Online Appendix

Addressing Measurement Errors in Ranking Questions for the Social Sciences

Appendix A Technical Discussions

We define a reference choice set, which will determine what permutation patterns such as 123 indicate. For example, let $\{a, b, c\}$ be our reference choice set, a finite and ordered set of $J = 3$ items. Then, a non-random response $(1, 3, 2)$ means that a respondent has the ranked preference $a \succ c \succ b$ Formally, let $\mathcal{P}_J$ be a *permutation space* of any reference set of size J that contains all $J!$ possible permutations of J numbers. When $J = 3$, for example, the permutation space is $\mathcal{P}_{J=3} = \{123, 132, 213, 231, 312, 321\}$, shorthand for all possible underlying preferences $\{a \succ b \succ c, a \succ c \succ b, b \succ a \succ c, b \succ c \succ a, c \succ a \succ b, c \succ b \succ a\}$. A *ranking function* Y_i for any respondent i $(i = 1, 2, \cdots, n)$ is defined as a mapping from the finite labeled set $\mathcal{A}$ to the permutation space $\mathcal{P}_J$. For simplicity, we denote $Y_i^{\text{obs}}(\mathcal{A})$ simply as Y_i^{obs}. Similarly, we will denote $g(Y_i^{\text{obs}}(\mathcal{A}))$ as $g(Y_i^{\text{obs}})$.

A.1 Bias from Random Responses

In this subsection, we quantify the bias from random responses and discuss its consequences in empirical studies.

Let Y_i^* and e_i be respondent i's $(i = 1, \cdots, N)$ non-random and random responses (or errors), respectively. Let z_i be a random variable denoting whether respondent i offers a non-random response $(z_i = 1)$ or otherwise $(z_i = 0)$. We denote respondent i's *observed* response by $Y_i^{\text{obs}} = Y_i^* z_i + e_i(1 - z_i)$.

We use a general notation $g(Y_i^*)$ to represent a ranking-based quantity of interest (QOI). We assume some linear operator for $g(\cdot)$. Here, $Y_i^* = (Y_{i,1}^*, Y_{i,2}^*, \cdots, Y_{i,J}^*)$, where $Y_{i,j}^*$ is the marginal rank of item j provided by i, e.g., $(1, 3, 2)$. Similarly, we denote $Y_i^{\text{obs}} = (Y_{i,1}^{\text{obs}}, Y_{i,2}^{\text{obs}}, \cdots, Y_{i,J}^{\text{obs}})$ to be their observed counterparts.

Researchers should not ignore random responses because the measurement error they induce will lead to a biased estimate of their quantity of interest. To demonstrate this bias, we focus on the expected value of item j's marginal rank (or average rank) among non-random responses as our quantity of interest. Formally, we can characterize the bias

as follows:

$$\text{Bias} = \underbrace{\mathbb{E}[\widehat{\mathbb{E}}(Y_{i,j}^{\text{obs}})]}_{\text{estimate based on raw data}} - \underbrace{\mathbb{E}[Y_{i,j}^{*}|z_i = 1]}_{\text{what we wish to study}} \tag{A.1a}$$

$$= \underbrace{-(\mathbb{E}[Y_{i,j}^{*}|z_i = 1] - \mathbb{E}[e_{i,j}|z_i = 0])}_{\text{difference between non-random and random responses}} \times \underbrace{\Pr(z_i = 0)}_{\text{prop. random responses}} \tag{A.1b}$$

The bias grows as (1) the difference between non-random and random responses increases or (2) the proportion of random responses increases.

In empirical applications, this is highly problematic because the bias is unpredictable (i.e., it depends on both the unknown quantity of interest and the unknown distribution offered by random responses) and, thus, the measurement error can bias researchers' estimates in any direction (e.g., may over or underestimate the average rank of a certain item). Consequently, unless we assume that random responses do not exist at all, if researchers do not address random responses, their conclusion will be based on a quantity that differs from their QOI. Moreover, they have no way to ascertain how much their estimates deviate from the QOI.

A.2 The Role of Item Order Randomization

This subsection provides proof for our result that the rankings based on random responses follow a uniform distribution under item order randomization.

Let $\mathcal{A}$ be the reference choice set (i.e., reference item order), and let $\mathcal{A}_i$ be an *observed* choice set (i.e., observed item order) for respondent i. We use the terms "choice set" and "item order" interchangeably. When an item order is fixed, the observed choice set is identical to the reference choice set for all respondents. That is, $\mathcal{A}_i = \mathcal{A}$ for all i. For example, when all respondents may see $\{a, b, c\}$, ranking (1, 2, 3) means that respondents prefer a to b to c, i.e., $a \succ b \succ c$.

We now introduce two closely related concepts. First, we continue using Y_i to represent the *observed ranking* for respondent i with respect to the reference choice set. Second, we use R_i to denote the *recorded response* for respondent i—a provided ranking (or ordering) with respect to i's observed choice set a^*. For example, a recorded response may

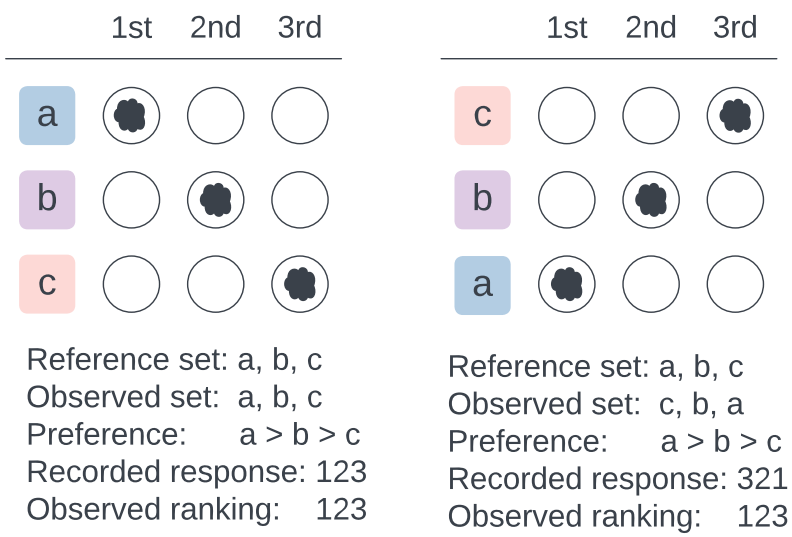


Figure A.1: Difference Between the Reference Choice Set, Observed Choice Set, Recorded Response, and Observed Ranking

be (3, 2, 1) with respect to the observed choice set $\{c, b, a\}$, whereas the reference choice set is $\{a, b, c\}$. Then, the corresponding observed ranking is (1, 2, 3), indicating that the respondent prefers a to b to c. In ranking data analysis, we wish to study observed rankings and not recorded responses. Figure A.1 shows the difference between the reference choice set, respondent-by-respondent observed choice set, recorded response, and observed ranking.

Recorded responses are identical to observed rankings when item orders are fixed. However, when an item order is not fixed, the observed choice set differs across respondents. Formally, we express this idea by writing $\mathcal{A}_i = o_i(\mathcal{A})$, where $o_i(\cdot)$ is a known function that reorders the available items in the reference choice set for respondent i.

Let $h(\cdot)$ be a deterministic function that transforms recorded responses to observed rankings based on observed choice sets. That is, $Y_i = h(R_i, A_i)$. For example, some respondents may see $\{a, b, c\}$, while other respondents may see $\{c, a, b\}$. In this setting, recorded response (1, 2, 3) has two substantively different meanings. For the first group, it means that people prefer a to b to c. For the second group, it means that they prefer c to a to b. Again, our substantive interest lies in people's observed rankings instead of recorded responses.

Given the above notation, we now discuss the role of item order randomization. Item

order randomization allows us to make two key assumptions, which we leverage to learn about the distribution of observed rankings based on random responses. The first assumption states that respondents see each observed choice set with equal probability. The second assumption posits that observed choice sets and peoples' underlying preferences are statistically independent.

Let $a^* \in \mathcal{A}$ be a realized observed choice set from the set of all possible choice sets. Let $\underline{Y}_i$ be respondent i's underlying preference. We formally state the two assumptions as follows.

Assumption A.1 (*Random Item Order*)**.** Each possible observed choice set appears with equal probability. That is, with J items, $\Pr(\mathcal{A}_i = a^*) = \frac{1}{J!}$ for all a^*.

Assumption A.2 (*Order Ignorability*)**.** A presented item order is statistically independent of one's underlying preference. That is, $\underline{Y}_i \perp \mathcal{A}_i$ for all i.

With these ideas, we formally define random recorded responses as follows.

Definition 1 (*Random Recorded Responses*)**.** Random recorded responses are statistically independent of people's underlying preferences and observed choice sets. Formally, $R_i(z_i = 0) \perp (\underline{Y}_i, \mathcal{A}_i)$ and thus $e_i \perp (\underline{Y}_i, \mathcal{A}_i)$.

Finally, we present our key result—item order randomization makes random responses follow a uniform distribution. Suppose respondent i provides observed ranking y^* by offering random response r^* when exposed to item order a^*. Suppose also that the same respondent has underlying preference $\underline{y}^*$.

Then, the probability that respondent i offers observed ranking y^* (e.g., 312) when providing random response r^* is $\frac{1}{J!}$. We present proof below. Our result is general in that it holds regardless of people's underlying preference $\underline{y}^*$ (e.g., 123 vs. 231) and random response pattern r^* (e.g., diagonalizing vs. zig-zagging).

Proof.

$$\Pr(e_i = y^*) = \underbrace{\Pr(e_i = y^*, R_i = r^*, \underline{Y}_i = \underline{y}^*, \mathcal{A}_i = a^*)}_{\text{can be expanded via law of conditional probability}} \tag{A.2a}$$

$$= \underbrace{\Pr(e_i = y^* | R_i = r^*, \underline{Y}_i = \underline{y}^*, \mathcal{A}_i = a^*)}_{\text{can be simplified via Definition 1}} \Pr(R_i = r^*, \underline{Y}_i = \underline{y}^*, \mathcal{A}_i = a^*) \tag{A.2b}$$

$$= \Pr(e_i = y^* | R_i = r^*) \underbrace{\Pr(R_i = r^*, \underline{Y}_i = \underline{y}^*, \mathcal{A}_i = a^*)}_{\text{can be factorized via Definition 1}} \tag{A.2c}$$

$$= \Pr(e_i = y^* | R_i = r^*) \Pr(R_i = r^*) \quad \underbrace{\Pr(\underline{Y}_i = \underline{y}^*, \mathcal{A}_i = a^*)}_{\text{can be factorized via Assumption A.2}} \tag{A.2d}$$

$$= \underbrace{\Pr(e_i = y^* | R_i = r^*)}_{\text{1 by definition}} \underbrace{\Pr(R_i = r^*)}_{\text{1 by definition}} \underbrace{\Pr(\underline{Y}_i = \underline{y}^*)}_{\text{1 by definition}} \underbrace{\Pr(\mathcal{A}_i = a^*)}_{\frac{1}{J!} \text{via Assumption A.1}} \tag{A.2e}$$

$$= 1 \times 1 \times 1 \times \frac{1}{J!} \tag{A.2f}$$

$$= \frac{1}{J!} \tag{A.2g}$$

□

A.3 Unbiased Estimator of $\Pr(z_i^{\text{anc}} = 0)$

In this subsection, we provide proof for the unbiasedness of the proposed estimator of the proportion of random responses based on correct answers in an anchor question.

Let c_i be a binary random variable that takes 1 if respondent i offers the correct answer in the anchor and 0 otherwise. Let $\Pr(z_i^{\text{anc}} = z)$ be the probability that respondent i provides a non-random ($z = 1$) or random response ($z = 0$) in the anchor question, where $z \in \mathcal{Z} = \{0, 1\}$.

Let $A = \Pr(z_i^{\text{anc}} = 1)$ be the proportion of non-random responses in an anchor question. Let $B = \Pr(c_i = 1 | z_i^{\text{anc}} = 0)$ be the probability that random responses happen to be correct in the anchor. We assume that all non-random responses are the correct answer; namely, $\Pr(c_i = 1 | z_i^{\text{anc}} = 1) = 1$.

The probability that respondent i provides the correct answer can be expressed as

follows:

$$\Pr(c_i = 1) = \sum_{z \in \mathcal{Z}} \Pr(c_i = 1 | z_i^{\text{anc}} = z) \Pr(z_i^{\text{anc}} = z) \tag{A.3a}$$

$$= \Pr(c_i = 1 | z_i^{\text{anc}} = 1) \Pr(z_i^{\text{anc}} = 1) + \Pr(c_i = 1 | z_i^{\text{anc}} = 0) \Pr(z_i^{\text{anc}} = 0) \tag{A.3b}$$

Given that, we can rewrite the proportion of non-random responses A as follows:

$$\Pr(c_i = 1) = 1 \cdot A + B \cdot (1 - A) \tag{A.4a}$$

$$\Pr(c_i = 1) = A + B(1 - A) \tag{A.4b}$$

$$\Pr(c_i = 1) = A + B - BA \tag{A.4c}$$

$$\Pr(c_i = 1) - B = A - BA \tag{A.4d}$$

$$\Pr(c_i = 1) - B = (1 - B)A \tag{A.4e}$$

$$A = (\Pr(c_i = 1) - B)(1 - B)^{-1} \tag{A.4f}$$

From the above subsection (Equation A.2g), we also know the following:

$$B = \frac{1}{J!} \tag{A.5}$$

That is, the probability that random responses are the correct answer is $\frac{1}{J!}$.

Integrating the above results, we can rewrite the proportion of non-random responses as follows:

$$\Pr(z_i^{\text{anc}} = 1) = \left[\Pr(c_i = 1) - \frac{1}{J!} \right] \left(1 - \frac{1}{J!} \right)^{-1} \tag{A.6}$$

Finally, we propose an unbiased estimator of the proportion of non-random answers:

$$\widehat{\Pr}(z_i^{\text{anc}} = 1) = \left[\frac{\sum_{i=1}^{N} c_i}{N} - \frac{1}{J!} \right] \left(1 - \frac{1}{J!} \right)^{-1}, \tag{A.7}$$

where $\frac{\sum_{i=1}^{N} c_i}{N} \xrightarrow{p} \Pr(c_i = 1)$ via the weak law of large numbers.

The right-hand side of the above result has an intuitive substantive interpretation. Naively, one may think that the proportion of correct answers is the proportion of non-random responses. However, this idea does not account for the fact that some random responses happen to be correct by chance. Thus, one may overestimate the proportion of non-random responses by only using the proportion of correct answers. One can interpret each term in the above result as it adjusts for this overestimation.

The first part indicates that we need to adjust for the overestimation of non-random responses by adjusting the observed proportion of correct answers for the probability that random responses happen to be correct. The second part indicates that we need to normalize the resulting quantity so that it will be a proper probability.

Our result in Equation A.7 holds under the regularity condition of

$$\frac{\sum_{i=1}^{N} c_i}{N} > \frac{1}{J!} \tag{A.8}$$

For example, with four items ($J = 3$), the sample proportion of correct answers must be greater than $\frac{1}{3!} = \frac{1}{6} \approx 0.167$. Otherwise, our estimated proportion of non-random responses becomes negative as $\frac{\sum_{i=1}^{N} c_i}{N} - \frac{1}{J!} < 0$. The regularity condition can be violated when (1) the proportion of random responses is relatively high and (2) the sample proportion of correct answers among random responses is less than $\frac{1}{J!}$ due to sampling variability.

A.4 The Uniformity Test

In this subsection, we provide proof for the uniformity test under item order randomization. We first formally define non-random responses as follows:

Definition 2 (*Non-random Responses*)**.** Among non-random responses, observed rankings are a function of people's underlying preferences and observed choice sets. Formally, $R_i(z_i = 1) = h(\underline{Y}_i, \mathcal{A}_i)$ such that $\Pr(R_i = r^* | \underline{Y}_i = y^*, \mathcal{A}_i = a^*) = 1$.

Then, the probability that any recorded ranking among non-random responses is a specific ranking p^* is $\frac{1}{J!}$, given item order randomization. More generally, recorded responses provided by non-random responses follow a uniform distribution $\mathcal{U}_J$.

Proposition 1. Uniformity of Recorded Responses Among Non-random Responses

$$\mathbb{P}(R_i|z_i = 1) \sim \mathcal{U}_J = \left(\frac{1}{J!}, \frac{1}{J!}, \cdots, \frac{1}{J!}\right)$$

Proof.

$$\Pr(R_i = r^*|z_i = 1) = \underbrace{\Pr(R_i = r^*, \underline{Y_i^*} = y^*, \mathcal{A}_i = a^*)}_{\text{can be expanded via law of conditional probability}} \tag{A.9a}$$

$$= \Pr(R_i = r^*|\underline{Y_i^*} = y^*, \mathcal{A}_i = a^*) \underbrace{\Pr(\underline{Y_i^*} = y^*, \mathcal{A}_i = a^*)}_{\text{can be factorized via Assumption A.2}} \tag{A.9b}$$

$$= \underbrace{\Pr(R_i = r^*|\underline{Y_i^*} = y^*, \mathcal{A}_i = a^*)}_{1 \text{ via Definition 2}} \underbrace{\Pr(\underline{Y_i^*} = y^*)}_{1 \text{ by definition}} \underbrace{\Pr(\mathcal{A}_i = a^*)}_{\frac{1}{J!} \text{via Assumption A.1}} \tag{A.9c}$$

$$= 1 \times 1 \times \frac{1}{J!} \tag{A.9d}$$

$$= \frac{1}{J!} \tag{A.9e}$$

□

This suggests that when all respondents offer non-random responses under item order randomization, the distribution of recorded responses should follow a uniform distribution. To test whether all respondents are indeed non-random responses, researchers can then apply Pearson's χ^2 test for uniformity to the empirical distribution of observed rankings; that is, $\widehat{\mathbb{P}}(R_i)$. If the test rejects the null hypothesis of uniformity, observed data may contain random responses.

For completeness, we also show that recorded responses provided by random responses will *not* follow the uniform distribution and instead follow some unknown distribution, representing certain patterns provided by random responses.

Proposition 2. Non-uniformity of Recorded Responses Among Random Responses

$$\mathbb{P}(R_i|z_i = 0) \sim \mathcal{R}_J \neq \left(\frac{1}{J!}, \frac{1}{J!}, \cdots, \frac{1}{J!}\right)$$

The proof is straightforward.

Proof.

$$\Pr(R_i = r^*|z_i = 0) = \underbrace{\Pr(R_i = r^*, \underline{e_i} = y^*, \mathcal{A}_i = a^*)}_{\text{can be expanded via law of conditional probability}} \tag{A.10a}$$

$$= \Pr(R_i = r^*|\underline{e_i} = y^*, \mathcal{A}_i = a^*) \quad \underbrace{\Pr(\underline{e_i} = y^*, \mathcal{A}_i = a^*)}_{\text{can be factorized via Assumption A.2}} \tag{A.10b}$$

$$= \underbrace{\Pr(R_i = r^*|\underline{e_i} = y^*, \mathcal{A}_i = a^*)}_{\text{dependency goes away via Definition 1}} \underbrace{\Pr(\underline{e_i} = y^*)}_{\text{1 by definition}} \underbrace{\Pr(\mathcal{A}_i = a^*)}_{\frac{1}{J!} \text{via Assumption A.1}} \tag{A.10c}$$

$$= \Pr(R_i = r^*) \times 1 \times \frac{1}{J!} \tag{A.10d}$$

$$= \frac{\Pr(R_i = r^*)}{J!} \tag{A.10e}$$

□

Appendix B Illustrations of the Uniformity Test

In Appendix A.4, we show that when all respondents offer non-random responses under item order randomization, the distribution of recorded responses follows a uniform distribution with probability $\frac{1}{J!}$. In contrast, when all respondents offer random responses, the distribution of recorded responses represents the underlying patterns that respondents fall to. Using this result, researchers can detect the potential presence of random answers by checking whether the distribution of recorded responses follows a uniform via visualizations and uniformity tests such as Pearson's χ^2 test.[1] We call it the uniformity test, and this section illustrates it with simulated and empirical data.

B.1 Simulation Studies

First, we illustrate the uniformity test with simulated data. The uniformity test checks whether the distribution of recorded rankings $\{R_i\}_{i=1}^N$ follows a uniform distribution through data visualization and Pearson's χ^2 test. Figure B.2 presents an example with simulated ranking data. To demonstrate, we generated three survey data with $J = 3$ and $N = 2,000$ and visualized the empirical distributions of recorded responses. We set that all respondents have a strict preference $c \succ b \succ a$ with respect to the reference choice set $\mathcal{A} = \{a, b, c\}$ and thus (observed) ranking $\underline{Y}_i = 321$.

Panels A–C represent the distributions of recorded responses based on our simulated data researchers would observe in the absence of item order randomization. Thus, recorded responses are identical to observed rankings. We set all respondents to offer non-random responses in the first data (Panel A). In the second data (Panel B), we set that all respondents offer random responses and have an orientation towards some zig-zag random patterns. Finally, in the third data (Panel C), we set that 50% of respondents offer non-random responses and the other 50% offer random responses as in Panel B.

Suppose that Panel C is the most realistic example of actual survey data researchers would observe. The fundamental problem of contaminated sampling is that researchers can *never* know whether the result in Panel C is susceptible to random responses and, if so, to what extent (in the absence of item order randomization).

Next, Panels D–F represent the same distributions researchers would observe with

[1] Recently, Atsusaka (2024) leverages this idea to study ballot order effects in ranked-choice voting.

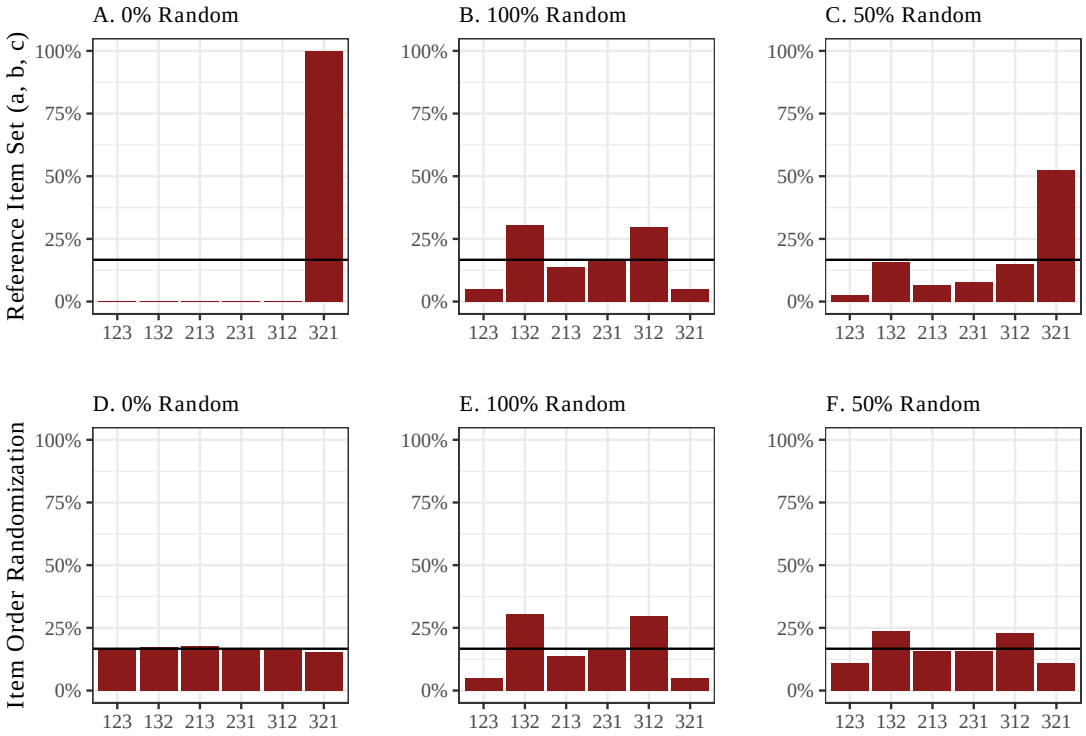


Figure B.2: Simulated Distributions of Recorded Responses with and without Item Order Randomization

Note: A vertical line is drawn at $1/6 \cdot 100\%$. In Panels A–C, rankings are defined with respect to a reference item set. In Panels D–F, rankings are defined with respect to observed item sets.

item order randomization. Here, recorded responses (visualized in the panels) may not be identical to observed rankings. Panel D shows that when all respondents offer non-random responses, the distribution of recorded responses follows the uniform distribution with $\frac{1}{J!}$, consistent with Proposition 1. Next, Panel E shows that the distribution is not uniform, and it is, in fact, the same as in Panel B. This means that random responses are independent of observed choice sets as assumed in Assumption A.2, and the result is consistent with Proposition 2.

Finally, Panel F shows the distribution of recorded responses when 50% of respondents offer random and non-random responses, respectively. In applied research, this is what researchers would observe with item order randomization. The goal of the uniformity test is to check whether this distribution follows a uniform distribution. While researchers are unable to conclude whether Panel C contains random responses, they are

able to tell that the data presented in Panel F are contaminated by random responses. This is because we can see that the distribution of recorded responses does not follow a uniform distribution. In addition to the visual inspection, Pearson's χ^2 test rejects the null hypothesis that the distribution in Panel F follows the uniform distribution at the $\alpha = 0.05$ level.

B.2 Empirical Studies

What random patterns emerge as prominent when non-uniform ranking patterns are discovered? To empirically investigate this, we include *no-context ranking questions* (with three and four items) in our survey, which are completely stripped of contextual information. Figure B.3 shows the design of our no-context question with three items. Respondents were prompted to provide a ranking over options that provided equally empty values after being informed that the options were left intentionally blank. Given that the options were not differentiable and contained no substantial information, respondents are freely allowed to respond randomly, and we can fully observe a distribution of random responses, albeit in an extreme case.

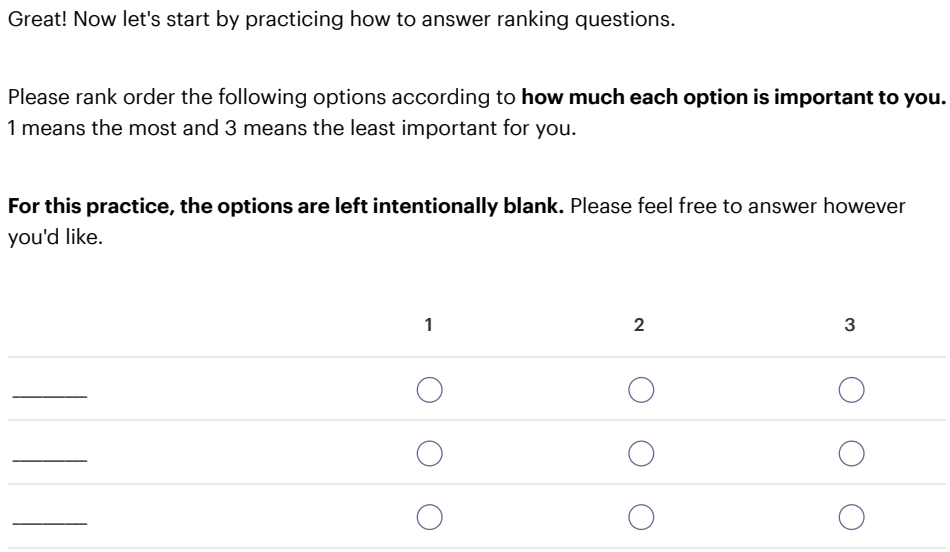
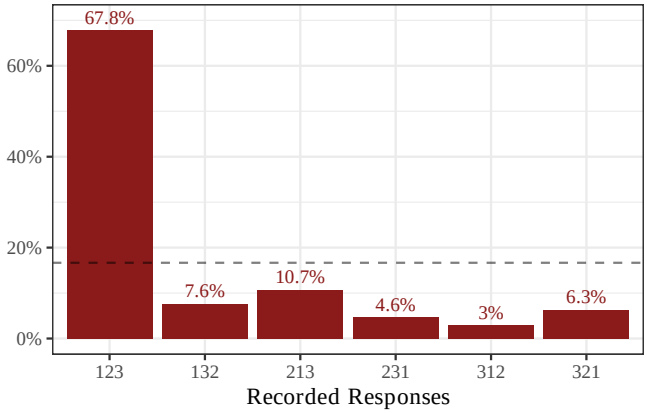


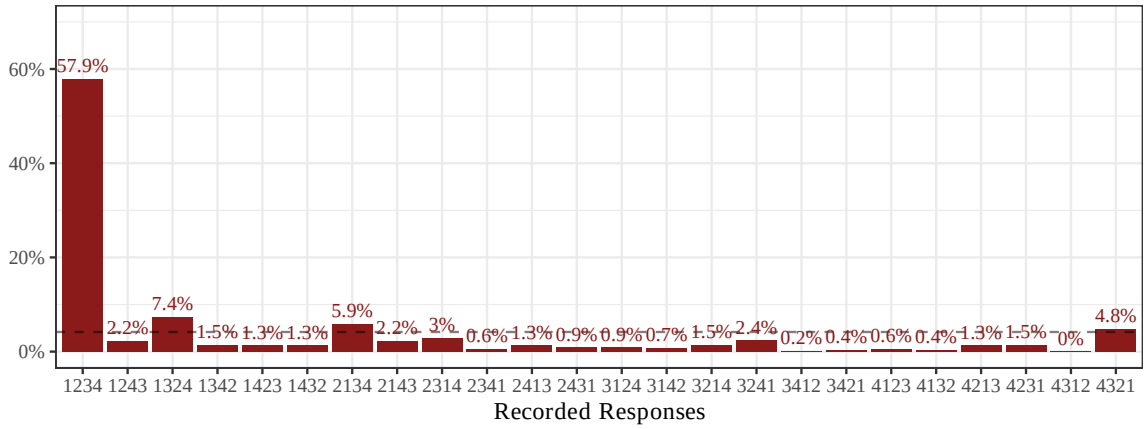
Figure B.3: Design of the No-context Ranking Question

Figure B.4 presents the distribution of recorded responses. We can see that the recorded responses show a clear non-uniform distribution ($p < 0.001$) and a clear prevalence of diagonalizing, which occurs more than 67% (Panel A) and 57% (Panel B) of the time. This aligns with the survey satisficing literature, as we could call diagonalizing

the ranking questions' equivalent of straightlining (Leiner, 2019; Reuning and Plutzer, 2020). We also see some unintuitive patterns, such as the prevalence of 213 in the $J = 3$ case or the lack of 4312 in the $J = 4$ case. Why some permutations are more salient than others might be an interesting analysis that we leave for future research.



(a) With 3 Items ($N = 540$)



(b) With 4 Items ($N = 542$)

Figure B.4: Distributions of Recorded Responses for No-Context Ranking Questions

Note: the horizontal line shows $\frac{1}{J!}$, a value expected if the responses were uniformly distributed across the permutation space. Each respondent was randomly exposed to either $J = 3$ or $J = 4$ questions.

Appendix C Empirical Comparisons of Anchor Questions and the Alternative Approaches

This section provides empirical comparisons of anchor questions and alternative approaches to address measurement errors discussed in the main text.

C.1 Design and Flow of Survey Items for Checking Random Responses

In Section 6.1, we briefly provided a theoretical discussion on the alternative approaches to identifying θ_z via listwise deletion (or potentially estimating the proportion of random responses). In Appendix C.1, we show how these actually manifested in empirical data that we have collected and what implications the results may have. We also analyze different types of anchor questions here. Given that these alternative measures include both instructional and factual manipulation checks, we will simply refer to them as checks or alternative designs altogether.

We consider the following six designs:

1. Main anchor question
2. Alphabetical anchor question
3. Exact-order anchor question
4. Attention check I (adopted from Ternovski and Orr (2022))
5. Attention check II (variation of one from Berinsky et al. (2014))
6. Repeated question

In Figure C.16, we show how six different alternative designs, including the main anchor question presented in Figure 4, differ in their estimates of average ranks of the items based on different strategies. In this section, we first show what might account for the variations in each of these types of checks, as well as how they relate to each other.

We have already demonstrated what the main anchor question looks like. Recently, a similar problem of measurement errors has been discussed in conjoint experiments (Clayton et al., 2023). Similar to this work, another potential solution would be to ask the main ranking question twice in a survey and estimate the proportion of random responses: the design is, of course, the same as the main ranking question (shown in Figure 1), and it is asked again verbatim after a few unrelated questions. We also have

two attention checks (Berinsky et al., 2012, 2014; Hauser and Schwarz, 2016; Alvarez et al., 2019; Alvarez and Li, 2022; Ternovski and Orr, 2022; Berinsky et al., 2024; Tyler et al., 2024), sometimes called instructional manipulation checks, screeners, or bogus questions. The rest of the alternative designs look as follows:

- In *anchor alphabet*, we ask respondents to rank the items according to their alphabetical order, as in Figure C.5.

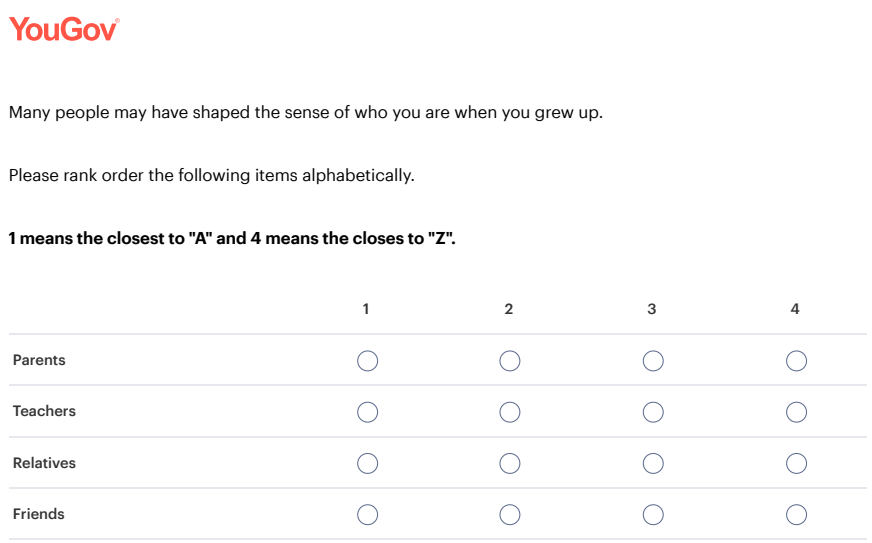


Figure C.5: Anchor Question Using an Alphabetical Order

- In *anchor exact*, we ask respondents to rank the items according to the exact order that we have specified in the question, as in Figure C.6.

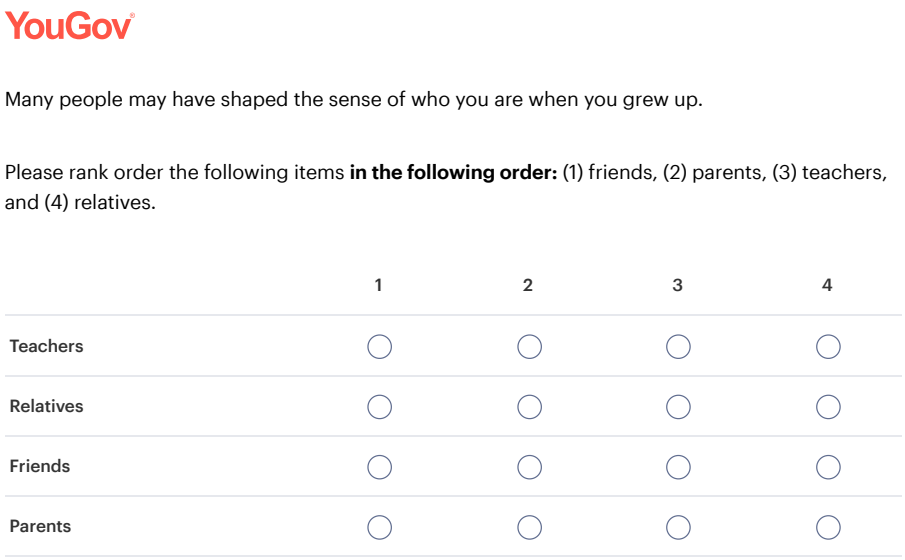


Figure C.6: Anchor Question Using an Exact Order

- In *Attention I*, we adopt an attention check from Ternovski and Orr (2022).

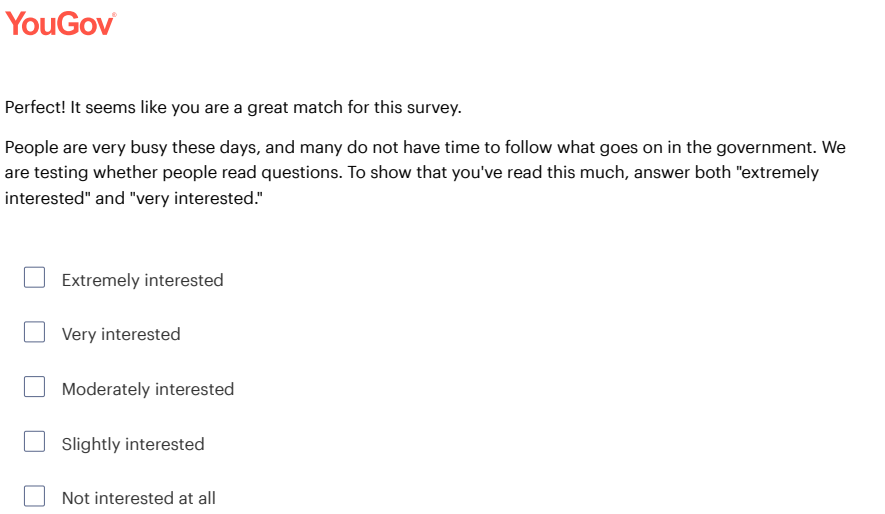


Figure C.7: Attention Check Adopted from Ternovski and Orr (2022)

- In *Attention II*, we adopt an attention check that is similar to the design shown in Berinsky et al. (2014), but adapted so that its contents are more related to the overall theme of the survey.

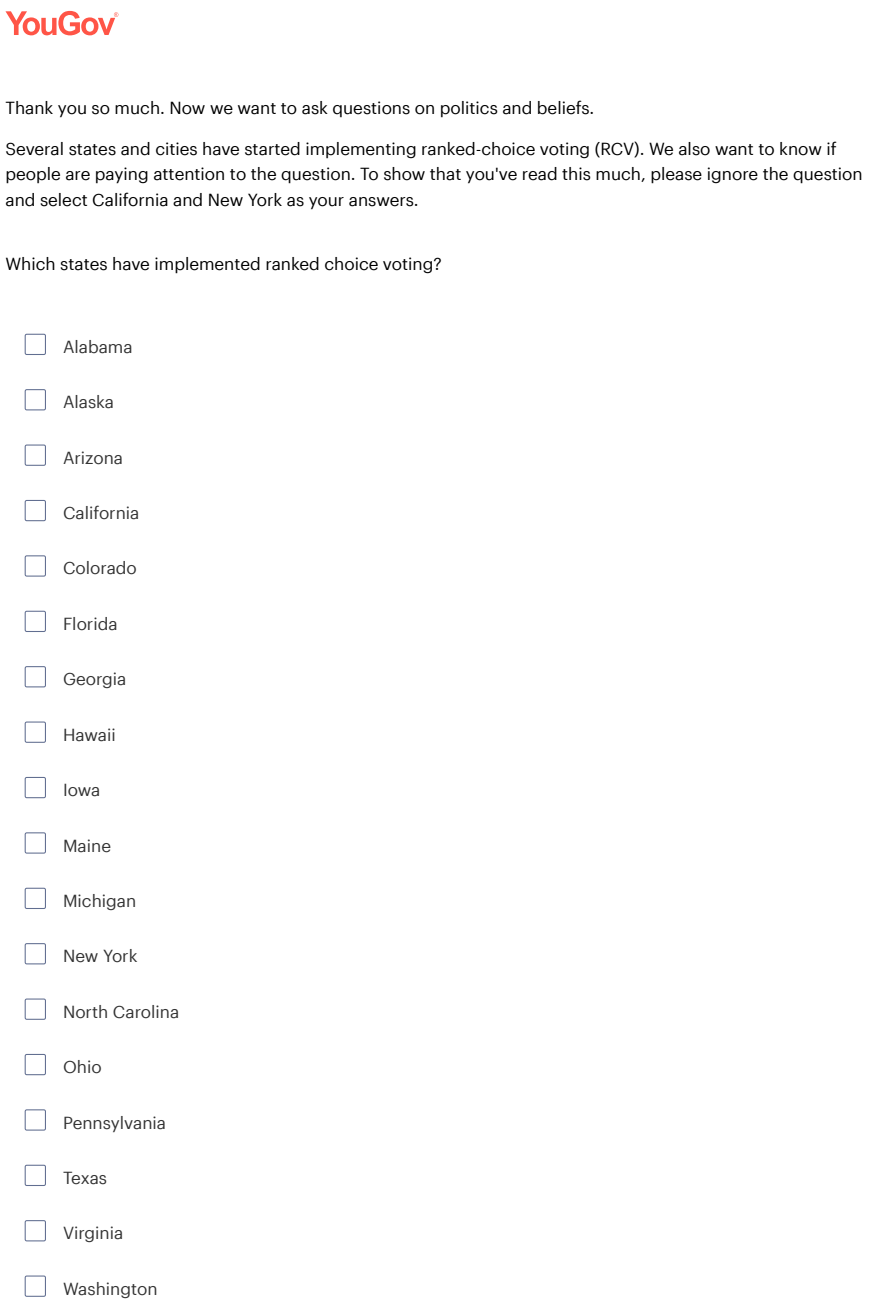


Figure C.8: Attention Check Similar to That of Berinsky et al. (2014)

Figure C.9 shows the order in which all alternative designs were tested in our YouGov survey. As can be seen, the anchor questions are situated adjacent to the main ranking question. Due to the survey length, only 50% of the respondents were exposed to the repeated question check. Respondents were randomly divided between the alphabet and the exact order ranking question. No statistical differences in the main ranking exercise

were detected between those who were exposed to the main anchor question and then the main ranking question versus the reverse.

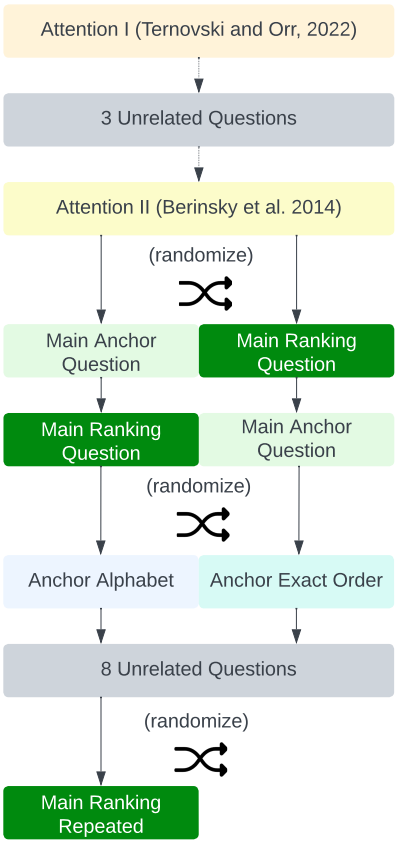


Figure C.9: Order of Survey Screeners for Checking Random Responses

Note that these are not the only alternative options that researchers can employ. Depending on the survey length, intention, and design, they may also come up with other domain-relevant methods. One such idea is to use latency measures (response time data) or frequencies of skipping a question.

C.2 Explaining Variations in Random Responses

Table C.1 predicts if any important demographic/sociopolitical variables predict the likelihood of randomness across all possible types of detections. Figure C.10 visualizes the regression results. Importantly, it shows the associations between covariates and random responses to these alternative designs and *not between covariates and underlying preferences*

	What Determins a Random Response?					
	Repeated	Attention I	Attention II	Anchor main	Anchor exact	Anchor alphabet
Independent	0.129	0.082	0.008	0.057	0.244*	0.139
	(0.100)	(0.058)	(0.059)	(0.075)	(0.107)	(0.094)
Republican	0.107	−0.042	−0.016	−0.090	−0.103	−0.121
	(0.090)	(0.052)	(0.054)	(0.060)	(0.086)	(0.077)
Ideology	−0.013	0.004	0.010	0.010	−0.016	0.047*
	(0.021)	(0.013)	(0.013)	(0.015)	(0.021)	(0.019)
Party ID intensity	0.025	0.014	0.024	0.046*	0.100**	0.071*
	(0.035)	(0.019)	(0.020)	(0.023)	(0.033)	(0.031)
Age	0.002	0.001	0.0004	−0.001	0.001	0.004*
	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Female	−0.039	−0.024	−0.028	−0.047	−0.040	0.018
	(0.054)	(0.033)	(0.032)	(0.038)	(0.056)	(0.051)
Neither male/female	−0.183	−0.060	−0.021	−0.222	−0.330	−0.106
	(0.137)	(0.100)	(0.101)	(0.118)	(0.225)	(0.161)
Black	0.165	0.030	0.025	0.116	0.111	0.173*
	(0.093)	(0.059)	(0.060)	(0.066)	(0.089)	(0.087)
Hispanic/Latino	0.091	−0.032	−0.011	−0.022	−0.131	0.057
	(0.092)	(0.049)	(0.046)	(0.058)	(0.084)	(0.098)
Other race	−0.160	−0.018	−0.009	0.014	0.215*	0.009
	(0.082)	(0.055)	(0.063)	(0.075)	(0.095)	(0.095)
Some college	−0.142*	−0.113**	−0.036	−0.044	−0.047	−0.198**
	(0.064)	(0.040)	(0.038)	(0.048)	(0.068)	(0.066)
College graduate	−0.161*	−0.077	0.034	−0.047	−0.079	−0.123
	(0.081)	(0.050)	(0.047)	(0.056)	(0.082)	(0.075)
Post-graduate	−0.122	−0.175***	−0.020	−0.209***	−0.113	−0.157
	(0.088)	(0.045)	(0.051)	(0.055)	(0.087)	(0.088)
Household income: 50-80k	−0.080	0.028	−0.023	0.030	0.024	−0.049
	(0.069)	(0.046)	(0.041)	(0.050)	(0.075)	(0.069)
Household income: 80-150k	−0.004	−0.014	−0.004	−0.026	0.028	−0.095
	(0.075)	(0.039)	(0.045)	(0.049)	(0.072)	(0.073)
Household income: 150k or more	−0.001	−0.003	0.007	0.073	0.009	−0.126
	(0.100)	(0.059)	(0.058)	(0.074)	(0.119)	(0.094)
Household income: prefer not to say	0.029	−0.093	−0.046	−0.056	−0.118	−0.270**
	(0.097)	(0.048)	(0.052)	(0.070)	(0.087)	(0.089)
Political interest	0.016	0.035	0.042*	0.045*	0.025	0.050
	(0.026)	(0.018)	(0.018)	(0.021)	(0.031)	(0.030)
Catholic	0.091	0.041	0.019	0.041	0.152	0.194*
	(0.087)	(0.051)	(0.051)	(0.060)	(0.089)	(0.080)
Jewish	0.360*	0.143	0.081	0.033	0.129	0.386**
	(0.150)	(0.089)	(0.108)	(0.098)	(0.124)	(0.139)
None/agnostic/atheist	0.017	−0.008	−0.032	0.014	−0.030	0.106
	(0.069)	(0.040)	(0.043)	(0.049)	(0.074)	(0.069)
Other religion	0.069	0.026	0.048	−0.019	0.075	0.053
	(0.085)	(0.050)	(0.059)	(0.060)	(0.087)	(0.091)
Region: midwest	−0.103	−0.016	−0.057	−0.056	−0.037	0.030
	(0.081)	(0.046)	(0.049)	(0.063)	(0.085)	(0.084)
Region: south	−0.017	0.002	−0.036	−0.102	−0.028	0.120
	(0.077)	(0.043)	(0.046)	(0.057)	(0.077)	(0.075)
Region: west	−0.029	0.085	−0.007	0.005	0.069	0.158
	(0.083)	(0.050)	(0.052)	(0.062)	(0.079)	(0.086)
Observations	549	1,081	1,081	1,081	545	536
R^2	0.096	0.067	0.038	0.075	0.105	0.152
Adjusted R^2	0.053	0.045	0.015	0.053	0.061	0.110
Note:						*p<0.05; **p<0.01; ***p<0.001

Table C.1: Predicting Random Responses to Anchor Questions, Failure to Answer Consistently for Repeated Questions, and Failure to Pass Attention Checks

within the target ranking question.

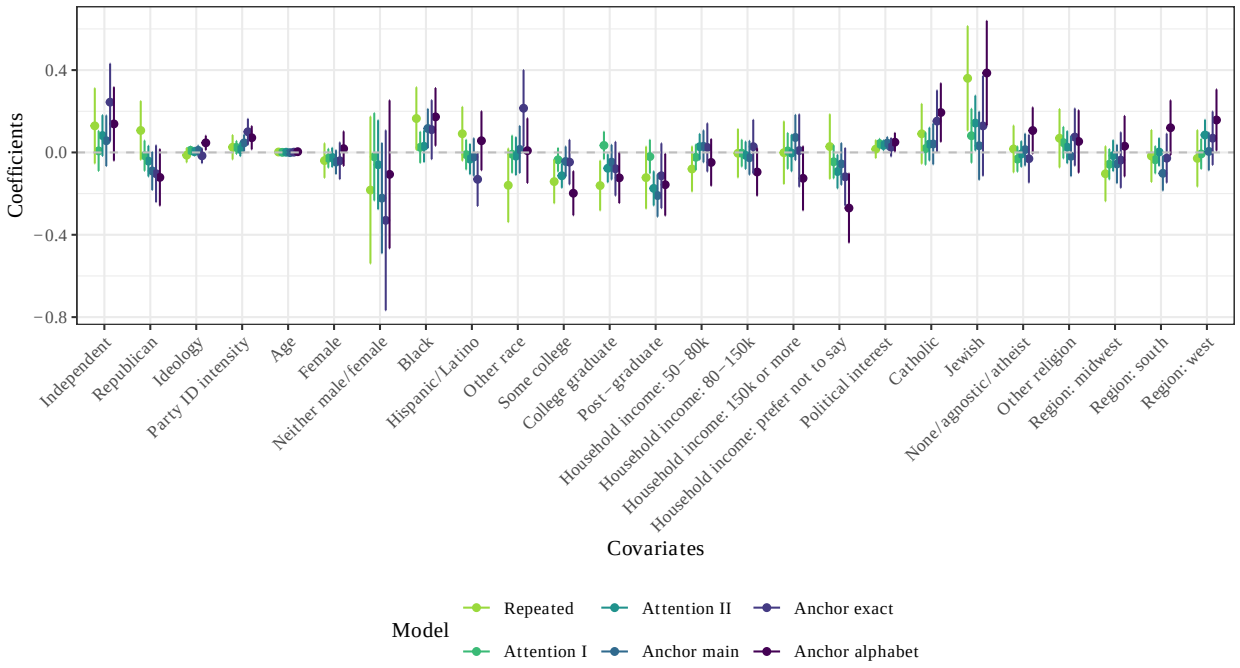


Figure C.10: Coefficient Plot for Table C.1

Some consistent patterns emerge: levels of education seem to be associated with response randomness. Having a post-graduate degrees is statistically significant in models for Attention II and the main anchor question. College graduate status is statistically significant for repeated questions, and similarly, some college degree for Attention II and the alphabetical order anchor question. All these have a negative effect on responding randomly. Additionally, Judaism was related to higher degrees of randomness for repeated questions and the alphabetical order anchor question.

On the other hand, a higher degree of randomness was reported if partisan identification intensity (e.g., higher for strong Democrats/Republicans than for moderate Democrats/Republicans) for all anchor questions, as well as higher political interest for Attention I and the main anchor question. We discuss an alternative assumption and estimation strategy we can adopt in light of these discoveries in Section 6.2.

C.3 Relationships Between Random Responses via Different Types of Checks

In this subsection, we examine the relationship between failing attention checks, failing repeated questions, and answering anchor questions randomly. First, we show the correlation plot between the items at hand. There is no correlation between *Anchor alphabet* and *Anchor exact* because they were exclusive treatments after randomization. Figure C.11 shows the visualization.

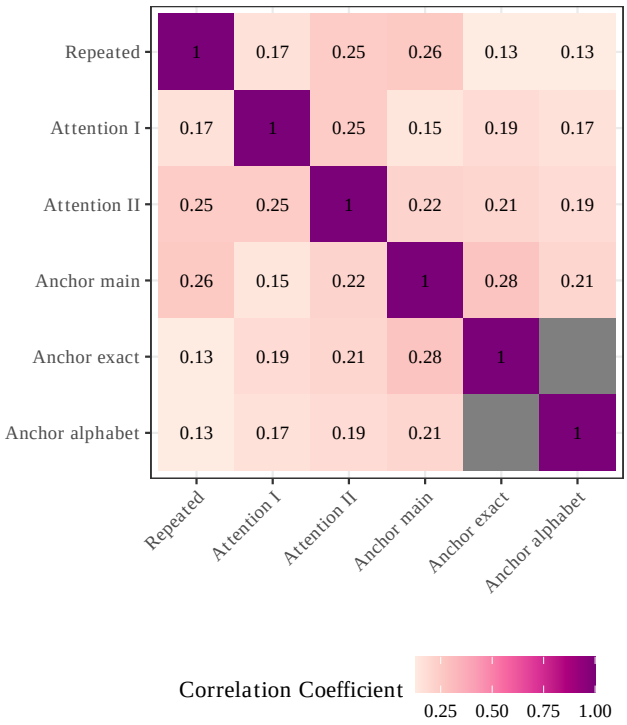


Figure C.11: Pairwise Correlation Between Checks for Random Responses

If people's attention does not fluctuate across questions, we should expect to see high correlations among these checks. In contrast, we find that the correlations between the items are quite low, none of them exceeding 0.3. This is not just a feature of anchor questions. For example, conditional on failing the first attention check (adopted from Ternovski and Orr (2022)), only 36.5% of the respondents fail the second attention check (variation of the check used in Berinsky et al. (2014)). Conversely, conditional on failing the second attention check, only 34.4% of the respondents fail the first attention check.

The highest correlations occur between the main anchor question and the exact order

anchor question, then the main anchor question and the repeated question check. In terms of conditional probabilities, conditional on randomly responding to the main anchor question, 35.4% of respondents randomly responded to the exact anchor question, and 45.7% vice versa. Conditional on randomly responding to the main anchor question, 27.4% of the respondents fail to provide a consistent answer when asked the same identity ranking question, and 47.6% vice versa. For a full table of conditional probabilities, see Table C.2.

Conditional on failing	Repeated	Attention I	Attention II	Anchor main	Anchor exact	Anchor alphabet
Repeated	–	21.2%	27.0%	47.6%	27.0%	31.7%
Attention I	27.0%	–	36.5%	47.3%	35.1%	39.2%
Attention II	32.5%	34.4%	–	54.8%	36.3%	39.5%
Anchor main	27.4%	21.3%	26.2%	–	35.4%	34.5%
Anchor exact	20.1%	20.5%	22.4%	45.7%	–	–
Anchor alphabet	19.9%	19.3%	20.6%	37.5%	–	–

Table C.2: Probabilities of Providing a Random Response (Failing the Check) Conditional on Failing Another

What does this tell us? First, this shows again that attention may fluctuate wildly and unstably within respondents even within a short survey (Berinsky et al., 2014). In Berinsky et al. (2014), the correlation between passage rates of different screeners were also low. Second, this may show that the main anchor question designed (ranking community size) is a more conservative filter for random responses than most other options. For another example, the estimation for the proportion of random responses is 31.6% for the main anchor question, while it is 48.6% for the exact order anchor question (and 58.5% for the alphabet anchor question).

At the very minimum, we believe that this indicates that (1) any checks to capture the proportion of random responses should be closely situated to the main ranking question, and (2) non-ranking questions may not sufficiently capture the degree of randomness that can emerge from ranking questions. For example, the second attention check, while sufficiently close to the main ranking question of interest, still displays low degrees of correlation with the main anchor question. While it is outside the scope of this study to analyze exactly *why* respondents display such instability in attention checks/screeners, this section may help inform future researchers who wish to choose between these six alternative designs—again, which we leave for their discretion.

C.4 Latency Measures

Next, we discuss response latency, or response speed, or completion time (Mulligan et al., 2003; Meade and Craig, 2012; Wood et al., 2017; Revilla and Ochoa, 2015), which may be an indicator of response quality. Although there are no accepted common rules of thumb, a response time that is too short (or sometimes even too long) may be interpreted as inattention. Given this, are respondents who failed the checks performing these checks faster, as we might expect?

Given how generally the latency measures are highly skewed, we use a Wilcoxon rank-sum test to compare the samples of check passers (i.e., providing a non-random response) and failers (providing a random response). All checks' latency measures are in the expected direction except the exact anchor, which shows almost no evidence that those who fail the checks show shorter response times.

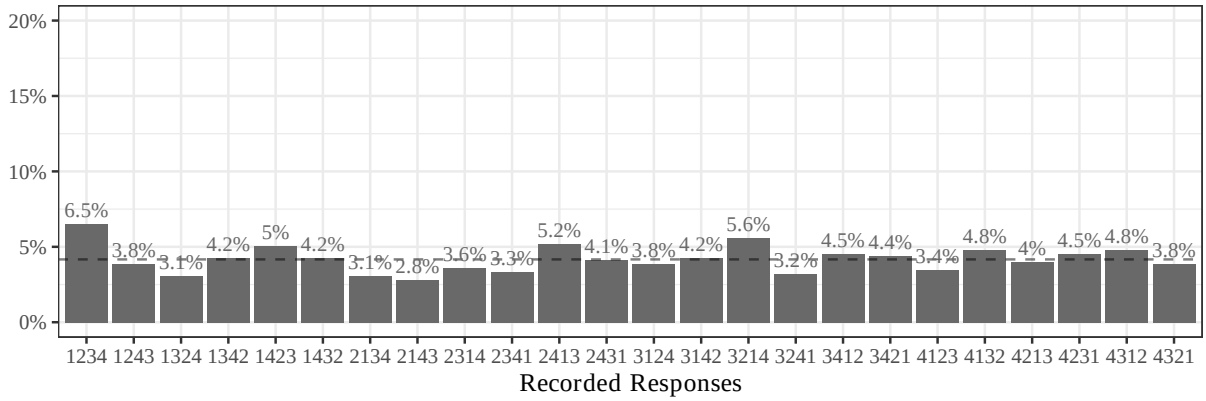
	Passers' Median Latency	Failers' Median Latency	Wilcoxon Rank-sum Test P-value
Repeat	18.5 seconds	16.4 seconds	0.0802
Attention I	15.8 seconds	9.2 seconds	< 0.0001
Attention II	24.1 seconds	13.7 seconds	< 0.0001
Anchor main	27.3 seconds	23.2 seconds	< 0.0001
Anchor exact	32.5 seconds	34.0 seconds	0.3825
Anchor alphabet	24.6 seconds	20.8 seconds	0.0070

Table C.3: Comparison of Response Latency of the Checks for Respondents Who Fail vs. Pass the Checks

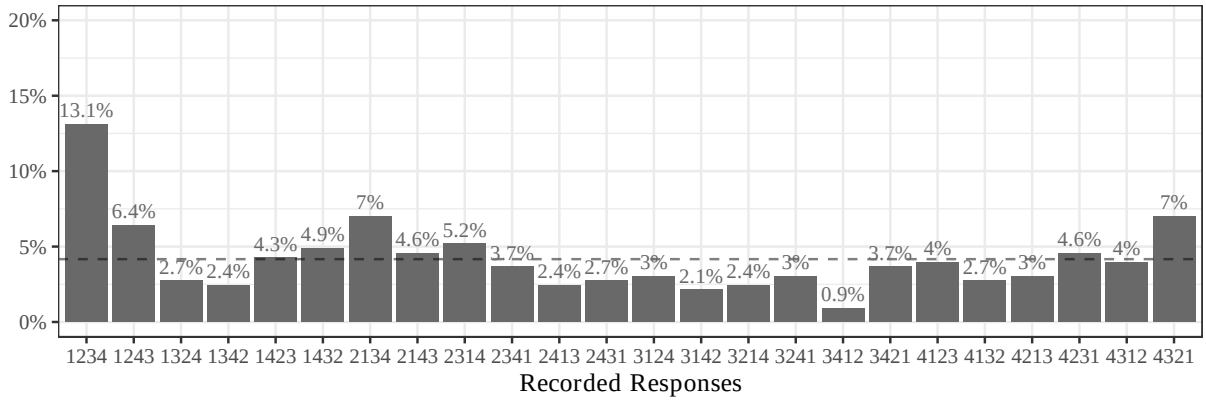
C.5 Uniformity Tests

In this subsection, we apply the uniformity test to some of the alternative questions to examine the variation across them. For comparison, Figure 6 is partly reproduced with a slightly adjusted y-axis in Figure C.12.

Figures C.13–C.13 are Figure C.12 equivalents but for the alphabet anchor, exact anchor, and the repeated question. For the distribution shown in Figure C.13a, the χ^2 test statistic is 17.02 with a p-value = 0.8084, while Figure C.13b has a χ^2 test statistic of 34.60 with a p-value = 0.0569, again largely aligned with our expectation that recorded responses will follow a uniform distribution in the absence of random responses, but only at a significance level of 0.1.

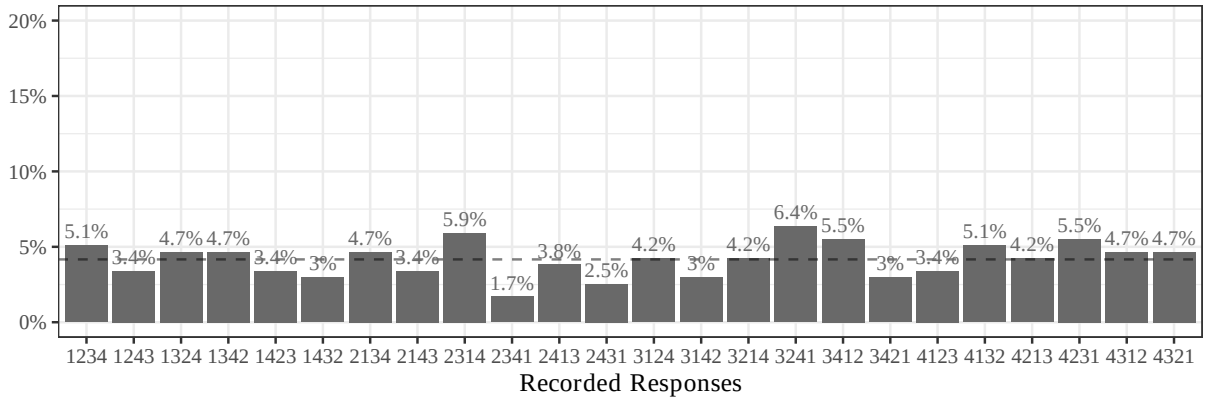


(a) Main Anchor Question, Respondents Who Passed the Anchor

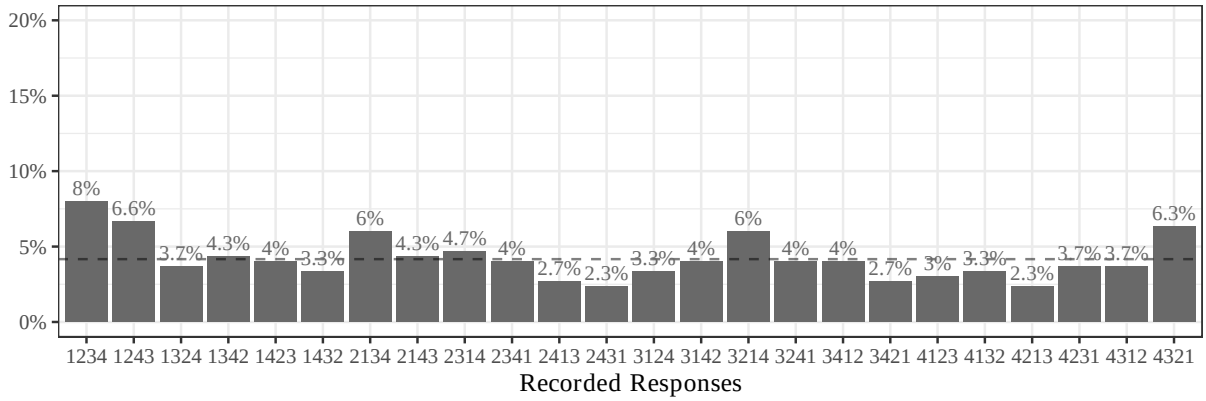


(b) Main Anchor Question, Respondents Who Failed the Anchor

Figure C.12: Uniformity Test Visualized: Distribution Over All Possible Recorded Rankings, Main Anchor Question



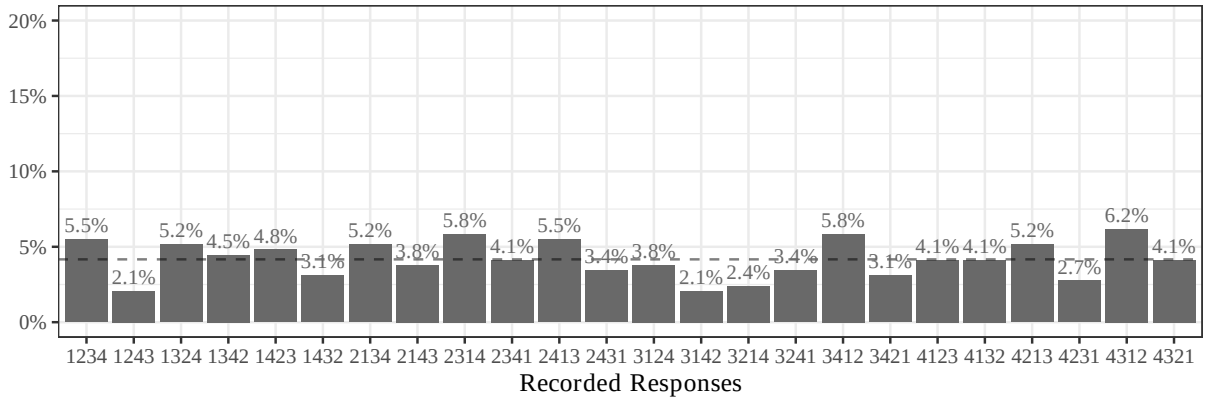
(a) Alphabet Anchor Question, Respondents Who Passed



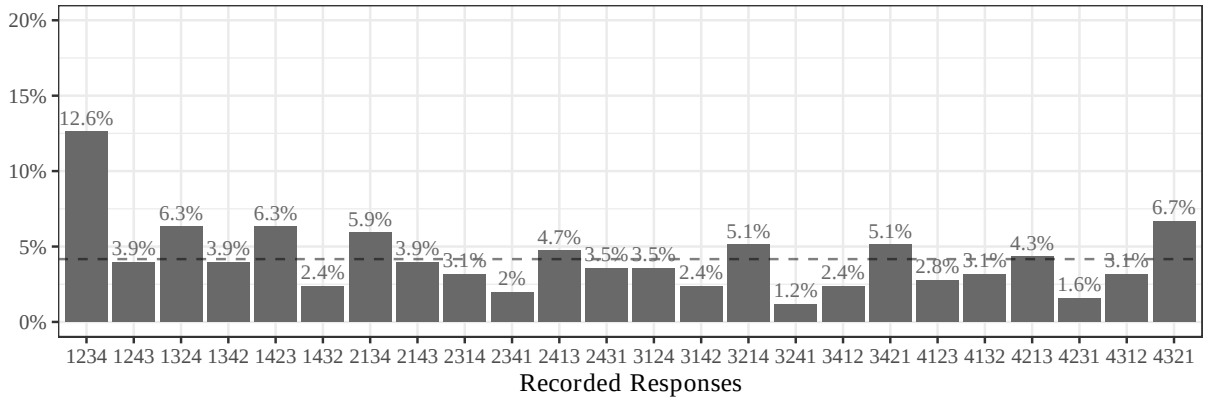
(b) Alphabet Anchor Question, Respondents Who Failed

Figure C.13: Uniformity Test Visualized: Distribution Over All Possible Recorded Rankings, Alphabet Anchor Question

Similarly, using the exact anchor question, Figure C.14a has a χ^2 test statistic of 23.97 with a p-value = 0.4055, while Figure C.14b has a χ^2 test statistic of 78.03 with a p-value < 0.001.



(a) Exact Anchor Question, Respondents Who Passed

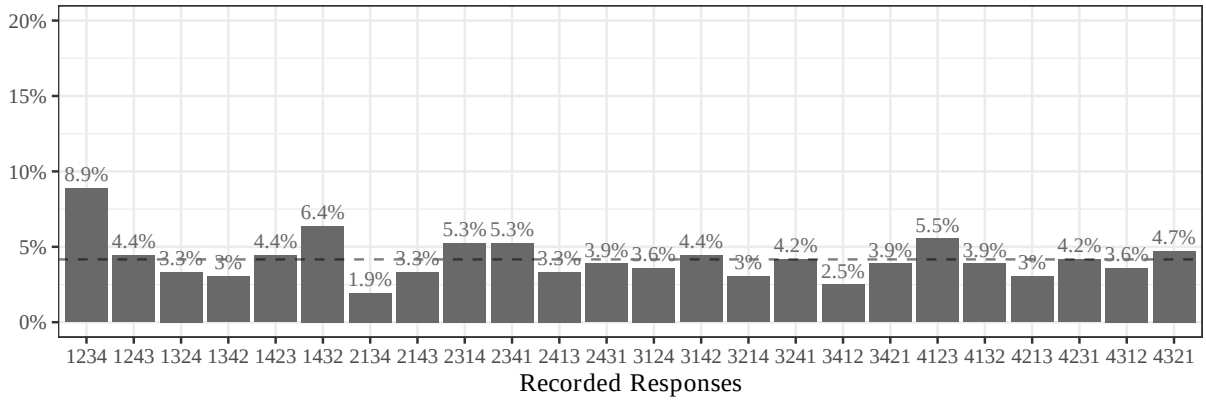


(b) Exact Anchor Question, Respondents Who Failed

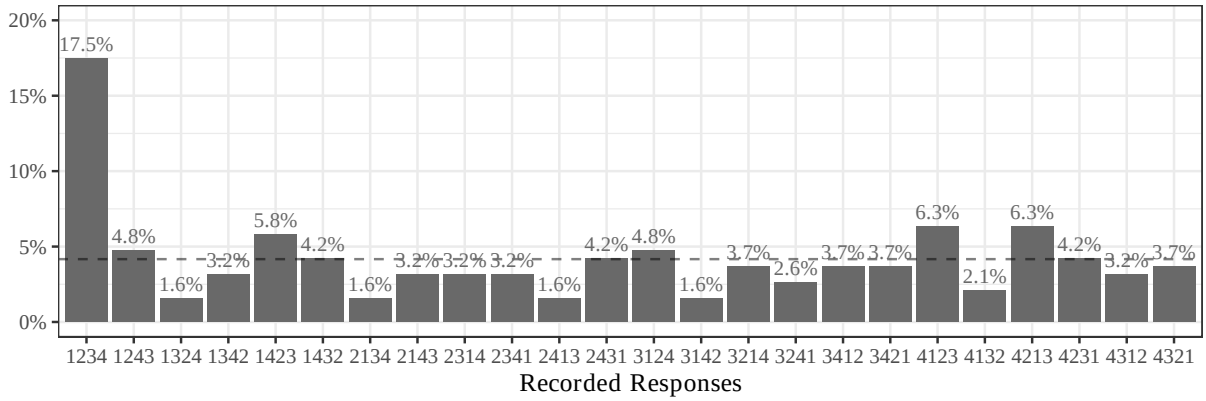
Figure C.14: Uniformity Test Visualized: Distribution Over All Possible Recorded Rankings, Exact Anchor Question

On the other hand, the story is different for repeated questions. For those who have passed (Figure C.15a, the χ^2 test statistic is 40.09 with p-value = 0.01503, while for those who have failed (Figure C.15b), the χ^2 test statistic is 103.70 with p-value < 0.0001. Thus, although the proportion of forward-facing diagonalization (1234) is almost halved among the passers compared to the failers, the recorded responses' distribution does not seem uniform. Although it is not entirely clear why, given these findings, our inclination is to put more trust in well-designed anchor questions among the checks presented.

Finally, note that uniformity tests are impossible for scenarios in which only attention



(a) Repeated Questions, Respondents Who Passed



(b) Repeated Questions, Respondents Who Failed

Figure C.15: Uniformity Test Visualized: Distribution Over All Possible Recorded Rankings, Repeated Questions

checks are employed because they are typically not in rank-order forms (as are the two attention checks that tested here). Note that we are intentionally *not* using the probability mass function over the main ranking question of interest, because we do not assume that randomness carries over at the respondent level—i.e., we do not assume that the distribution over ranking patterns will be uniform for those who pass the attention check and not for those who fail. Assumption 2 merely assumes that the proportion of random responses is the same across the main question and the check.

C.6 Estimated Proportion of Random Responses and Resulting Quantities of Interest

The upper panel of Figure C.16 presents the estimated proportion of random responses based on six alternative designs. Both attention checks estimate that about 14–15% of respondents provided random responses to our target question based on their responses to the checks that were embedded at the beginning of the survey. These are significantly lower than our original estimate (about 32%), potentially because (a) respondents were more attentive at the time of the attention checks, which was earlier in the survey, (b) respondents provided random responses to ranking questions than to the check questions, among others. In contrast, the estimates based on the two additional anchors are much higher (49% and 58%) compared to 32% in the main anchor. This pattern is potentially because many respondents recognized these anchors to be "trick" questions and decided not to respond truthfully. Finally, the repeated question appeared to produce a similar estimate to ours (0.34).

The lower panel of Figure C.16 shows the estimated average ranks based on our correction (three anchors) and listwise deletion (two attention checks and one repeated question). The gray bands represent the 95% CIs of unadjusted (raw-data) estimates. It demonstrates that listwise deletion yields results that are statistically and substantively similar to unadjusted estimates, while our methods correct for the measurement errors. Based on the evidence presented so far, we believe that the main anchor is the most reasonable choice among the other designs (see also Appendix D.2).

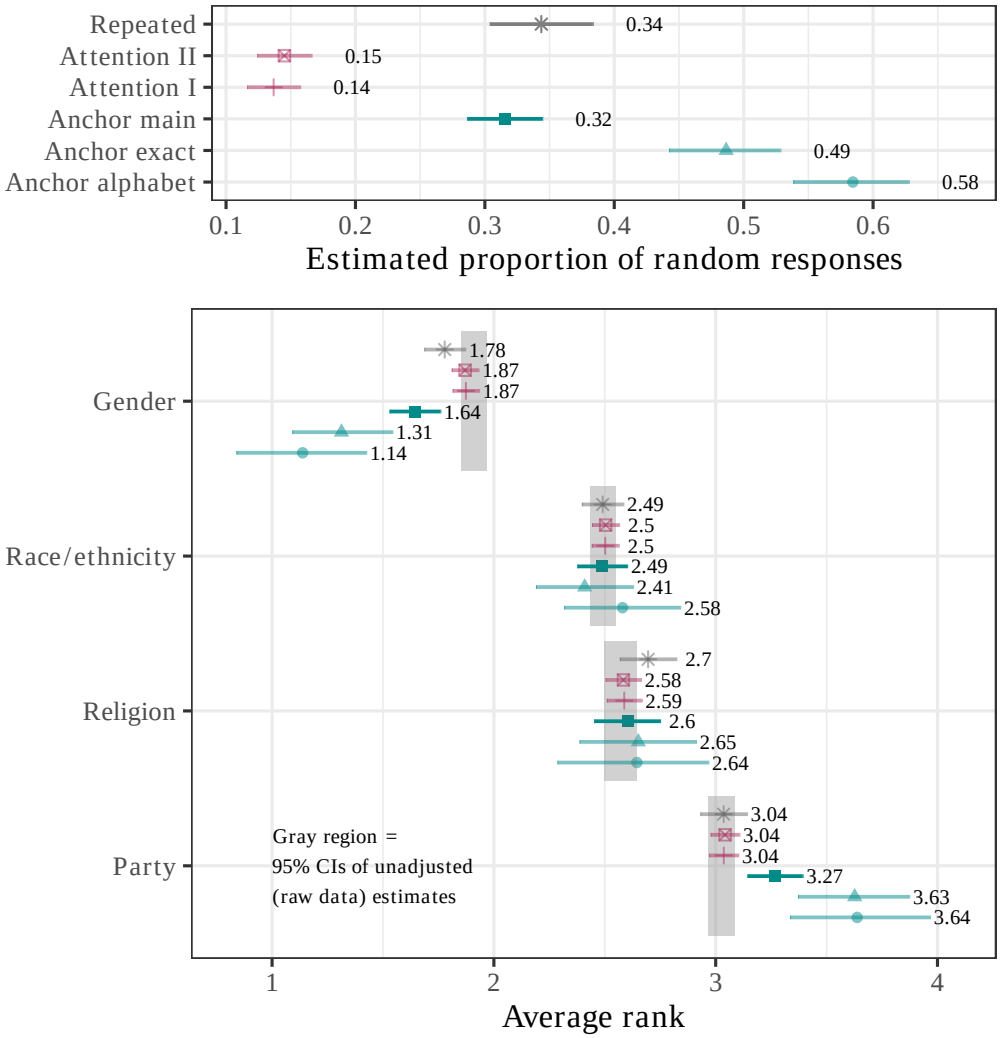


Figure C.16: Comparison of Average Ranks with Alternative Designs

Appendix D A Practical Guide for Designing Anchor Questions

In this section, we provide a practical guide for designing anchor questions. While our framework is statistical, the heart of our methodology calls for researchers' substantive knowledge related to their ranking questions of interest. The essence of our method is to give researchers the power to make their assumptions more plausible by designing (and re-designing during pretests) their anchor questions.

D.1 How Should We Construct Anchor Questions?

Where can researchers find topics for anchor questions for which they know the correct rankings? While there is no "correct" way to construct anchor questions, there are several natural orderings that researchers can take advantage of. We have already demonstrated three types of different anchor questions, using spatial ordering, alphabetical ordering, and exact ordering. Other forms of orderings can also be exploited, such as temporal and logical orderings, as long as they are related to the topic of the target ranking questions.

How can we design our anchor question so that we can *make* Assumption 1 more plausible? Our generic suggestion for researchers is to make their anchor questions as close as possible to their target ranking questions. For example, we recommend that researchers choose a choice set that is relevant and connected to the target ranking questions. It is also important for researchers to craft their anchor questions by using similar lengths of instructions, item descriptions, and tones of questions. Furthermore, if researchers believe that random responses are related to respondents' political knowledge, they may want to craft their anchor questions that require respondents to have the same level of political knowledge necessary to answer the main ranking questions correctly.

D.2 Easier vs. Harder Anchor Questions

Here, we highlight one important practical concern: what if an anchor question is easier than the target question? Researchers may wonder if their anchor questions may be "too easy" compared to their target ranking question, thereby underestimating the degree of random responses. While this is a reasonable concern, we discuss that such a situation

allows us to draw a more conscious conclusion by under-correcting for measurement errors.

Here, we assume that researchers are interested in finding sharper (i.e., non-uniform) ranked preferences in many applications. The main statistical concern with this setup is that we may falsely find "more interesting" or less uniform results—false positives. Now, suppose that when an anchor question is actually "easier" than the target question, fewer respondents offer random responses to the anchor question. Consequently, the proportion of random responses in the target question will be estimated to be lower than it is.

Under this scenario, the proposed estimator will give researchers a biased estimate of their quantity of interest. However, it does so by *under*-correcting measurement errors. Substantively, this means that easier anchors offer point estimates closer to the uniform than otherwise. In other words, by underestimating the magnitude of measurement errors, easier anchors will give us more conservative estimates about or lower bounds of our quantities of interest. For this reason, we recommend that analysts consider "easier" anchor questions than "harder" ones compared to their target ranking questions unless there is good reason per domain expertise.

Having said that, we also find evidence for the exact opposite effect of easier questions. We find that some anchor questions that might seem easier to the researcher may actually have more random responses. Figure C.16 shows that the proportion of getting the main anchor question correct was 69.7%, while for the alphabet and exact anchors, both of which are seemingly easier than the identity question, it was only 43.9% and 53.4%, respectively. One potential explanation is that some respondents spotted the latter two anchor questions as "trick" questions and decide to give random responses. Thus, while easier anchors can be helpful if they appear to be substantively relevant ranking questions, they may have the opposite effect when they seem to be manipulation checks.

D.3 Pilots and Pretests

We highly recommend that the researchers try out a pilot test of the anchor questions before fielding it to the full sample. As we have demonstrated above in Appendix C and Appendix D.2, even with the best design intentions, respondents may behave in

ways that are inconsistent with the researcher's ex-ante beliefs—such as having a lower proportion of correct answers for the alphabet anchor. There also may be more effective alternatives depending on the situation. For example, see the postmaterialism value question in the World Values Survey (2017–2021 World Values Survey Wave 7, Master Survey Questionnaire, pages 11–12), where respondents rank a different list of four items out of twelve items for three different questions. In this case, respondents answer the question three times with parallel questions, providing a built-in method to boost the signal relative to random noise. Breaking rankings with multiple items into subtasks with a smaller number of items to rank may reduce the cognitive load and, potentially, random responses.

In pretesting the survey, one useful instrument may be to follow up on the respondents who have failed the anchor question to ask why they have given the answers they have. For example, if there is an unintuitive answer to the main anchor question, we may write the survey flow to follow up with a question such as "You have ranked your city as larger than your state. Can you tell us why?" Such answers may yield insights as to whether respondents think certain types of answers are also "correct" answers and may help calibrate the proportion of random responses.

Appendix E Existing Scholarship on Ranking Questions

This section briefly reviews scholarly discussions around ranking questions to better situate our work.

Ranking questions often are used to elicit preferences from survey respondents to learn about individual preferences (and sometimes aggregate them into group-level choices) in various platforms (Yu et al., 2019; Marden, 1996; Alvo and Yu, 2014). For example, political scientists seek to understand who citizens blame when the government fails to fulfill people's needs (Malhotra and Kuo, 2008). Election reformers consider ranked-choice voting as an alternative to the simple plurality rule (Drutman, 2020) and test its applications via ranking survey questions (Crowder-Meyer et al., 2021). Marketers want their customers to provide rankings over flavors of yogurt for product development (Bolton and Brace, 2008). Sociologists study what parents value as qualities in their children (Kohn, 1977). Data scientists build effective recommendation systems for movies, songs, and food menus based on people's rankings (Xia, 2019).

In modern surveys, ranking questions can be asked in several different formats, including, but not limited to, *drag-and-drop*, *radio buttons*, *numeric entry*,[2] and *select box*. For a comparison of different forms of ranking, see Fabbris (2013). Our framework addresses measurement errors regardless of the format, although what types of random responses are incurred may be different across designs. Genter et al. (2022) assess the drag-and-drop design versus the numeric entry design and find no significant differences in terms of task time or distribution of rankings. Smyth et al. (2018) also explore the level of usable data provided in two different ranking question formats and concludes that the numbering format provides more useful data than the most-second-most grid format.

Regardless of format, ranking questions are a complex version of more typical discrete choice questions, where respondents are asked to make *multiple choices* on *multiple items*. In other discrete choice questions, respondents are only asked to select a single item between two (e.g., "money" or "freedom"), choose a single item among several (e.g., "Trump," "Biden," "Someone else"), or provide an ordered-response to all available items. Given the relative complexity of ranking questions and expected cognitive difficulty, scholars have debated whether ranking questions can offer informative data

[2]Qualtrics uses the name "text box" for numeric entry.

for social science research (Alwin and Krosnick, 1985; Dillman et al., 2014; Smyth et al., 2018; Kaufman et al., 2021; Plutzer and Berkman, 2022).

On the one hand, several studies have cast doubt on survey respondents' ability to answer ranking questions properly and, thus, on the quality of observed ranking data. Krosnick and Alwin (1987), for example, find that items presented early were disproportionately likely to be picked. Moreover, Serenko and Bontis (2013) also find that when respondents are asked to rank journals on a seven-point Likert scale, peoples' responses are affected by journal order. Recently, Atsusaka (2024) finds that some respondents use random geometric patterns instead of their underlying preference when answering ranking questions in the context of ranked-choice voting elections. Because of these potential problems, some studies suggest that ranking questions must be replaced with rating or pairwise choice questions (McCarty and Shrum, 2000). While the literature has found no significant difference between ranking and rating (Ovadia, 2004), researchers have overwhelmingly preferred rating over ranking. McCarty and Shrum (2000, 274) suggest that "[t]his preference for rating likely stems from the advantages it affords for statistical analyses, in spite of the potential disadvantages with respect to its measurement properties." They also note that "the ordinal nature of ranked data limits analysis to the use of nonparametric statistical procedures" (McCarty and Shrum, 2000, 273). This may be slightly misleading as parametric models for ranking data have a long history in statistics (Fligner and Verducci, 1993; Liu et al., 2019). Thus, the relative unpopularity of ranking over rating may be based on the challenge that many applied researchers are unfamiliar with ranking data analysis (c.f., for rating data, OLS can be used, albeit without accounting for the dependency among items) and *not* based on the properties of ranking survey questions.

On the other hand, other studies support the utility of ranking questions. For example, Krosnick (1999) provides a meta-analysis showing that ranking questions provide better data than rating questions in terms of nonresponsiveness, reliability, and discriminant validity. In particular, rankings can reduce the data collection workload while ensuring that the respondents use the same dimension of attributes for comparison. Additionally, Kaufman et al. (2021, p. 539) show that when asking survey respondents to compare differently shaped legislative districts for a measure of compactness, pairwise comparison "utterly fails" while the ranking approach works well. Similarly, Yannakakis

and Martínez (2015) claim that "ratings are overrated" and discuss how ranking-based questions can overcome the limitations of rating-based questions.

While both camps carefully assess the value of ranking questions, what has been missing in the debate is *what researchers can do* when their ranking questions induce measurement errors. More specifically, what should analysts do when some respondents answer rankings questions randomly and not based on their underlying preferences? In the absence of systematic analysis, researchers have been left with little guidance about how random responses may affect their analysis, let alone what they can do about it. The statistical framework and the proposed solution in this study will transform the complexity of ranking questions into its strength. We provide means to estimate and correct measurement errors while designing survey questions accordingly. This, in turn, allows researchers to study their quantities of interest without bias using ranking questions.

References

Alvarez, R. Michael, Lonna Rae Atkeson, Ines Levin, and Yimeng Li (2019). Paying Attention to Inattentive Survey Respondents. *Political Analysis 27*(2), 145–162.

Alvarez, R Michael and Yimeng Li (2022). Survey Attention and Self-Reported Political Behavior. *Public Opinion Quarterly 86*(4), 793–811.

Alvo, Mayer and Philip L. H. Yu (2014). *Statistical Methods for Ranking Data*. Springer.

Alwin, Duane F. and Jon A. Krosnick (1985). The Measurement of Values in Surveys: A Comparison of Ratings and Rankings. *Public Opinion Quarterly 49*(4), 535–552.

Atsusaka, Yuki (2024). Analyzing ballot order effects in ranked-choice voting. *Political Analysis*.

Berinsky, Adam J., Alejandro Frydman, Michele F. Margolis, Michael W. Sances, and Diana Camilla Valerio (2024). Measuring Attentiveness in Self-Administered Surveys. *Public Opinion Quarterly*, nfae004.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis 20*(3), 351–368.

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), 739–753.

Bolton, Kate and Ian Brace (2008). *Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research*. Kogan Page.

Clayton, Katherine, Yusaku Horiuch, Aaron R. Kaufman, Gary King, and Mayya Komisarchik (2023). Correcting Measurement Error Bias in Conjoint Survey Experiments.

Crowder-Meyer, Melody, Shana Kushner Gadarian, and Jessica Trounstine (2021). Ranking Candidates in Local Elections: Neither Panacea nor Catastrophe.

Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley & Sons.

Drutman, Lee (2020). *Breaking the Two-Party Doom Loop: The Case for Multiparty Democracy in America*. Oxford University Press.

Fabbris, Luigi (2013). Measurement Scales for Scoring or Ranking Sets of Interrelated Items. In C. Davino and L. Fabbris (Eds.), *Survey Data Collection and Integration*, pp. 21–43. Berlin, Heidelberg: Springer.

Fligner, Michael A and Joseph S Verducci (1993). *Probability models and statistical analyses for ranking data*, Volume 80. Springer.

Genter, Shaun, Yazmín García Trejo, and Elizabeth Nichols (2022). Drag-and-Drop Versus Numeric Entry Options: A Comparison of Survey Ranking Questions in Qualtrics. *Journal of User Experience 17*(3), 117–130.

Hauser, David J. and Norbert Schwarz (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods 48*(1), 400–407.

Kaufman, Aaron R., Gary King, and Mayya Komisarchik (2021). How to Measure Legislative District Compactness If You Only Know It When You See It. *American Journal of Political Science 65*(3), 533–550.

Kohn, Melvin (1977). *Class and Conformity: A Study in Values*. University of Chicago Press.

Krosnick, Jon A. (1999). Survey Research. *Annual Review of Psychology 50*(1), 537–567.

Krosnick, Jon A. and Duane F. Alwin (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly 51*(2), 201–219.

Leiner, Dominik Johannes (2019). Too Fast, too Straight, too Weird: Non-Reactive Indicators for Meaningless Data in Internet Surveys. *Survey Research Methods 13*(3), 229–248.

Liu, Qinghua, Marta Crispino, Ida Scheel, Valeria Vitelli, and Arnoldo Frigessi (2019). Model-based learning from preference data. *Annual review of statistics and its application 6*, 329–354.

Malhotra, Neil and Alexander G. Kuo (2008). Attributing Blame: The Public's Response to Hurricane Katrina. *The Journal of Politics 70*(1), 120–135.

Marden, John I. (1996). *Analyzing and Modeling Rank Data*. CRC Press.

McCarty, John A. and L. J. Shrum (2000). The Measurement of Personal Values in Survey Research: A Test of Alternative Rating Procedures*. *Public Opinion Quarterly 64*(3), 271–298.

Meade, Adam W. and S. Bartholomew Craig (2012). Identifying careless responses in survey data. *Psychological Methods 17*(3), 437–455.

Mulligan, Kenneth, J. Tobin Grant, Stephen T. Mockabee, and Joseph Quin Monson (2003). Response Latency Methodology for Survey Research: Measurement and Modeling Strategies. *Political Analysis 11*(3), 289–301.

Ovadia, Seth (2004). Ratings and rankings: reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology 7*(5), 403–414.

Plutzer, Eric and Michael B. Berkman (2022). Scaled Paired Comparisons as an Alternative to Ratings and Rankings for Measuring Values.

Reuning, Kevin and Eric Plutzer (2020). Valid vs. Invalid Straightlining: The Complex Relationship Between Straightlining and Data Quality. *Survey Research Methods 14*(5), 439–459.

Revilla, Melanie and Carlos Ochoa (2015). What are the Links in a Web Survey Among Response Time, Quality, and Auto-Evaluation of the Efforts Done? *Social Science Computer Review 33*(1), 97–114.

Serenko, Alexander and Nick Bontis (2013). First in, best dressed: The presence of order-effect bias in journal ranking surveys. *Journal of Informetrics 7*(1), 138–144.

Smyth, Jolene D., Kristen Olson, and Allison Burke (2018). Comparing survey ranking question formats in mail surveys. *International Journal of Market Research 60*(5), 502–516.

Ternovski, John and Lilla Orr (2022). A Note on Increases in Inattentive Online Survey-Takers Since 2020. *Journal of Quantitative Description: Digital Media 2*.

Tyler, Matthew, Justin Grimmer, and Sean J. Westwood (2024). A Statistical Framework to Engage the Problem of Disengaged Survey Respondents: Measuring Public Supposrt for Partisan Violence.

Wood, Dustin, P. D. Harms, Graham H. Lowman, and Justin A. DeSimone (2017). Response Speed and Response Consistency as Mutually Validating Indicators of Data Quality in Online Samples. *Social Psychological and Personality Science 8*(4), 454–464.

Xia, Lirong (2019). *Learning and Decision-Making from Rank Data*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Cham: Springer International Publishing.

Yannakakis, Georgios N. and Héctor P. Martínez (2015). Ratings are Overrated! *Frontiers in ICT 2*.

Yu, Philip L. H., Jiaqi Gu, and Hang Xu (2019). Analysis of ranking data. *WIREs Computational Statistics 11*(6), e1483.