile-level choices narrow the substantive questions that can be asked. Choice-level analysis enables researchers to choose from the three types of profile-pair choice attributes described in the Data section above. In contrast, profile-level analysis restricts substantive questions to only one of the three types, "independent attributes" (at least without even more complicated and unnecessary induced problems and corrections).

Despite the weaknesses of profile-level data structures, and the lack of good reasons to continue to use them, many prior political science articles have used this strategy and so we pause to define two relevant quantities that can be coded in this way. First is the *profile-level MM*. To define this, we first add separate notation to distinguish the attribute

of interest levels chosen for the two candidate profiles by letting $a_\ell = (a_0, a_1)$. Then, the profile-level MM is the proportion choosing Candidate 0 among those with level a_0 of the attribute of interest for Candidate 0 — averaged over all individuals i, all possible combinations of levels a_1 for Candidate 1 (including even values such that $a_1 = a_0$), and all other attributes $a_{-\ell}$ for both candidates:

$$\bar{\rho}(a_0) = \underset{a_1}{\text{mean}} \big\{ \rho(a_0, a_1) \big\}. \tag{3}$$

And second, the *profile-level AMCE* is the difference between two profile-level MMs:

$$\begin{aligned} \bar{\theta}(a_0, a_0') &= \bar{\rho}(a_0) - \bar{\rho}(a_0') \\ &= \underset{a_1}{\text{mean}} \big\{ \rho(a_0, a_1) \big\} - \underset{a_1}{\text{mean}} \big\{ \rho(a_0', a_1) \big\}. \end{aligned} \tag{4}$$

As the choice-level MM and AMCE allow for simpler statistical analyses, are easier to understand and applicable to a wider array of substantive questions excluded by profile-level quantities, and can be used as building blocks to create other quantities, we use these quantities for our illustrations below. However, our measurement error corrections can be used in the same way with choice- or profile-level analysis.

Correcting Measurement Error Induced Bias

We now introduce notation for the possibility that respondents' observed choices are not always equal to their (unobservable) preferences, detail the biases that result from this measurement error, and show how to correct these biases.

Observation Mechanism

Most applications of conjoint and other types of survey research explicitly acknowledge the presence of measurement error and the resulting potential biases. Kane and Barabas, 2019; Berinsky, Margolis, and Sances, 2014; Curran, 2016; Ward and Meade, 2023. Researchers usually try to mitigate these biases via different types of attention checks or other filters. We follow the literature and use these checks but relax the assumption that those who pass a check are free of measurement error. To do this, we recognize that

conjoint data may have what we call *swapping error*, where some of the respondents' reported answers to a binary question with attributes a reflect their (true) preferences $\rho_i(a)$, but other answers are swapped with the wrong ones $1 - \rho_i(a)$. (This can occur with any binary outcome variable, even if not from a conjoint.) To formalize this idea, we define each respondent's reported *choice* between the two candidates, with attribute vector a, as

$$C_i(a) = \begin{cases} \rho_i(a) & \text{w.p. } 1 - \tau_i(a) \\ 1 - \rho_i(a) & \text{w.p. } \tau_i(a), \end{cases} \tag{5}$$

where $\tau_i(a)$ is the probability of swapping error (i.e., when the respondent's choice does not reflect their true preference $\rho_i(a)$) and "w.p." is standard mathematical notation for "with probability". We also make the reasonable assumption that some information exists in the data, and so choices are not made by flipping coins, $\tau \in [0, 0.5)$. Most prior conjoint research implicitly assumes the absence of measurement error, $\tau_i(a) = 0$ for all a and i, which we show below is not justified.[5]

Consequences of Ignoring Measurement Error

As introductory linear regression textbooks commonly explain, measurement error in a continuous outcome variable causes no coefficient bias as long as the error is random with zero mean. This consequence is easy to see: it is equivalent to a regression with no measurement error in the outcome variable but higher residual error. However, swapping error in a binary outcome variable cannot be mean zero as it swaps some zeros with ones and ones with zeros. Thus, ignoring swapping error biases inferences. The bias induced by swapping error cannot be ignored and also cannot be corrected by general purpose methods for correcting measurement error bias that assume the error is mean zero. e.g., Blackwell, Honaker, and King, 2017.

[5] Although we have not found evidence for other types of measurement error and thus use only Equation 5 in this paper, many other observation mechanisms are possible and so are worthy subjects for future research. For example, consider a survey where measurement error affects the choice of Candidate B, perhaps due to social desirability bias, but not A. For another, we may conceptualize continuous unobserved propensities of the candidate choice (as in biostatistics) or continuous unobserved utilities for each choice (as in econometrics, with differences in utilities equaling candidate choice propensities) — in either case with error. As we show in Supplementary Appendix B (p. 2-4), some of these alternative observation mechanisms are equivalent to Equation 5 whereas others may generate different types of biases and require different correction methods.

Formally, the standard estimators of the choice-level MM, $\rho(a)$, and AMCE, $\theta(a, a')$,

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

are unbiased if $\tau_{it}(a) = 0$ for all i, t, and a. However, they are biased in the presence of non-zero swapping error, which we show by taking the expectation over the random assignment of profiles and swapping error, with true potential preferences fixed:

$$\begin{aligned} E[\hat{\rho}(a)] &= E[\underset{it}{\text{mean}}(C_{it})] \\ &= \underset{it}{\text{mean}} \left[\rho_{it}(a)\big(1 - \tau_{it}(a)\big) + (1 - \rho_{it}(a))\tau_{it}(a)\right] \\ &= \rho(a) + \underset{it}{\text{mean}} \left(\tau_{it}(a) - \rho_{it}(a)\tau_{it}(a)\right) \\ &\neq \rho(a) \end{aligned} \tag{6}$$

and thus

$$\begin{aligned} E[\hat{\theta}(a, a')] &= E[\hat{\rho}(a)] - E[\hat{\rho}(a')] \\ &= \theta(a, a') + 2\left[\underset{it}{\text{mean}} \left(\rho_{it}(a)\tau_{it}(a)\right) - \underset{it}{\text{mean}} \left(\rho_{it}(a')\tau_{it}(a)\right)\right] \\ &\neq \theta(a, a'). \end{aligned} \tag{7}$$

Note that these expressions do not assume that swapping error is fixed over i, t, or a, subjects we address below. When estimating the marginal mean (or average preference) or AMCE by subgroups (defined by P_i), all of our results hold within each subset.

Estimating Swapping Error

As we demonstrate in the Correcting Measurement Error Bias section, correcting measurement error bias is straightforward if we have estimates of swapping error, $\tau_{it}(a)$. However, in principle, $\tau_{it}(a)$ may vary over individuals i, tasks t, and attributes a, which would seem to require that researchers obtain at least $N \times T$ estimates of the probability of swapping error. By showing below that swapping error is mathematically related to Intra-Respondent Reliability (IRR), we have a way to estimate one of these quantities. But 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 these problems in three steps.

First, we provide extensive empirical evidence in our section on Conjoint Induced Patterns in Measurement Error that, within a conjoint survey, IRR(a) is not a function of the attributes a (including potential carryover effects or profile orders). In other words, $\tau_{it}(a) \approx \tau$ and so is unlikely to vary systematically with different attribute levels. This empirical finding simplifies the estimation of swapping error probability down to a single parameter. We find that IRR can vary somewhat across applications (and respondent characteristics), and so we need to estimate it for every conjoint survey. If researchers are interested in subgroup comparisons, they also need to estimate IRR for each subgroup.

Second, the Estimating the Intra-Respondent Reliability section offers several easy ways of estimating IRR at the design stage or after data is collected. Our preferred way involves repeating the first conjoint task at the end of the task list for each individual (usually including about 5 tasks) and then calculating the proportion agreement between these two identical conjoint questions. This approach takes advantage of two unusual benefits of conjoint survey designs: (a) Asking the two questions so close together makes it unlikely that different "considerations" (i.e., unmeasured confounders; see Zaller 1992) account for differences in respondent choices.[6] (b) Even though the repeated questions are asked only moments apart, we find that respondents virtually never remember having been asked an identical question previously. Across all of our studies, zero out of 9,472 survey respondents reported noticing being asked identical questions — even when prompted in open-ended questions to carefully explain how and why they chose the profile they did (see our Supplementary Appendix for details, p. 24).[7]

Finally, these advantages of conjoint questions enable us to obtain a clearer indicator of measurement error than is possible with traditional survey questions. The key assump-

[6] In contrast, the venerable literature on issue knowledge and survey response instability in the electorate almost exclusively relies on repeated questions asked from weeks to years apart; see Achen 1975; Lazarsfeld 1948; Converse 2000; Zaller and Feldman 1992.

[7] If profile order or carryover effects exist, they will be subsumed under what we are calling swapping error. However, perhaps because conjoint questions are sufficiently complicated, and because our two questions are separated by several distractors, we find no evidence of learning or memory effects from one question to its repeat through "carryover effects". Ham, Imai, and Janson, 2022. the largest p-value from formal tests across the eight studies we analyzed is 0.18.

tion required is that, conditional on preferences, the choices made in the pair of identical questions are independent: $C_{i1}(a) \perp\!\!\!\perp C_{iT}(a) \mid \rho_i(a)$. Then, from Equation 5, we set IRR equal to the probability that a respondent gives the same answer to the two identical questions:

$$\text{IRR} = 1 - 2\tau(1-\tau).$$

We then solve this expression for τ,

$$\hat{\tau} = \frac{1 - \sqrt{1 - 2(1 - \text{IRR})}}{2}, \tag{8}$$

which we use as an estimator for the probability of swapping error.

Correcting Measurement Error Bias

Although only a single parameter is required for each bias correction, we do not need to assume $\tau_{it}(a) = \tau$. Instead, as we now demonstrate, we only require the less restrictive assumption that swapping error probabilities are linearly unrelated to respondent preferences in sample: If $\text{Cov}\big(\rho_{it}(a), \tau_{it}(a)\big) = 0$ (where $\text{Cov}(\cdot)$ is the sample covariance operator), then $\text{mean}_{A_{it,\ell}=a}\left(\rho_{it}(a)\tau_{it}(a)\right) = \rho(a)\tau(a)$. Even when researchers cannot distinguish between these assumptions in a particular application, this less restrictive assumption is useful in demonstrating that estimated results will be robust to the independence assumption.

We begin by simplifying the bias expressions by assuming the in-sample zero covariance in Equations 6 and 7, respectively, as

$$E\left[\hat{\rho}(a) \;\middle|\; \text{Cov}(\rho_{it}(a), \tau_{it}(a)) = 0\right] = \rho(a) \cdot (1 - 2\tau) + \tau \tag{9}$$

$$E\left[\hat{\theta}(a, a') \;\middle|\; \text{Cov}(\rho_{it}(a), \tau_{it}(a)) = 0\right] = \theta(a, a') \cdot (1 - 2\tau) \tag{10}$$

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

$$\tilde{\rho}(a) = \frac{\hat{\rho}(a) - \tau}{1 - 2\tau}, \qquad \tilde{\theta}(a, a') = \frac{\hat{\theta}(a, a')}{1 - 2\tau}, \tag{11}$$

which are unbiased if τ is known, $E[\tilde{\rho}(a)] = \rho(a)$ and $E[\tilde{\theta}(a, a')] = \theta(a, a')$. They are consistent as long as an estimate of τ is consistent (the Finite Sample Properties and Empirical Examples section 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 attribute of interest to refer to more than one element of the attribute vector.

Equations 11 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 $\hat{\rho} < 0.5$, the corrected estimate will be smaller than the biased estimate, but if $\hat{\rho} > 0.5$, the corrected estimate will be larger. This can be seen by solving for the difference between the corrected and uncorrected estimates as

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

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 $\tilde{\rho}(a)$ and $\tilde{\theta}(a, a')$ requires an extra step because of the uncertainty in $\hat{\tau}$. We show how to do this in Supplementary Appendix C (p. 5-8) in three different ways that optimize for speed, convenience, or familiarity, all of which are implemented in our open-source software. Finally, although we recommend using the more general choice-level definition of MMs and AMCEs, our methods of correcting measurement error bias described here are also applicable to special cases such as the profile MMs and AMCEs.

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 published conjoint studies; (2) estimating the IRR 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 IRR across attribute combinations within studies; and (5) showing how IRR varies over the personal characteristics used for subgroup estima-

tion.[8]

Eight Replications

We choose eight published political science conjoint studies to replicate, with a preference for those in major journals and substantively diverse topics. The studies 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. These articles included conjoint about housing developments, climate agreements, political candidates, immigrants, and others (see our Supplementary Appendix and replication file for details).

We then fielded a series of survey experiments using U.S. samples with nationally representative quotas based on age, gender, race, ethnicity, and region (from Lucid Marketplace; see Coppock and McClellan 2019). Although only Bechtel and Scheve. 2013. report using attention checks among the eight studies we replicate, we give conservative results on IRR by dropping respondents who failed an attention check administered prior to our conjoint task (see our Supplementary Appendix p. 26-27 for details; we find little evidence that respondent inattentiveness explains low IRR).

Most replication studies in the literature begin with the data and methods from a published article and try to 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 170 AMCEs computed in the eight studies (all of which are profile AMCEs). Figure 1 presents a scatterplot of estimates of these quantities from the original studies plotted horizontally and our replication of each AMCE from 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).

[8]The studies we describe in this section and in the Supplementary Appendix received IRB approval prior to data collection. Respondents recruited from Lucid, Prolific, and Mechanical Turk were paid minimum wage or higher for participating in the survey. Respondents recruited from a university-based source were online volunteers who chose to participate for fun and/or to contribute to social science research. Respondents provided informed consent prior to participating, their confidentiality was maintained at all times, and the survey involved no harm or deception.

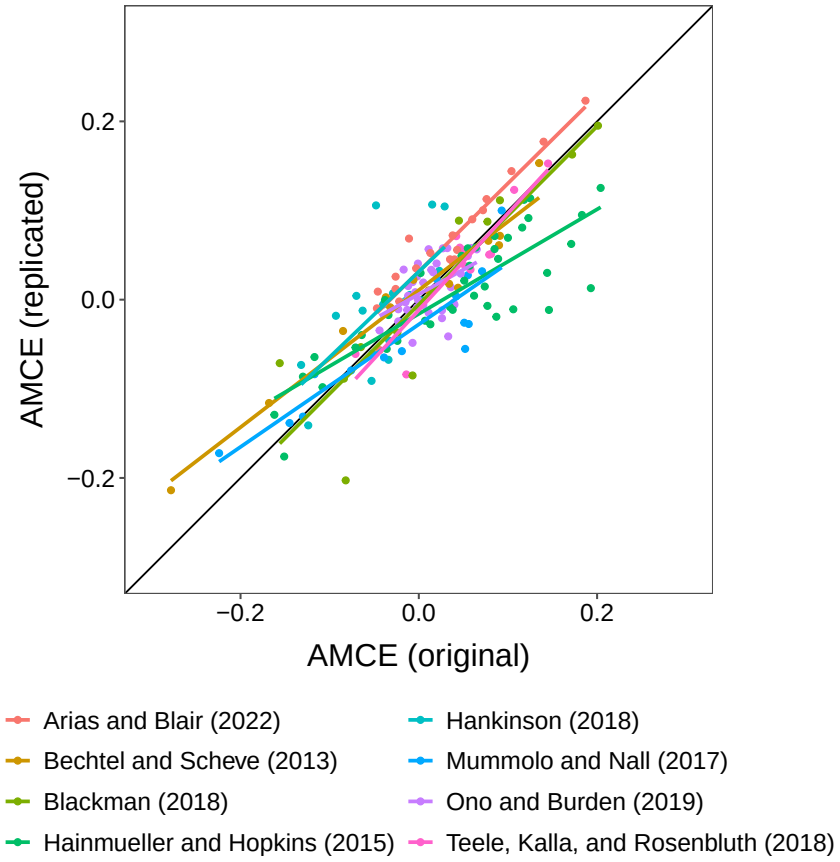

Figure 1: Original vs. Replicated AMCEs in Eight Prior Studies. Notes: Points are represented as a scatterplot of AMCEs from the original studies (plotted horizontally) by estimates from our replications in new data (plotted vertically), colorcoded by article, along with a regression line fit to all estimates from each study.

Despite the differences across the survey platforms, the sampling periods, details of survey implementation, and sample characteristics, the results in Figure 1 reveal a close correspondence between the originally published estimates and the estimates based on our replications of these studies. This can be seen in the distance between each of the points and the 45-degree line, or the eight (different colored) regression lines, all also fairly close to the (black) 45-degree 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 prevalence of replication failures across scientific fields in recent years. Gilbert et al., 2016; Open Science Collaboration, 2015. researchers should be be proud to see the uniformly high levels of transparency and scientific rigor achieved in the literature on conjoint-based political science experiments displayed in Figure 1.

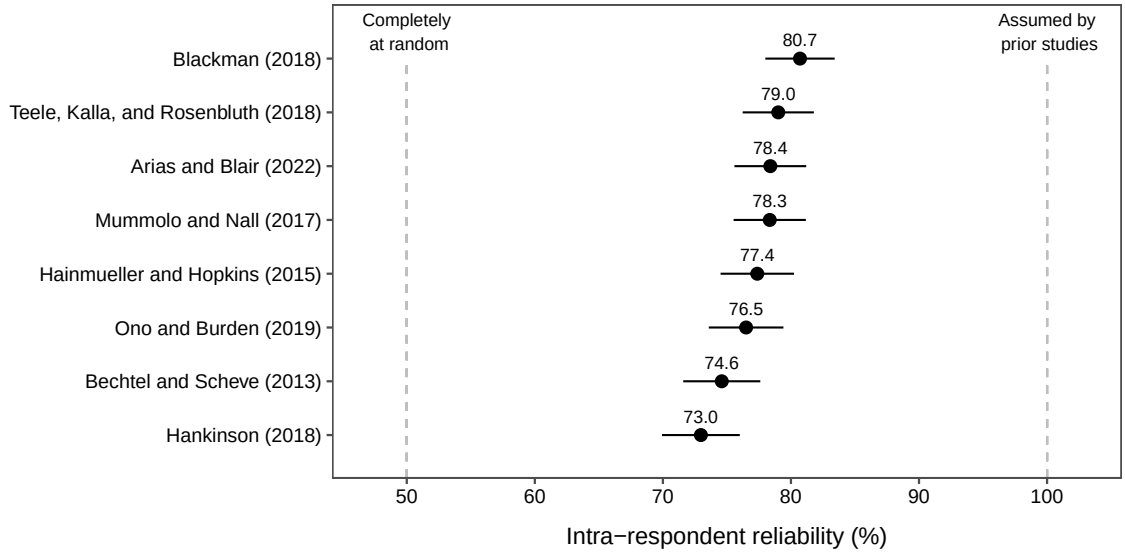

Figure 2: Inter-Respondent Reliability of Eight Prior Studies. Notes: Point estimates appear as dots with 95% confidence intervals as a horizontal line.

Estimates of Intra-Respondent Reliability

Second, we estimate the IRR for each of the eight studies. We do this by randomly assigning two of the eight original studies to each of the 3,289 respondents. We standardize the number of tasks (conjoint survey questions) per respondent across our eight replications to five (the mean, median, and mode of the studies we replicate) and then add a sixth conjoint task that repeats the first (randomly selected) task at the end. That is, just a few moments after a respondent chooses between two profiles, we ask this same person the same question a second time and see whether their answer is the same. Then, our estimate of IRR is the proportion agreement between these first and last (repeated) responses.[9]

Results appear in Figure 2. The horizontal axis in Figure 2 indicates IRR, ranging between respondents flipping coins (50%, at the left) and perfect agreement (namely,

[9]To minimize the possibility that respondents notice that the same information is included in the first and last tasks, we switch the profile order (i.e., between the left and right columns). (Our Supplementary Appendix includes information on three additional surveys we conducted to study the effect of this procedure. We found, in two of the three surveys, that the not switching estimate was slightly smaller, small enough that differences in our bias corrections would not be substantively meaningful. We also conducted a more formal randomization test with similar results; see Ham, Imai, and Janson 2022.) We also gave each respondent two different randomly selected sets of tasks, following the experimental protocol in the text, to increase efficiency. This means that each respondent received $12 = (5 + 1) \times 2$ conjoint tasks. The results from only the first of these two sets of tasks, which was as close as possible to what the original articles used, are not materially different from the second set.

perfect reliability) as is assumed by most conjoint applications (100%, at the right). Our point estimates appear as dots, with 95% confidence intervals as horizontal lines. IRR 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%). For IRR $=$ 0.75, we know from Equation 8 that $\tau = 0.15$, meaning that about 15% of reported choices do not reflect respondent preferences (swapping 0 for 1 or 1 for 0). If these 15% could be chosen by an adversary, almost any type of bias can be induced; nature is not always this unkind, but substantive conclusions should obviously be based on evidence where possible.

Reliability Comparisons with Traditional Survey Questions

Third, we provide evidence that, as might be expected, IRR is higher for conjoint questions designed to analyze inherently complicated real world decision making than for traditional multiple-choice survey questions with similar content abstracted from these real world choices. Of course, we are only measuring reliability; we would expect the usual abstract survey questions to have lower validity than conjoint designs (especially when corrected for measurement error biases).

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), with the order of the two types of questions randomized across respondents. 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 IRR for the conjoint tasks vs. the traditional survey questions. (See our Supplementary Appendix for details.)

Results appear in Figure 3: While IRR (on the horizontal axis) is 79.5% in the conjoint experiment (at the bottom), all three survey questions have, as would be expected,

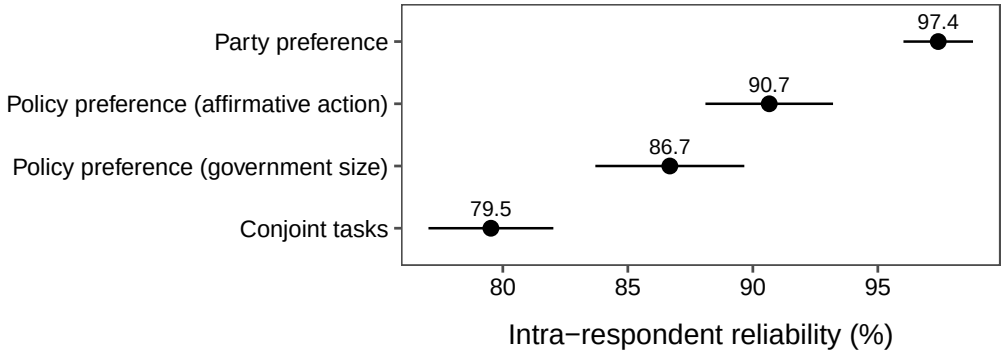

Figure 3: Intra-Respondent Reliability of Traditional Surveys vs. Conjoint. Notes: Intra-respondent reliability across each question type appears as dots with 95% confidence intervals as a horizontal line.

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 may suggest that the lower IRR in conjoint experiments is inherent in the more complex real world nature of the conjoint analysis.

No Reliability Variation by Attributes

We now present evidence that the reliability of conjoint survey questions does not vary systematically with the pairs of attribute combinations (i.e., the information contained in conjoint tables presented to the respondent). We do this from both a top-down theoretical approach, which we describe here (with empirical evidence in Supplementary Appendix D, p. 8-16), and a bottom-up empirical approach that we present next (with additional information in Supplementary Appendix E, p. 16-20).

Top-down approach

We apply the literature on survey best practices to conjoint studies to study how reliability may be reduced as a function of the content of the profile-pairs presented to respondents. We develop the following three hypotheses that explain IRR and test them empirically.

First, *inconsistency* is the level of disagreement across attributes within a profile when interpreted from the perspective of its most prominent dimension. see Bansak and Jenke, 2023. 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 decreasing IRR.

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 decrease IRR.

Finally, *divergence* refers to the degree of dissimilarity between the profiles. Attribute levels with small differences between profiles within a pair may encourage respondents to assess options essentially at random, increasing IRR. For example, in a candidate conjoint experiment, "moderate Democrat" versus "moderate Republican" is less divergent than "extreme Democrat" versus "extreme Republican"; in a conjoint experiment on housing developments, "3 units versus 5 units" is less divergent than "3 units vs 50 units." And because attributes are randomly assigned, some attributes will have identical levels and zero divergence. We hypothesize that respondents will have an easier time choosing between candidates with larger differences.

As Supplementary Appendix D (p. 8-16) shows, through numerous survey experiments, we find no systematic evidence that inconsistency, complexity, or divergence accounts for the variation in IRR. On the theory that the exception proves the rule, we were able to construct highly artificial and unrealistic conjoint studies that affected IRR, but not by enough to make a difference in real cases. We also went further and studied the consequence of attribute sets with a single (and again artificially extreme) dominant attribute and failed to find a systematic substantive explanation for IRR there either, except in the most extreme and unrealistic cases.

Thus, we find that in realistic conjoint experiments, with the types of attributes and levels used in social science applications and with variation one would see in reality, IRR rarely varies in substantial ways as a function of the attribute levels. An advantage of conjoint analysis is the ability of researchers to present respondents with difficult options, which people often face in real-world shopping, dating, and voting decisions. In part

because of this difficulty, the IRR is not 100%, although it tends to be unrelated to the attributes used in defining the profile-pairs.

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 IRR, we also present respondents with another six profile-pairs, identical to the first six (with the profile appearing on the left and right flipped).[10] We then collect about 50 respondents for each of our 48 profile-pair combinations (Sample 1). To reduce uncertainty, we repeat the entire experiment with 100 responses for each combination in a separate survey (Sample 2) and present the results for the samples separately to allow for sampling and population changes. These two studies yield two IRR estimates for each combination. (See Supplementary Appendix E, p. 16-18 for additional design details.)

Our estimates of IRR from these experiments appear in Figure 4, Panel (a). Each point represents one profile-pair combination for a choice, with IRR estimated from Sample 1 plotted horizontally and Sample 2 plotted vertically. The mean in each sample is about the same as the means 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 E, p. 16-20).

If the IRR 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 in Panel (a) of Figure 4: the empirical correlation of the points in the graph is 0.23 with an

[10] In addition to excluding ties, we avoid showing the same conjoint table consecutively. Specifically, we fix the first task in each set (while randomizing the order of the other five tasks) so that the last task in the first set and the first task in the second set are always different.

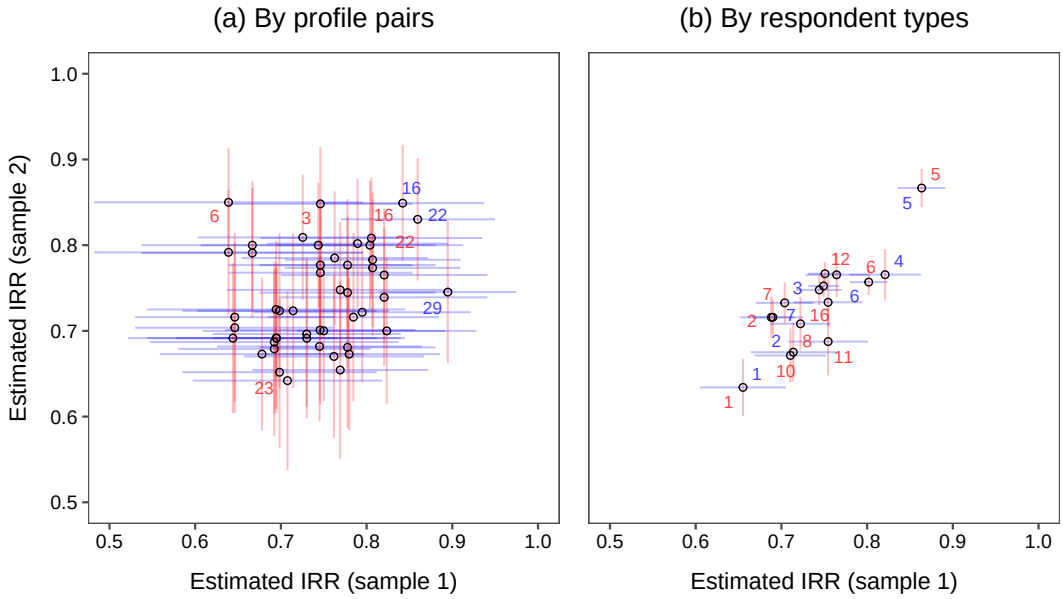

Figure 4: Variation in Intra-Respondent Reliability over (a) Attributes and (b) Personal Characteristics. Notes: A key to the numbers shown in each plot appears in our Supplementary Appendix (p. 21). Points that differ from the mean (for each sample) at the 95% level are given a numeric code (with blue for Sample 1 and red for Sample 2).

(insignificant) p-value for a difference from zero of 0.112. Moreover, if IRR estimates differed from their mean only due to sampling error, we would expect to see, on average, 2.4 of these 48 points "significant" at the 95% level. In fact, we see three in Sample 1 and five in Sample 2. 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), we see no evidence for systematic patterns, at least not large enough to make a difference in our bias corrections. Even via post hoc interpretations of the data, 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 E, p. 16-20).

Thus, all the evidence on this question seems to point in the same direction: If predictable differences in IRR exist as a function of the profile-pairs with randomly assigned attributes, they are unlikely to be large enough to matter substantively.

IRR Variation by Personal Characteristics

Finally, we use the same methodology from Panel (a) of Figure 4 to demonstrate that IRR 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 our Supplementary Appendix, p. 21), 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 IRR is indeed systematic. Although these effects vary over studies, we often find that younger, non-white, and male respondents tend to have lower levels of reliability.

These results indicate that assuming constant IRR over attributes is usually justified. However, researchers should use separate IRR estimates for descriptive and causal analyses that are analyzed within subgroups defined by personal characteristics.

Estimating the Intra-Respondent Reliability

We propose here four methods of estimating the IRR, which is required to estimate the swapping error parameter, τ (see Equation 8). Two are for new conjoint projects that work via simple adjustments to the survey design (the following section), while the other two are for analyses of existing conjoint datasets where new data collection is infeasible (the Estimation without Additional Data Section).

Estimation via New Survey Data

Conjoint studies still in the design stage can be easily modified to estimate IRR using one of the following two procedures. The first, which we recommend for most researchers and use in the Estimates of Intra-Respondent Reliability section, is to estimate only the average IRR by adding one extra task at the end of a conjoint survey that repeats the first task but with the order of profiles flipped between left and right. We find no evidence that respondents notice the repetition, which makes this a simple, inexpensive, and widely applicable approach to estimate IRR, and to infer swapping error.[11]

[11] 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

Estimating the overall average IRR is useful for researchers willing to make the assumption we justify empirically in the Patterns in Conjoint-Induced Measurement Error section. Researchers who prefer not to make this assumption can instead choose a second, more extensive procedure, by estimating IRR for different values of the attributes. Researchers studying subgroup effects may also wish to estimate IRR within each subgroup.

Estimation without Additional Data

We now offer two methods of estimating IRR from a pre-existing conjoint survey without any new data collection or survey design changes. Avoiding new data collection obviously saves costs, but it has additional benefits. 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 reliability. Of course, collecting more data is always preferable to these approaches and should be pursued whenever feasible.

In most situations, we recommend using both of the following methods when new data collection is impossible. The first approach is to choose a value for IRR based on substantively similar studies for which it has already been estimated, such as some of the articles we replicated (see Figure 2). Uncertain estimates from less similar studies can be studied via sensitivity testing by repeating the bias correction for a range of IRR values.

The second approach involves estimating IRR 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 pair of tasks 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.[12]

For example, Hankinson. 2018. one of the studies we replicate, includes seven attributes. This means that a pair of tasks can differ in profile-pair attribute levels for a total of 0, 1, 2, 3, 4, 5, 6, or 7 attributes. The unobserved proportion agreement in a pair of

few other survey questions should be used between the pair of repeated questions to estimate IRR. These additional questions ensure respondents do not recall they are being asked the identical question twice.

[12]We develop this approach by adapting the methodology in Gakidou and King. 2006. for estimating mortality rates from surveys of people about their siblings.

tasks with 0 differences (which is unobserved in this dataset) 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 only one, 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 profile 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 three and four attribute-value differences). Next, in this same panel, we plot a weighted least squares regression line fit (dotted red) to these five data points (at 3, . . . , 7). Weights are calculated from the standard errors of each of the points, which differ because they are based on different numbers of observations. We then use the weighted least square estimates to extrapolate to 0 on the horizontal axis, the object of our inference. (We could extend the procedure with a more fine-grained task pair difference metric, such as by recognizing that some levels are ordered or interval scaled.)

Our estimate for IRR, the proportion agreement with no attribute differences, is the constant term in the regression. We plot this extrapolated estimate of the IRR 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 linear regression. King and Zeng, 2006.

Also in this top left panel is a triangle, which is our direct estimate of IRR 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 repeat this sample procedure for all eight of the studies we replicated, and each one is close to the direct estimate (each of which is in a separate panel in Figure 5). Obviously, we can offer

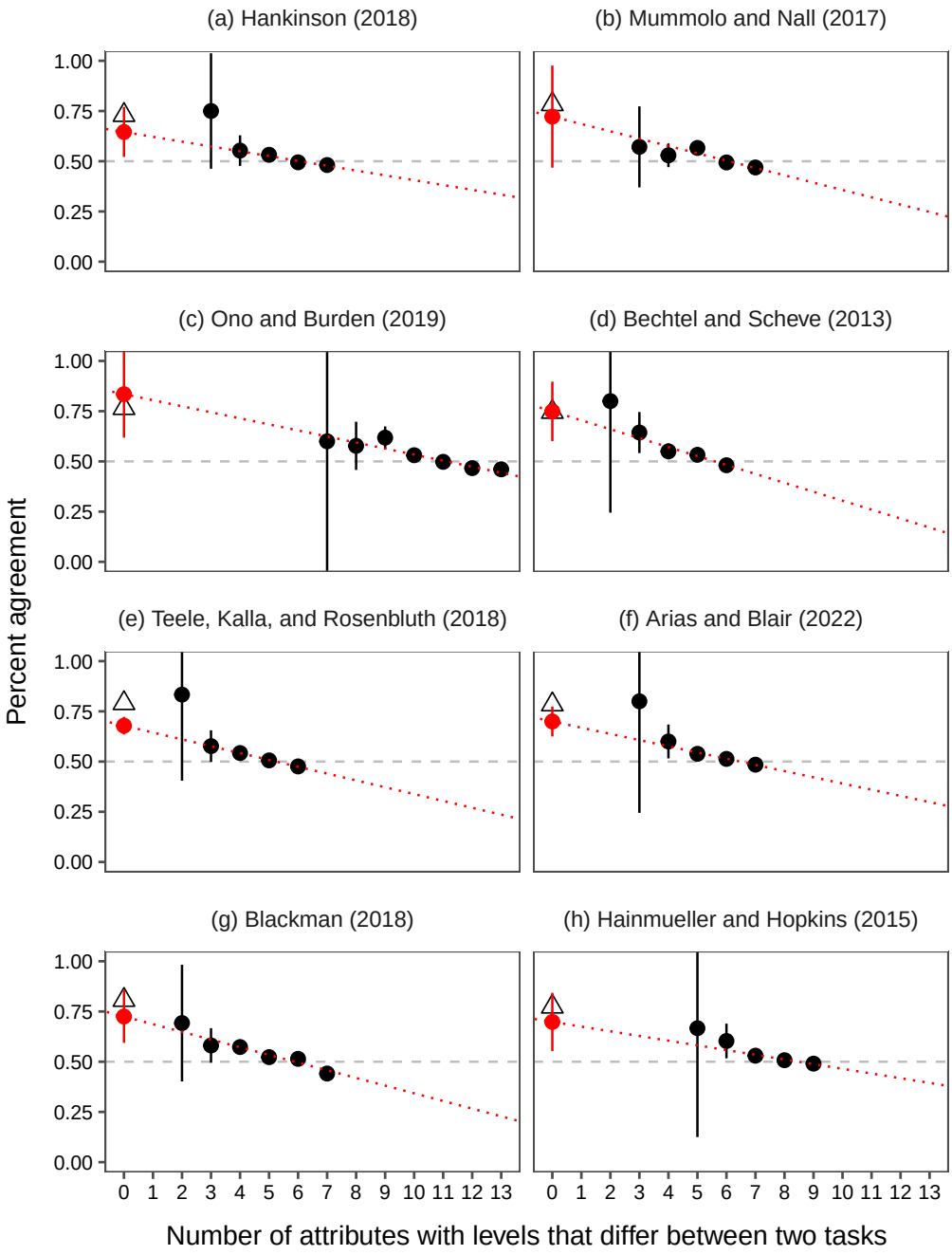

Figure 5: Estimating Intra-Respondent Reliability from Data without Repeated Tasks. Notes: The red dotted line extrapolates the black dots, representing percent agreement conditional on the number of attribute-value differences within task pairs, to the 0 difference point (see the red dot and 95% confidence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

no guarantee that this method will work as well in all future datasets, but it certainly is encouraging.

We could also extend this procedure by combining the extrapolation estimate with an estimate from one or more of our replication studies if they are substantively similar to the study being conducted. With any of these approaches, we would need to estimate the implied uncertainty of τ or, more simply, present uncertainty estimates of the MMs and AMCEs conditional on $\hat{\tau}$, supplemented with some sensitivity estimates.

Finite Sample Properties and Empirical Examples

The section entitled Correcting Measurement Error Bias offers estimators for the choice-level MM and AMCE corrected for measurement error (see Equation 11) and shows mathematically that the estimators are unbiased when τ is known and statistically consistent when τ is consistently estimated. The Patterns in Conjoint-Induced Measurement Error section shores up the key simplifying assumption in these estimators. To complement those analyses, we first undertake a Monte Carlo simulation and show 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 estimators. These results suggest that, in applications, researchers should use our bias correction because it corrects bias (see Equation 11). The mean square error result suggests that the slight increase in standard errors is less of a matter of concern compared to the point estimates and the substantive consequences of the corrections. We show below the pattern across estimates from our replications of our corrections decreasing, increasing, and flipping signs of subgroup differences.

Simulation

We begin with a population of 100,000 individuals with known true preferences, the true marginal mean $\rho(a)$, and AMCE $\theta(a, a')$. 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 $\tau = \{0.1, 0.15, 0.2\}$ by using the observation mechanism in Equation 5.

Finally, in each simulated dataset, we compute the uncorrected estimates (used throughout the literature) and our alternative corrected estimates 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 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 estimators (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 (using a color-blind-friendly palette, with values indicated in the figure legend). As anticipated by the mathematical results in the Correcting Measurement Error Induced Bias section, for both the marginal mean and the AMCE, bias increases as τ 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 MMs and AMCEs, our new estimators are approximately unbiased, which can be seen by all the lines appearing 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 estimators in the last row of panels. In almost all cases, the RMSE is lower for our corrected estimator than the uncorrected one. (When the quantities of interest are only slightly different from the null value (i.e., MM ≈ 0.5 and AMCE ≈ 0) but the measurement error is unusually large, the noise in the data is overwhelming the signal and bias corrections will be quite uncertain.) Every line for all simulations with different quantities of interest (indicated by different colors 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 the MMs and AMCEs.

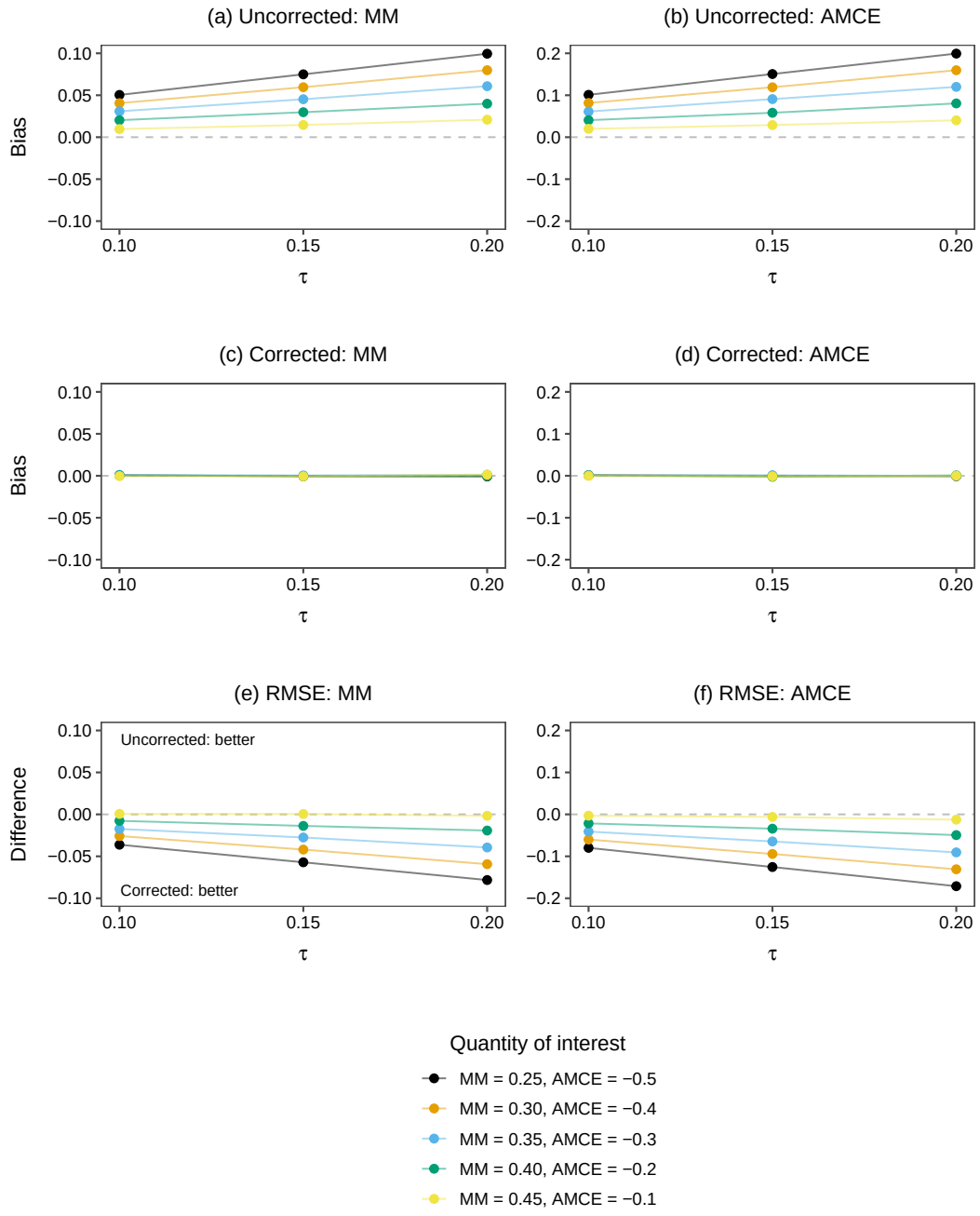

Figure 6: Reducing Bias and Mean Square Error in MMs and AMCEs. Notes: Bias and root mean square error are shown for marginal means in the first column of panels and for AMCEs in the second column, with the degree of bias for uncorrected estimators for each value of τ in the first row, corrected estimators in the second row, and a comparison between them in the third row, with values for the two quantities interest plotted by color.

Empirical Examples

Equation 5 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 differ-

ences 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 replication studies (see the Eigh Replications section). 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 available in our replication datasets. They include 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 produce the differences between uncorrected and corrected estimates 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 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 AMCEs and MMs, we find that the bias correction increases the subgroup difference for about 82% of the estimates, decreases it in about 12%, and switches the sign in about 5%. The size of the effect in each category has a wide range relative to the size of the original estimate.

We now offer three examples of substantive changes in real studies that result from correcting for measurement error. (Of course, authors should not be held responsible for "ignoring" correction methods that had not been invented at the time their articles were

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

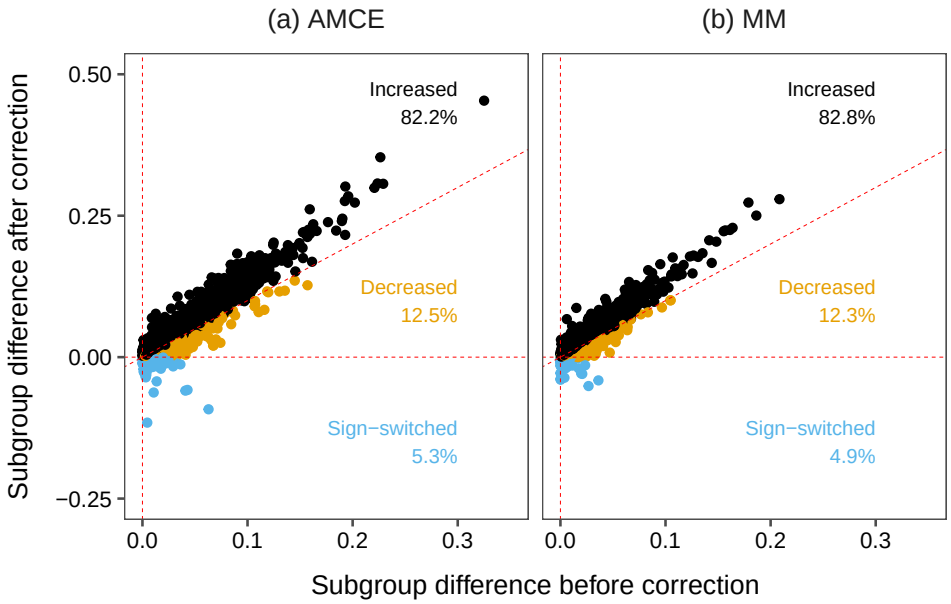

Figure 7: Consequences of Bias Correction in Eight Studies. Notes: The horizontal axis is the positive difference between the two subgroups, and the vertical axis is the corrected value for AMCEs (left panel) and MMs (right panel).

published.) First, Mummolo and Nall (2016) find that that Democrats and Republicans have small (and statistically insignificant) differences in preferences over neighborhood types. However, correcting their estimates for measurement error reveals that Democrats are in fact about four percentage points more opposed to communities in the top income bracket (with CI [0.03, 0.05]). Second, Arias and Blair (2022) reported that Republicans are more opposed to admitting hypothetical migrants to the country who are house cleaners than Democrats, but correcting for measurement error reveals that the reverse is actually the case. Finally, Hainmueller and Hopkins (2015), find statistically significant gender differences in preferences for immigrants who had previously entered the country without authorization (-0.04, with CI $[-0.05, -0.04]$), but evidence for this claim vanishes after correcting for measurement error (-0.005 with CI $[-0.01, 0.005]$).

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

Best Practices For Conjoint Analyses

Survey experiments can be greatly improved with conjoint designs if appropriately corrected for measurement error. Although hypothetical conjoint experiments without measurement error are more efficient than typical single task randomized experiments, the real world aspects of conjoint designs introduce more measurement error than traditional survey questions, reducing some of the efficiency gains. Furthermore, measurement error in conjoint analysis can induce substantial bias if ignored. Particularly when researchers are interested in subgroup comparisons, the bias may attenuate, exaggerate, or flip signs of the differences in MMs or AMCEs.

For all these reasons, researchers designing and fielding conjoint experiments should understand the nature of measurement error in conjoint designs and address this issue explicitly. In this section, we make four practical recommendations.

First, *researchers should try to reduce measurement error in the design phase whenever feasible*. To do this, they should follow best practices in standard survey design, such as via "cognitive debriefing," where researchers administer a draft survey to a small sample of respondents and immediately go back to the start of the survey and ask the same respondents what they understood each question to mean. Researchers should repeat this procedure while continuously adjusting their survey instruments, perhaps multiple times. Conjoint analyses are more complicated to understand than traditional survey questions, making this standard advice especially valuable.

Second, *researchers who wish to use conjoint designs other than binary choice should conduct further measurement error studies*. Our research shows how to correct measurement error bias for one type of conjoint analysis, a forced binary choice, which is by far the most commonly used in political science and marketing literatures. Other designs, such as rating the profiles individually, ranking the profiles, or choosing a single profile out of more than two, may also be valuable, depending on the research questions. However, we would expect that measurement error bias to be worse in all of these other types of designs

and so researchers who wish to try these alternative question types should consider how to measure, evaluate, and correct measurement error in those designs as well.

Finally, *researchers who administer binary choice conjoint experiments should use one of the bias correction methods proposed in this paper*. Mean squared error can be reduced by estimating IRR and applying the simple correction methods for MMs and AMCEs (see Equation 11). Specifically, we suggest that researchers choose among four approaches to estimating IRR (ordered by the simplicity of application):

1. If your research topic is similar enough to one or more of the studies we replicate—in both substantive content and target population—use the corresponding estimate of IRR 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. Still, one should clearly justify the choice of a particular value of IRR (e.g., 0.75).

2. You can estimate IRR from an existing conjoint without new data collection by extrapolating patterns from existing data, as we show in the Estimation Wihout Additional Data section. An estimate from this method can also be qualitatively averaged with the first option if one of the studies we replicated is similar to the one you are analyzing. Our software implements these and other methods for estimating IRR.

3. If you are in the planning stage of a conjoint study, we recommend adding a repeat of the first task presented at the end (and with two profiles and the order of the two columns switched). This enables researchers to estimate IRR by simply computing the percent agreement between the first and last questions and averaging over all respondents or the relevant subgroup. To use this design to correct bias, the researcher would rely on the extensive empirical evidence we offer that IRR 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 IRR 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 on the order of n for each estimate.

Although we recommend that most researchers adopt the third strategy, these researchers can still check whether IRR varies over selected types of profile-pair combinations by grouping them in different ways. Researchers should also use our replication packages, which include a large number of observations for many types of studies. Examining the replication data in greater detail may help them discover patterns of IRR related to their research.

Concluding Remarks

Through theoretical, simulation-based, and empirical evidence, we show that measurement error in conjoint designs can induce substantial bias in estimates of descriptive and causal effects—on average, within subgroups, and for subgroup differences. We show that measurement 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 measurement error-induced biases. These bias corrections will often make effects larger, but not in all situations. In particular, measurement error can lead to attenuation, exaggeration, or sign switches when researchers compare MMs or AMCEs between subgroups of respondents.

Our approach applies only to the most common type of conjoint design with a binary choice outcome variable. Valuable future research would include studying 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 demands these alternative conjoint designs place on respondents may lead to even higher levels and more complicated forms of measurement error than for binary choice outcomes. But, at this point, using these alternative conjoint designs without this research would put a researcher's results and conclusions at unnecessary risk.

All research conducted in this paper meet the American Political Science Association's 2020 statement on *Principles and Guidance for Human Subjects Research*. See Footnote 8 and our replication dataset.

References

Abramson, Scott F, Korhan Kocak, Asya Magazinnik, and Anton Strezhnev. 2023. *Detecting Preference Cycles in Forced-Choice Conjoint Experiments*. Working paper. URL: https://osf.io/preprints/socarxiv/xjre9_v1.

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

Achen, Christopher H. 1975. "Mass political attitudes and the survey response". *American Political Science Review* 69.4, pp. 1218–1231.

Allenby, Greg M, Nino Hardt, and Peter E Rossi. 2019. "Economic foundations of conjoint analysis". *Handbook of the Economics of Marketing*. Vol. 1. Amsterdam: Elsevier, pp. 151–192.

Arias, Sabrina B. and Christopher W. Blair. 2022. "Changing Tides: Public Attitudes on Climate Migration". *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". *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". *Political Analysis* 26.1, pp. 112–119.

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

Bansak, Kirk and Libby Jenke. 2023. "Odd Profiles in Conjoint Experimental Designs: Effects on Survey-Taking Attention and Behavior". *Political Analysis*, pp. 1–24.

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

Berinsky, Adam J, Michele F Margolis, and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys". *American journal of political science* 58.3, pp. 739–753.

Blackman, Alexandra Domike. 2018. "Religion and foreign aid". *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". *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". *Health Economics* 9.5, pp. 385–395.

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

Converse, Philip E. 2000. "Assessing the capacity of mass electorates". *Annual review of political science* 3.1, pp. 331–353.

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

Curran, Paul G. 2016. "Methods for the detection of carelessly invalid responses in survey data". *Journal of Experimental Social Psychology* 66, pp. 4–19.

De la Cuesta, Brandon, Naoki Egami, and Kosuke Imai. 2022. "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution". *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". *Demography* 43.3, pp. 569–585.

Ganter, Flavien. 2023. "Identification of preferences in forced-choice conjoint experiments: Reassessing the quantity of interest". *Political Analysis* 31.1, pp. 98–112.

Gilbert, Daniel, Gary King, Stephen Pettigrew, and Timothy Wilson. 2016. "Comment on 'Estimating the Reproducibility of Psychological Science'". *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". *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". *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". *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". *Political Analysis* 22.1, pp. 1–30.

Ham, Dae Woong, Kosuke Imai, and Lucas Janson. 2022. "Using machine learning to test causal hypotheses in conjoint analysis". *Political Analysis*, pp. 1–16.

Hankinson, Michael. 2018. "When do renters behave like homeowners? High rent, price anxiety, and NIMBYism". *American Political Science Review* 112.3, pp. 473–493.

Horiuchi, Yusaku, Shiro Kuriwaki, and Daniel M. Smith. Forthcoming. "Winning Elections with Unpopular Policies: Valence Advantage and Single-Party Dominance in Japan". *Quarterly Journal of Political Science*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4371978.

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

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. "Misunderstandings Among Experimentalists and Observationalists about Causal Inference". *Journal of the Royal Statistical Society, Series A* 171, part 2, pp. 481–502. URL: j.mp/misunEO.

Imbens, Guido W. 2004. "Nonparametric estimation of average treatment effects under exogeneity: a review". *Review of Economics and Statistics* 86.1, pp. 4–29.

Jenke, Libby, Kirk Bansak, Jens Hainmueller, and Dominik Hangartner. 2021. "Using Eye-Tracking to Understand Decision-Making in Conjoint Experiments". *Political Analysis* 29.1, pp. 75–101.

Kane, John V and Jason Barabas. 2019. "No harm in checking: Using factual manipulation checks to assess attentiveness in experiments". *American Journal of Political Science* 63.1, pp. 234–249.

King, Gary. 1995. "Replication, Replication". *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". *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". *Political Analysis* 14.2, pp. 131–159. URL: j.mp/dangerEC.

Lazarsfeld, Paul F. 1948. “The use of panels in social research”. *Proceedings of the American Philosophical Society* 92.5, pp. 405–410.

Leeper, Thomas J, Sara B. Hobolt, and James Tilley. 2020. “Measuring subgroup preferences in conjoint experiments”. *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”. *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”. *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”. *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”. *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”. *Political Behavior* 41.3, pp. 583–607.

Open Science Collaboration. 2015. “Estimating the reproducibility of psychological science”. *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”. *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”. *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". *American Political Science Review* 112.3, pp. 525–541.

Ward, MK and Adam W Meade. 2023. "Dealing with careless responding in survey data: Prevention, identification, and recommended best practices". *Annual Review of Psychology* 74, pp. 577–596.

Zaller, John and Stanley Feldman. 1992. "A simple theory of the survey response: Answering questions versus revealing preferences". *American journal of political science*, pp. 579–616.

Zaller, John R. 1992. *The nature and origins of mass opinion*. Cambridge: Cambridge university press.

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