Ranked-choice conjoint experiments*

Mats Ahrenshop† Spyros Kosmidis‡ Thomas S. Robinson§

Early draft – presented at EPSA 2024

Abstract

Forced-choice conjoint designs have become a staple method in the experimentalist's toolkit. Conjoint experiments are useful for estimating multiple marginal effects simultaneously, but the forced-choice outcome is neither always consistent with the types of choices individuals make in real political contexts, nor is it statistically efficient. In this paper, we show how political scientists can, instead, ask subjects to *rank* conjoint profiles. We demonstrate that this ranked-choice conjoint is statistically more efficient and, with minimal new assumptions, does not affect the broader validity of the conjoint design. We validate the utility of this design in a survey experiment, randomising subjects into forced-choice and ranked-choice versions of two conjoint experiments: one choosing between candidates, and another choosing between budget proposals. Alongside this paper, we provide `R` implementations of the methods needed to handle ranked-choice conjoints in the **cjbart** package [NB: coming soon!].

*We would like to thank Ben Lauderdale and Andy Eggers for helpful comments on this project. This project underwent Ethics Review at the London School of Economics (approved May 2023, ref. 192578).

†Department of Politics and International Relations, University of Oxford.
(email: `mats.ahrenshop@politics.ox.ac.uk`)

‡Department of Politics and International Relations, University of Oxford.
(email: `spyros.kosmidis@politics.ox.ac.uk`)

§Department of Methodology, London School of Economics.
(email: `t.robinson7@lse.ac.uk`)

1 Introduction

Since their introduction to political science by Hainmueller et al. (2014), conjoint experiments have become a staple of political scientists' research toolkit. In these designs, subjects are faced with a choice between (typically two) profiles. The "conjoint" aspect of the design stems from the fact that each profile is described by multiple attributes, the values of which are randomised. Within political science, conjoint experiments have helped understand preferences over tax policy (Ballard-Rosa et al. 2017), voters expectations over corruption (Spencer & Theodoridis 2020), and emigration decisions (Duch et al. 2022), to name a few.

The core benefits of the conjoint design stem from two primary features. First, conjoint experiments allow researchers to estimate treatment effects less intrusively than direct preference measures. By averaging across choices, profiles, and the values of other attributes, researchers can estimate the average marginal component effect (AMCE) of attribute-levels (relative to some set of baselines). Second, and unlike simpler survey experiments, the conjoint design allows multiple dimensions to be randomized at once–allowing researchers both to test multiple hypotheses more efficiently, and making the treatment better resemble the information contexts individuals likely face in real political situations (Hainmueller et al. 2014).

Notwithstanding these advantages, conjoint experiments are not a silver bullet. Beyond questions over which conjoint *estimands* are informative (Abramson et al. 2019), and difficulties in interpreting coefficients *across* models (Leeper et al. 2020), conjoint experiments typically require subjects to complete many, repetitive rounds in order to estimate precisely the marginal effects of interest (Bansak et al. 2018). In doing so, and even if survey satisficing remains low, conjoint experiments can be costly to run due to the number of rounds subjects have to consider.

In this paper, we introduce an alternative to the typical conjoint design, replacing the forced-choice component with a ranking exercise. While we are not the first to propose the use of rankings in conjoint experiments (Vermeulen et al. 2011), to the best of our knowledge, this design has not received any attention by political scientists (nor social scientists more generally), despite its appealing properties and compatibility with many types of political decisions individuals make.

From a statistical perspective, we show that using a rankings-based outcome allows researchers to recover more information more quickly. In other words, for a fixed number of randomised profiles,

the ranked-choice design requires fewer rounds to be shown to subjects and, for each of those rounds, we recover more comparisons. Thus, and as we demonstrate in this paper, ranking profiles can lead to more efficient decision-making by subjects. Moreover, we show that the resulting estimates, for a fixed number of profiles, will be more precise, if we assume that that subjects' rankings of profiles are transitive.

To assess the performance of ranked-choice conjoints, we conduct a bespoke survey experiment in which we randomise subjects into taking either forced-choice or ranked-choice conjoint experiments. We replicate two common political science contexts in which conjoints have been used previously: budget spending (similar, but not identical, to Bansak et al. 2021) and a candidate election vignette Kirkland & Coppock (2018). We find that, consistent with our expectations, ranked-choice conjoints yield more precise, but reassuringly similar, estimates of the average marginal component effect (AMCE). Using an out-of-sample data, we also find some evidence that, in electoral vignettes, the ranked-choice conjoint data enables us to learn a better response surface of subjects' behaviours.

The remainder of this paper is organised as follows. We first set out how a rankings-based conjoint differs from forced-choice conjoints, and how we recover the conventional estimates from data with ranked outcomes. We then describe our pre-registered hypotheses, and the design of our survey experiment, as well as the specific conjoint experiments we replicate. We then present the results from this experiment, fielded in June 2024 on a representative sample of voting-age US subjects, before discussing their implications.

2 Rankings as an alternative to forced choices

While conjoint experiments have been used to study preferences since the 1970s, the method was popularised in political science by Hainmueller et al. (2014). Subjects typically face a choice between two "profiles", which are bundles of randomised treatments along a number of dimensions, typically referred to as "attributes". Each subject completes multiple of these tasks, where each "round" re-randomises the descriptions of each profile. For the sake of consistency, let a typical conjoint design be posed to i individuals, who complete j rounds, each of which has k profiles (typically two), and each profile has l attributes. The number of alternative attribute-levels within each attribute is denoted as v_l.

The outcome in conjoint experiments is measured at the subject-round-profile level. In the typical forced-choice experiment, this measure is a binary indicator for whether a profile was chosen (i.e. $y_{ijk} \in \{0, 1\}$). In the vignette, subjects are prompted with a question of the form "If you had to choose between these two profiles,..." Beyond the ability to test multiple preference dimensions simultaneously, therefore, one major advantage of conjoint experiments is that for a single vignette the researcher recovers k observations, in comparison to simpler randomised designs where one typically gets a single observation per vignette.

Nothing theoretical prevents k being equal to 1 (e.g Duch et al. Working paper) or larger than 2. In these cases, under the forced-choice design, $y_{ijk'} = 1$, for the chosen profile, and the other $k - 1$ profiles will all be recorded with outcome 0. Hence, as the number of profiles increases per vignette, so too does the number of observations. While this change would seem to imply an increase in power (holding constant the number of vignettes), the amount of additional *information* is less because, for any two profiles not chosen by the respondent, the outcome is 0 for both and thus non-discriminatory.

2.1 Ranked outcomes

An intuitive alternative is to rank $k > 2$ profiles. Subjects assign 1 to their most preferred profile, 2 to their second most preferred, and so on. As a result, rankings are far denser in terms of their information than forced-choice outcomes. Rankings identify not only the top choice (as in the forced-choice measure) but also indicate subjects' relative preferences over all other profiles too (which is lost in the typical conjoint design).

We are not the first to propose the use of rankings in conjoint experiments (Vermeulen et al. 2011), although to our knowledge we are the first to explore this design within political science. This lack of attention in our discipline is surprising, given the ubiquity of ranking in the political contexts we study (Atsusaka 2022). In studies of candidate elections, for example, and outside the minority of states that use majoritarian electoral systems, ranking candidates is a more natural analogue of the task that voters actually perform at the ballot box (Golder 2005, Atsusaka 2023). In policymaking contexts, too, lawmakers or and citizens may be presented with multiple choices, beyond two, and it may be more natural to rank these outcomes than simply choose one over all others.

Beyond its consistency with many real world behaviours, existing evidence suggests ranking alternative can have distinct measurement benefits over forced-choice, pairwise comparisons. For example, Kaufman et al. (2021) find that, when determining the compactness of electoral district boundaries, the intracoder reliability is considerably higher when subjects rank district shapes collectively, rather than make repeated forced-choice designs over pairs of districts. Indeed, in this case, the authors note pairwise comparison "utterly fails to accomplish its goal" (p.539).

Given the inherent complexity of conjoint experiments, which can lead to considerable measurement error in the parameter estimates (Clayton et al. Working Paper), rankings may offer advantages in terms of how voters comprehend the data. As Kaufman et al. (2021) suggest, rankings seem to perform better than pairwise comparisons because ranking tasks better engage subjects with the multi-dimensional nature of the concepts being considered. These purported benefits may also apply in the case of conjoint experiments, and asking subjects to rank multiple profiles may help address some of the measurement issues inherent to forced-choice tasks.

2.2 Recovering observations from ranked-choice conjoints

We denote subject i's ranking in round j of the kth profile as r_{ijk}. As measured, and similar to the forced-choice design, the ranked-choice outcome yields k observations per vignette. With one additional assumption, however, we can "expand" the measured rankings back to binary, pairwise comparisons between each combination of two profiles in that vignette. For example, if a subject ranks profile A higher than B, and B higher than C, then in a forced-choice context between A and B, we would expect the subject to choose A. Similarly, in a choice between A and C, we would expect the subject to choose A. Figure 1 depicts this expansion visually.

To be valid, this expansion relies on one further assumption in addition to those already needed for valid conjoint analysis (see Section X Hainmueller et al. 2014):

Assumption of transitivity *For a set of profiles* $\{P_1, \ldots, P_k\}$ *and subject* i*'s ranking of these profiles,* $\{r_{i1}, \ldots, r_{ik}\}$*, if subject* i *has preferences* $P_x \succ P_y$ *and* $P_y \succ P_z$*, then* $r_{ix} > r_{iy} > r_{iz}$ *for all* x, y, z*.*

In vignettes with few profiles, this assumption is, we argue, justifiable. It is unlikely that subjects would be inconsistent in their rankings of only three choices, for example. As the number of profiles

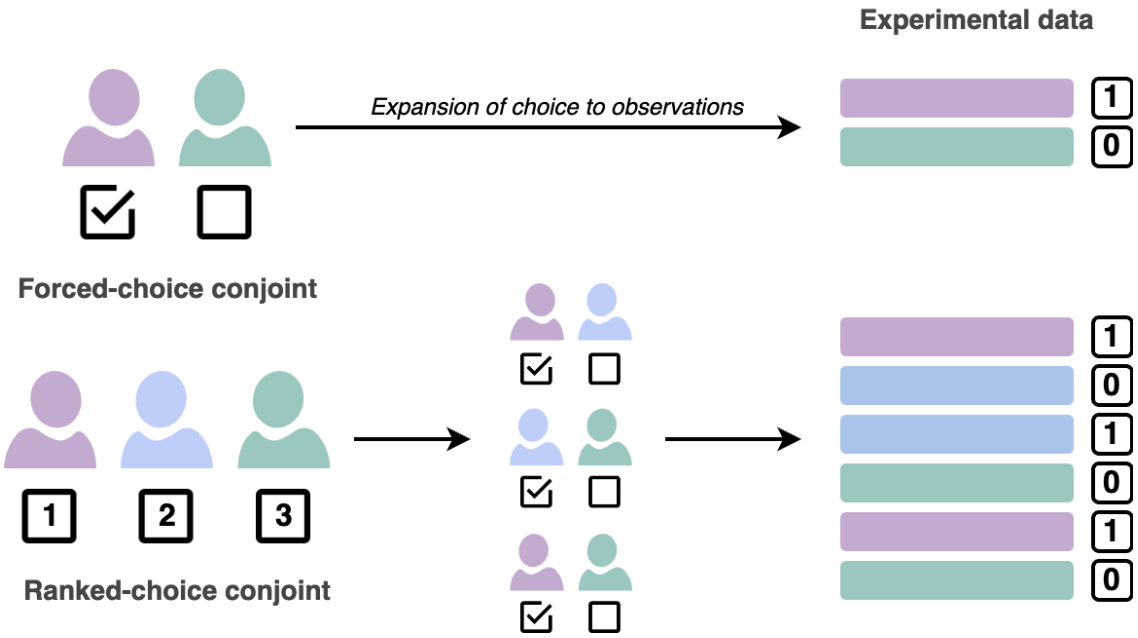

Figure 1: Difference in how forced-choice versus ranked-choice decisions expand to experimental data.

increases, however, we suspect that greater care would need to be taken to check for violations of transitivity. In the face of ordering many different profiles, subjects could plausibly struggle to maintain a consistent ranking across the full choice set.

In these cases, it may be useful to include manipulation checks posing back to subjects forced choice comparisons of a randomly selected ranking from the main conjoint. If the assumption of transitivity holds, then subjects' behaviour in the forced-choice check should be consistent with the expansion of their original rankings.

The statistical advantage of this expansion is the increase in power entailed by the larger number of observations. The forced-choice design scales at a constant rate: adding an extra profile in a vignette increases the number of observations by one. In the ranked-choice design, adding an extra profile increases the number of comparisons by one but increases the number of *observations* non-linearly. the For three profiles we recover 3 comparisons and thus 6 observations; for four profiles we recover 6 comparisons and thus 12 observations; and for five profiles, 20 observations and so forth.[1] In Section 2.4, we demonstrate by simulation the extent to which this increase in observations improves the power of the conjoint model.

[1]The number of comparisons is $2 \times T_{k-1}$, where T_n is the nth triangular number.

2.3 Estimand and estimating procedure

Since the expanded pairwise comparisons maintain the same structure as the forced-choice design, we can employ exactly the same estimating strategy to recover an estimate of the *same* estimand. Thus, aside from adapting the vignettes, the researcher has to make few adaptations to their analysis framework.[2]

That is, we recover estimates of the AMCEs by estimating a linear probability model of the form:

$$y_{ijk} = \sum_{l} \boldsymbol{X}_{l}\beta_{l} + e_{ijk},$$

where $\boldsymbol{X}_{l}$ is the $n \times (v_l - 1)$ indicator matrix of randomised treatments for the lth attribute, and β_{l} is the corresponding vector of AMCEs.

2.4 Simulation evidence of ranking efficiency

To simulate the efficiency advantage of the ranked-choice design, we ran Monte Carlo simulations comparing the statistical power of the designs. These simulations implement a hypothetical, fully-randomised conjoint experiment with six attributes (A-F), each of which has two levels. A non-heterogeneous utility function is defined as a linear combination of attribute-level specific coefficients:

$$U_{ijk} = \gamma \boldsymbol{X}_{ijk} + \varepsilon_{ijk},$$

where these coefficients (γ) are increasingly large to proxy for the varying importance of attributes in subject decision-making. $\boldsymbol{X}_{ijk}$ is an indicator matrix of attribute-level assignments for profile j in round k shown to subject i, and $\varepsilon_{ijk} \sim \mathcal{N}(0, 1)$.[3]

We generate hypothetical conjoint data for 500 subjects exposed to three rounds. For the RCC design, each round contains three profiles and, for the FCC design, each profile contains 2 profiles. We assign profile ranks within rounds for each subject based on the calculated utilities. To recover a binary choice outcome, we "expand" the data to assess the implied choice a subject

[2] Modelling the ranked choice outcomes directly is possible, building on recent advances in incorporate ranking outcomes into the potential outcomes framework (Atsusaka 2022). That said, we do not focus on this approach in this paper because, in doing so, we would necessary shift the estimand away from probabilistic statements about "choosing" one profile over another. From a modelling perspective, too, this shift in attention would be non-trivial: it entails either a linear regression model that assumes cardinality to the rankings, or the use of generalised linear models liked the ordered-logit, which yields more convoluted estimated parameters.

[3] $\gamma = [0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.15, 0.2, 0.25, 0.5, 1, 2]$

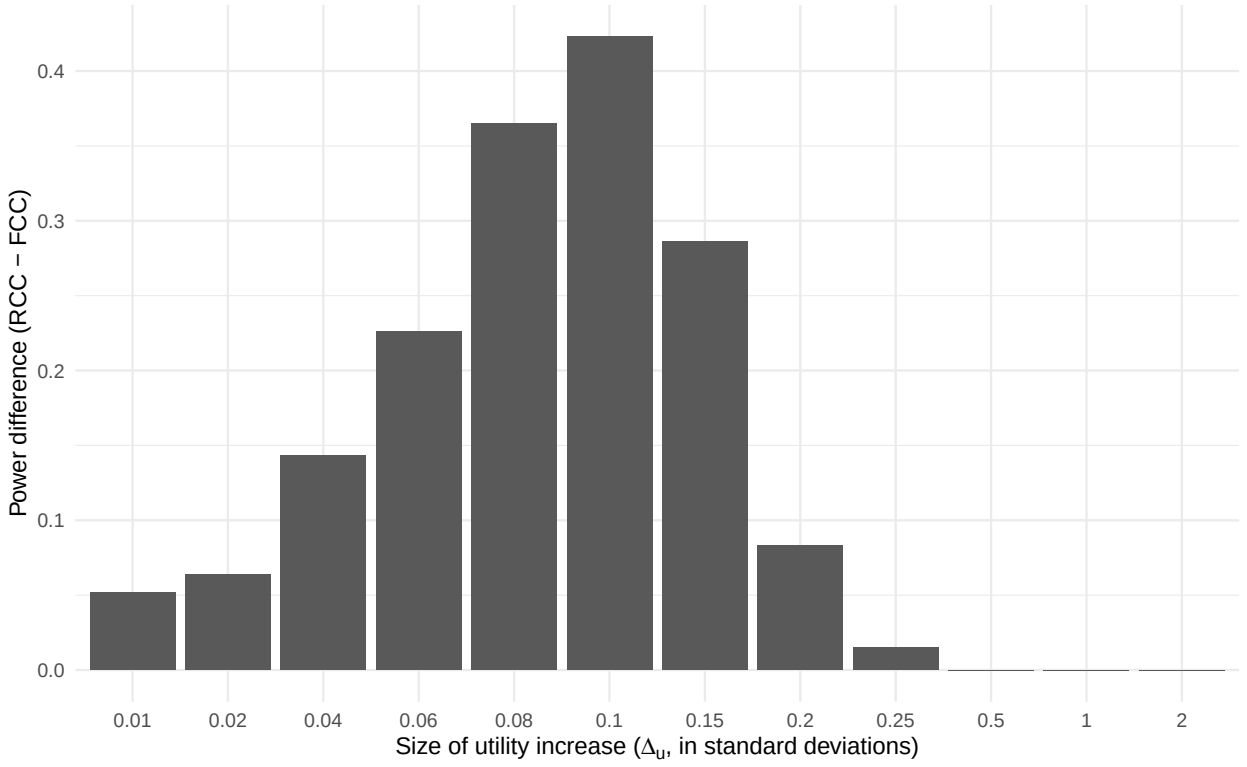

Figure 2: Estimated difference in power between RCC and FCC designs, holding constant the number of subjects and rounds

would make between each combination of two profiles within each round using the rankings. For the FCC condition, we generate data for the same number of subjects and rounds, but showing only 2 profiles (and thus the binary outcome is simply whichever profile has the higher utility).

In both conditions, we estimate linear probability models to estimate of the average marginal component effect (AMCE) of each non-reference attribute level, and repeat this exercise 1000 times. Figure 2 confirms that, as a consequence of lower variance estimates, RCC experiments have higher power. For very small/large utility changes, the difference in power is not substantial, but for moderate size increases–around 0.06 to 0.2 standard deviations of the residual variance–the differences are large. These moderate *utility* increases correspond to conventional AMCE magnitudes: a utility increase of 0.1 standard deviations (in our setup) corresponds to an AMCE of about 0.02 – a two percent increase in the probability of choosing a profile.[4]

[4]In Appendix Figure A1, we show that the density of AMCE estimates for both RCC and FCC designs are centred on the same value, although the RCC has a tighter distribution.

3 Empirical comparison of forced versus ranked choice designs

In this section, we introduce our empirical design to assess the performance of the ranked-choice conjoint design. We first summarise our (pre-registered) hypotheses, that stem from the discussion in the preceding section. We then describe the survey experiment itself, including randomisation procedure and sample.

3.1 Hypotheses

Survey duration In our pre-registration, we hypothesised that ranking profiles may reduce the time needed for subjects to complete experiments. Holding constant the number of profiles shown to subjects, ranked conjoint designs involve fewer vignettes. If there are $j = 2$ ranking tasks involving $k = 3$ profiles to rank, subjects evaluate 6 profiles in 2 vignettes. However, in a forced-choice design with $k = 2$ profiles per task, respondents would need to perform 3 choice tasks to evaluate the same amount of profiles and provide the same amount of preference information. At larger values of k, the comparative number of vignettes increases non-linearly. We hypothesized it would take respondents less time to rank the profiles compared to multiple single choice tasks, holding the total number of evaluated profiles constant across the two designs.

Sample attrition and satisficing In theory, ranking tasks require higher engagement and so may result in less survey satisficing. This may manifest itself in several ways. First, if subjects are more engaged, they may be more likely to finish the experiment. Secondly, higher engagement may mean that the experimental manipulations are better understood and processed by subjects. Thus, in experimental manipulation checks, candidates in the ranked-choice condition may perform better.

Efficiency With equal numbers of respondents and choice tasks, more profiles are evaluated in any single conjoint study under a ranked-choice design. As shown in our simulations, this yields tighter distributions of estimates and higher statistical power. We therefore expect in empirical settings, too, that estimation precision is greater under ranked designs, all else equal.

Recovering Preferences Finally, if ranking tasks prove more engaging to subjects, then not only may the experimental manipulations be more effective, but we may also recover a better estimate of subjects' latent preferences. Therefore, we hypothesize that the AMCEs estimated from ranked-choice designs better reflect the preferences of subjects than AMCEs recovered from forced-choice designs.[5]

To summarise, our pre-registered hypotheses are:

H1 *Subjects complete surveys more quickly under ranked-choice than under forced-choice designs.*

H2 *Attrition and survey satisficing are less under ranked-choice than under forced-choice designs.*

H3 *Estimation efficiency is greater under ranked-choice than under forced-choice designs.*

H4 *Accuracy of attribute importance is greater under ranked-choice than under forced-choice designs.*

3.2 Experimental design

To assess these hypotheses, we compare the ranked-choice conjoint (RCC) design against the standard forced-choice conjoint (FCC) by running parallel randomized experiments. At the beginning of the experiment, subjects are randomly assigned to one of two conditions: either completing a series of forced-choice conjoint experiments or a series of ranked-choice conjoint experiments.

In practise, conjoint experiments are deployed to help answer research question in a wide variety of settings. It may be that the advantages of the ranked-choice design are more apparent in some of these contexts compared to others. For example, ranking tasks may be easier and more engaging, compared to forced-choice tasks, when the task is more familiar (e.g. voting).

Therefore, in our survey, we field two separate conjoint experiments. The first experiment is a typical *candidate election* context: we present subjects with randomised descriptions of candidates running for the US House of Representatives, and ask who they would prefer to vote for. The candidate attributes, and values, are a replication of Kirkland & Coppock (2018). Within each

[5] In this paper, we set aside one thorny dimension of the external validity of these "preferences". As is well documented, the AMCE is not necessarily an estimate of the "real world" preferences subjects possess, because the AMCE is an avergae over the distribution of the attribute-levels for the other attributes (Abramson et al. 2019, De la Cuesta et al. 2022). Therefore, if the real world joint distribution of these values differs to the distribution in the experiment (which is likely), then the AMCE might not reflect what a voter, or individual more generally, may *prefer* in the real world.

If you had to rank the different options, where rank 1 is your most preferred option and rank 3 your least preferred option, how would you rank the options?

	Budget A	Budget B	Budget C
Health and Medicare spending	Remain the same	Increase	Decrease
Military spending	Increase	Increase	Increase
Education spending	Increase	Decrease	Remain the same
Environmental protection spending	Increase	Decrease	Increase
Social security spending	Decrease	Remain the same	Decrease

Figure 3: Example of the ranked-choice budget vignette

profile we describe the candidate's age, race, gender, previous job, previous political experience, and their partisanship. A full description of the attribute-levels can be found in Table 1.

Table 1: Conjoint attributes and levels: Conjoint I

Race	Political experience	Career experience	Gender	Age	Party
White	None	Educator	Female	35	Independent
Hispanic	School board president	Stay-at-home Mom/Dad	Male	45	Democrat
Black	City council member	Small business owner		55	Republican
Asian	State legislator	Police officer		65	
	Representative in congress	Electrician			
	Mayor	Business executive			
		Attorney			

The second experiment is a budget proposal experiment. We ask subjects to consider descriptions of *federal budget proposals* where we outline whether spending will "increase", "decrease", or "stay

the same" along each of five funding areas: health and medicare, military, education, environmental protection, and social security spending. Our design is inspired by the austerity conjoint posed by Bansak et al. (2021), except we avoid this framing given that balancing budgets implies trade-offs across spending areas, which we and others (e.g. Barnes et al. 2022) think are not well-suited to conjoint designs. The conjoint attributes are summarised in Table 2.

Table 2: Conjoint attributes and levels: Conjoint II

Education spending	Environmental protection spending	Health and Medicare spending	Military spending	Social security spending
Decrease	Decrease	Decrease	Decrease	Decrease
Remain the same	Remain the same	Remain the same	Remain the same	Remain the same
Increase	Increase	Increase	Increase	Increase

All subjects completed both the candidate and policy conjoint experiments, and the order of the two experiments was randomised at the subject-level. In each experiment, those in the forced-choice arm completed $j = 6$ tasks with $k = 2$ profiles per vignette. In the ranked-choice arm, respondents completed $j = 4$ tasks with $k = 3$ profiles each. As a result, subjects in both treatment arms evaluated 12 randomised profiles per experiment. Figure 3 shows an example ranked-choice vignette from the budget policy experiment.

Conjoint outcome measure In the forced-choice arm, subjects are asked which candidate they would vote for and which policy they most prefer, using radio boxes to indicate their preference. In the ranked-choice arm, subjects are asked to rank the profiles by dragging labels for the three profiles into descending order of preference.

Pre-treatment questions and manipulation checks Prior to completing either conjoint experiment, we collect basic demographic questions about each subject, and some general questions on their political attitudes. After completing *each* experiment, we also ask a series of factual manipulation check questions (discussed below). Separately, we present a single, constant conjoint profile and ask whether they would vote for/accept this proposal (the response choices are binary regardless of treatment arm). We use these observations to assess latent preferences, which we discuss in more detail below.

Sample The survey was fielded in June 2024, to a representative sample of the United States voting-age population. Subjects were invited to complete our survey using the Prolific platform.

	Total time taken			Time per observation		
	Candidate	Policy	Combined	Candidate	Policy	Combined
Ranked-choice	44.40***	134.26**	44.40	−2.72***	−0.04	−2.72
	(7.02)	(50.39)	(35.97)	(0.35)	(2.11)	(1.51)
Policy			25.67			2.14
			(35.70)			(1.50)
Ranked x Policy			89.86			2.67
			(50.87)			(2.14)
(Intercept)	109.67***	135.34***	109.67***	9.14***	11.28***	9.14***
	(4.93)	(35.36)	(25.25)	(0.25)	(1.48)	(1.06)
R^2	0.02	0.00	0.01	0.03	0.00	0.00
Adj. R^2	0.02	0.00	0.00	0.03	−0.00	0.00
Num. obs.	2211	2211	4422	2211	2211	4422

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 3: Effect of ranked choice (compared to forced choice) on completion time

Prolific subjects have been shown to have higher rates of attention than other online survey platforms (Douglas et al. 2023, Albert & Smilek 2023), and our own results are consistent with relatively high levels of attention. Subjects were paid $5.98/hr for completing the survey, based on an expected completion time of 8 minutes (in line with Prolific's compensation guidelines). In total, 2211 subjects completed our survey.

4 Results

4.1 Survey duration

Our first hypothesis concerns the time it took subjects to complete the conjoint experiments. For each subject, we recorded the duration it took to complete each vignette and calculated the total time to complete the experiment as a whole. We test the difference in the means of the total time, across conditions, using simple linear regression.

The first three columns of Table 3 reports these results, for both measures, across both experiments. We find that those completing ranked-choice tasks spend, on average, 44 seconds longer on the candidate experiment, and just over two minutes longer on the policy task. Despite the substantial difference between these two estimates, when we pool the data we do not find a statistically significant difference between the time taken to perform the policy and candidate experiments, for

those in the ranked condition.

From a data efficiency perspective, since we recover twice the number of observations in the three-profile ranked-choice design, we also estimate the time taken *per effective observation*. The second three columns of Table 3 report these results. In the candidate experiment, we find an efficiency advantage: per observation gained, subjects are almost 3 seconds quicker in the ranked-choice condition. In the policy condition, the time difference is statistically insignificant (and substantially very small).

Hence, contrary to our pre-registered hypothesis, and holding constant the number of profiles evaluated, the forced-choice design takes less time to complete. Reassuringly, however, given the additional number of observations gained from the ranked-design, we do find some small time efficiencies in the context of the candidate experiment. In the policy experiment, clearly the ranked choice task proved difficult, with subjects taking considerably longer to complete the experiment. In this case, we do not find a time efficiency at the observation-level, although we do not find a time inefficiency either.

4.2 Sample attrition and satisficing

All subjects who started our experiment (and did not screen themselves out via the consent questions), completed the experiment and so in neither condition did we observe attrition. This may be explained by the slightly more attentive Prolific subject pool compared to other online platforms, and the relatively short length of our survey. As a result, we cannot test whether attrition rates differ across conditions.

To check for satisficing, we employed a series of factual manipulation checks (FMCs) as proposed by Kane & Barabas (2019), using both treatment-irrelevant and treatment-relevant measures, to capture attentiveness to the experimental manipulation of the designs. These FMCs differ slightly across the two experiments. In both experiments, we asked subjects to report how many attributes they saw per profile (scored as the absolute difference between the true number and the subject's answer), and which of three example vignettes looked most similar to what they saw in the previous experiment. In the candidate election context, we also asked subjects to select which of a series of attributes were shown (of which 2 out of the 5 were not in the vignettes). In the budget experiment, subjects were also asked to identify the policy area considered in the experiment (federal budget

	Number of attributes		Identified vignette		Dimensions	
	Candidate	Policy	Candidate	Policy	Candidate	Policy
Ranked-choice	0.28^{***}	-0.02	-0.04^{*}	-0.03^{**}	-0.03	-0.01
	(0.03)	(0.03)	(0.02)	(0.01)	(0.03)	(0.01)
(Intercept)	0.65^{***}	0.52^{***}	0.78^{***}	0.97^{***}	2.56^{***}	0.98^{***}
	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
R^2	0.03	0.00	0.00	0.00	0.00	0.00
Adj. R^2	0.03	-0.00	0.00	0.00	-0.00	0.00
Num. obs.	2160	2197	2209	2209	2211	2210

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Effect of ranked choice (compared to forced choice) on manipulation checks

spending).

Table 4 reports the estimated difference in means by regressing subjects' performance on the manipulation test on whether they completed the ranked-choice task, for each FMC separately. When identifying the number of attributes in the candidate experiment, ranked-choice subjects are marginally more likely to make a mistake, although this difference does not hold in the policy experiment. In both experiments, subjects in the ranked choice condition are marginally less likely to choose the right vignette, by 4 and 4 percentage points respectively. Finally, in both experiments, there is no distinguishable difference in subjects' ability to identify the types of attributes shown.

Again, our pre-registered hypothesis was the subjects' performance would be better under the ranked-choice design. We do not find evidence to support this claim. Instead, and if anything, ranked-choice subjects perform slightly worse, although in all cases the differences are not substantially large.

4.3 Efficiency

Figures 4 and 5 plot the estimated AMCEs from linear probability models, estimated on the ranked-choice and forced-choice experimental data separately. Overall, we find that the substantive interpretation of the AMCEs are very similar: in almost all cases, AMCEs share the same sign and are on the same order of magnitude. If anything, we see a slight moderation in the effect sizes for ranked-choice AMCEs, compared to their forced-choice equivalents.

What is notable from these results models, however, is that in every case the confidence intervals

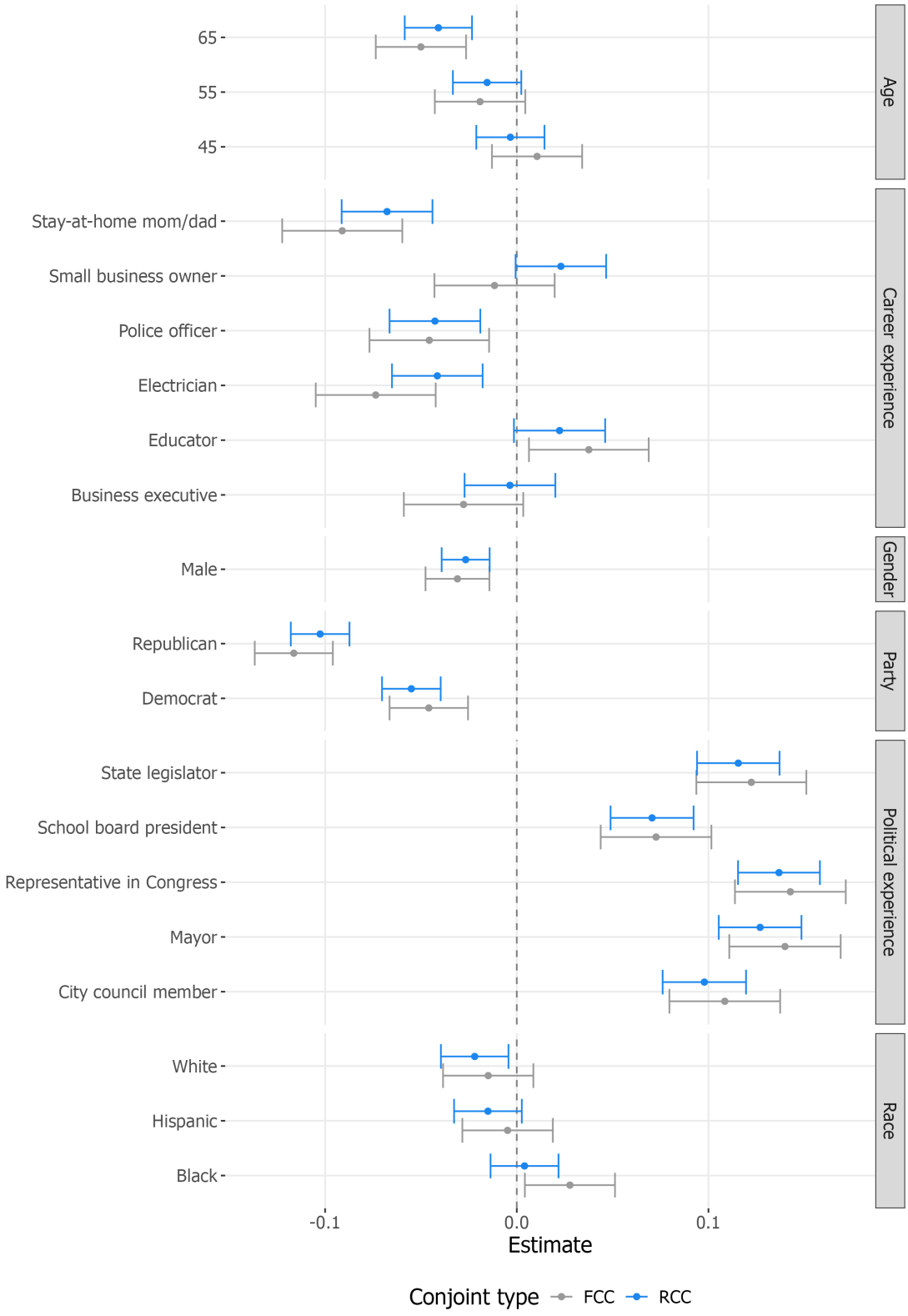

Figure 4: Estimated AMCEs from the candidate experiment using ranked and forced choice conjoints

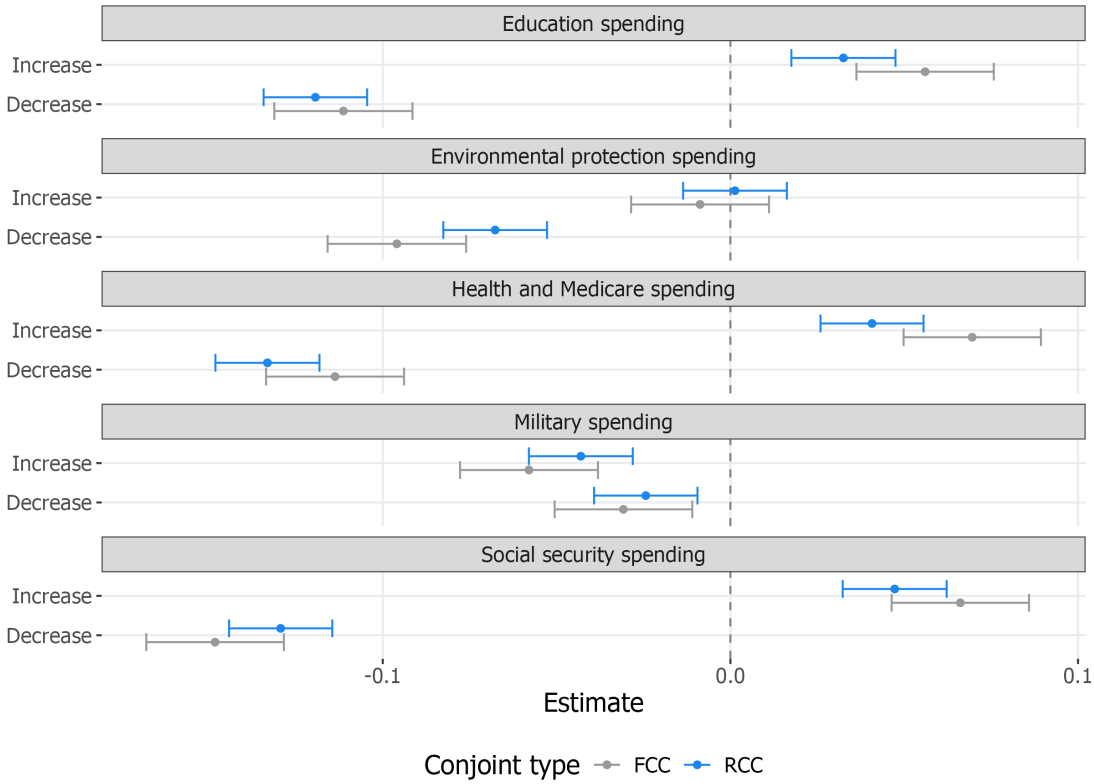

Figure 5: Estimated AMCEs from the policy experiment using ranked and forced choice conjoints

are narrower for the ranked-choice coefficient. These reductions are sizable: on average, and for both experiments, the RCC design yields standard errors that are 24.8% smaller than the forced-choice equivalent. The rate of reduction is very consistent across coefficients: the standard deviation of these reductions is just 0.006 in the case of the candidate experiment, and 0.003 in the case of the budget experiment.

Consistent with our hypothesis, and simulation results, the ranked-conjoint design does yield more precise estimates. This is a mechanical advantage, that stems from expanding pairwise comparisons from ranked choices. When estimating marginal shifts in behaviour that may be quite small, this increase in precision can help us detect these effects more easily.

4.4 Preferences

4.4.1 Implicit versus explicit measures of importance

We evaluate whether subjects' preferences are better captured under the ranked-choice design in two ways. First, we assess whether the ranked design yields different implicit measures of importance compared to the forced-choice design, and whether both of these diverge from explicit measures of importance.

In the post-treatment portion of our survey, we asked subjects to score the importance to their decision-making of each attribute in the conjoint experiments. The sum of these scores was constrained to equal 100 exactly. We compare these scores, to an implicit measure of importance, calculated using the estimated AMCEs *conjoint experiment*:

$$\text{Importance}_l = 100 \times \frac{\frac{1}{v_l} \sum_v |\beta_{lv}|}{\sum_l \frac{1}{v_l} \sum_v |\beta_{lv}|},$$

which reflects the average absolute effect size for attribute l, normalised by the total sum of these averages over *all* attributes so that they also vary between 0 and 100. To recover an estimate of the uncertainty of the implicit importance scores, we bootstrap these estimates 1000 times.

Figure 6 plots a comparison of the implicit importance scores, estimated via the AMCEs, against the explicit post-treatment survey item in our experiment. We do not find substantively meaningful differences between the two conjoint conditions. Interestingly, we do find that both AMCE-derived measures diverge from the explicit measure for some attributes. For example, while respondents

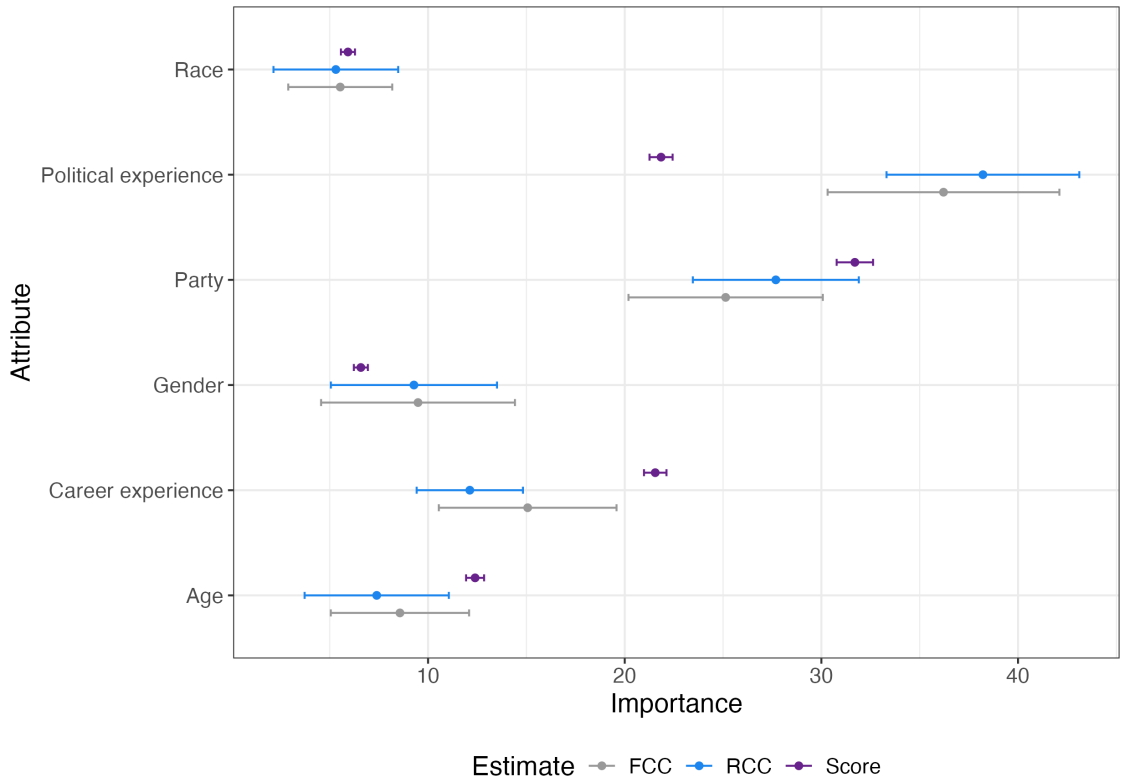

Figure 6: Estimated candidate attribute importance scores, compared against explicit survey measure

downplay the role of political experience in the explicit post-treatment item, the average absolute size of the political experience attributes is larger than any other attribute.

The results from the budget experiment are similar, as shown in Figure 7. The two implicit measures of policy importance do not substantially diverge between ranked- and forced-choice conditions, but again, for social security spending, we find that the explicit measure downplays the importance of the dimension relative to how voters actually behave.

The divergence between implicit and explicit measures is, in a sense, reassuring: conjoint experiments, like many behavioural instruments, are often used precisely because we suspect direct elicitation of preferences may be prone to bias. Thus, this component of our study provides some reassurance that these stated preference experiments, under both ranked- and forced-choice designs, are capturing something different to simpler, direct measures.

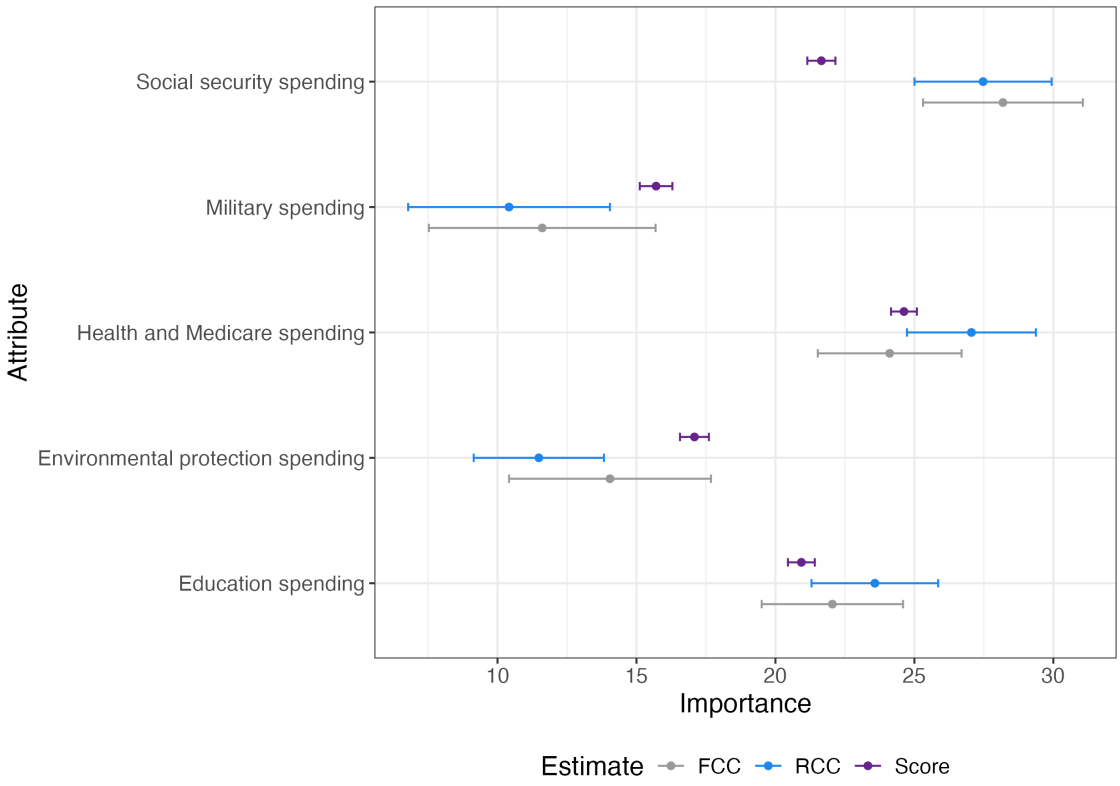

Figure 7: Estimated important scores for budget policy dimensions, compared against explicit survey measure

4.4.2 Estimated preference response-surfaces

It may be the case that the similarity in *A*MCEs is masking differential levels of information in the data about how these preferences vary across subjects. For example, in both experimental contexts we consider in this paper, it is likely that those with different ideological viewpoints will react differently to the *same* attribute-level values. Most generally, we can think about estimating response surfaces across the decisions in the conjoint that vary based on subjects' characteristics.

To assess this possibility, and following a similar method used by Bansak et al. (2021) to estimate how conjoint preferences translate to real world decisions, we use machine learning to model the conjoint response surface. Using our experimental data, we train a gradient boosted tree model for each combination of conjoint design and experiment context. We include in these models, subjects' pre-treatment answers to a battery of demographic and political preference questions.[6]

In the post-treatment sections of the experiment, we asked respondents to consider a single candidate/policy bundle profile, and to indicate whether they would vote for/accept this proposal (a binary outcome). The attribute-levels are identical for all subjects, and the outcome is binary regardless of which conjoint condition the subject was randomised into.

We use subjects' responses to these prompts as "test" set data for our models. We take our trained models and predict subjects' behaviour on the single, post-treatment profile, then calculate the classification performance of the model. Since we tune these models separately, such that the optimal classification threshold may vary, we use the area under the receiver-operator curve (AUC) as our measure of how well the models have estimated the response surface. We conduct a significance test of the difference in AUC scores, across conjoint design, using the boostrapped version of DeLong's test. Our expectation is that, if the RCC enables subjects to better engage/express their latent preferences, then the ROC scores will be higher than under the FCC condition.

Table 5 reports the results of this test. Interestingly, in both cases, the models significantly differ in their modelling performance, but in opposite directions. In the candidate experiment, and as we hypothesised, the ranked choice experiment performs substantially better than the forced-choice

[6] All models are tuned separately, using a 5-fold cross-validation process and classification accuracy metric. We grid search across the *nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight, and subsample* hyperparameters. Given the prohibitively high dimensionality of this search, we follow established practise by tuning these hyperparameters sequentially, using the optimal value in all subsequent tuning steps. The number of rounds is re-tested in every step.

	ROC Area Under Curve			
Experiment	*FCC*	*RCC*	**D**	***p*-value**
Candidate	0.51	0.64	-6.62	0.00
Budget	0.69	0.63	3.89	0.00

Table 5: Comparison and significant test of classification performance on out-of-sample conjoint profiles

model. In fact, although the AMCE estimates are very similar in these two designs, it appears that encoding more information via ranking gives the model a considerable advantage in learning how behaviours vary across subjects.

That said, in the policy experiment, the forced-choice model is statistically significantly better than the ranked-choice alternative. The difference in performance in this context is less stark however, with both models being reasonably predictive and the difference in AUC's being less than half the difference in the candidate experiment.

5 Discussion

In this paper, we introduce how ranked choices can be incorporated into conjoint experiments. This work builds on recent contributions in political science that point to the direct relevance, and potential performance benefits, of using rankings in the study of political phenomena (e.g. Kaufman et al. 2021, Atsusaka 2022, 2023). Under an additional transitivity of preferences assumption, and an expansion of ranks back to implicit pairwise comparisons, we can preserve the design mechanics of the traditional conjoint experiment while collecting more information from the same number of randomised profiles.

The major mechanical advantage of rankings is that, for every extra profile, we recover *multiple* new observations. Moreover, this scaling increases non-linearly as the number of profiles increases, due to the inflated number of pairwise comparisons. Of course, this expansion requires subjects' rankings to be transitive. In practise, we sugggest researchers can include validation profiles as part of the post-treatment phase of a survey to help assess any violations of this assumption.

Experimentally, we find that the AMCEs are reassuringly consistent with their forced-choice alternative. Due to the increased effective number of observations, however, the ranked-choice

AMCEs are substantially more precisely estimated. While subjects take longer to complete the conjoint experiment, for the amount of data one recovers, the time taken is, at worst, as efficient as in the forced-choice design. We do not find that subjects in the ranked condition pay more attention than in the forced-choice, but there is indicative evidence that ranked choice conjoint allows us to learn a more accurate response surface, in the context of elections.

One interesting, and recurring, feature of our results is the difference in subject performance across the two political contexts we consider: candidate elections and federal budget planning. Subjects were considerably slower at completing the budget policy experiment (in absolute terms), and we also find that the data from this experiment yields a worse-performing model of the response surface. These results are suggestive that the (dis)advantages of ranked choice methods vary across context. It may be that the federal budget context is just harder to rank: the consistent attribute-levels across dimensions may make it harder to realise one's ranking, for example. In the case of the candidate experiment, however, we do see both efficiency and measurement advantages, which is promising given the direct analogue to very common electoral systems.

We do not consider in this experiment the effects of *varying* the number of ranked profiles. Future research may want to consider whether, given the rapid increase in effective observations as the number of profiles increases, one may be able to trade-off the number of profiles against the extra absolute time required for subjects to rank these choices.

References

Abramson, B. S. F., Koçak, K. & Magazinnik, A. (2019), Do We Learn About Voter Preferences From Conjoint Experiments?

Albert, D. A. & Smilek, D. (2023), 'Comparing attentional disengagement between prolific and mturk samples', *Scientific Reports* **13**(1), 20574.

Atsusaka, Y. (2022), 'Causal inference with ranking data: Application to blame attribution in police violence and ballot order effects in ranked-choice voting', *arXiv preprint arXiv:2207.07005* .

Atsusaka, Y. (2023), 'Analyzing ballot order effects when voters rank candidates'.
URL: *osf.io/6mp3n*

Ballard-Rosa, C., Martin, L. & Scheve, K. (2017), 'The structure of american income tax policy preferences', *The Journal of Politics* **79**(1), 1–16.

Bansak, K., Bechtel, M. M. & Margalit, Y. (2021), 'Why austerity? the mass politics of a contested policy', *American Political Science Review* **115**(2), 486–505.

Bansak, K., Hainmueller, J. & Hopkins, D. J. (2018), 'The Number of Choice Tasks and Survey Satisficing in Conjoint Experiments', *Political Analysis* **26**, 112–119.

Barnes, L., Blumenau, J. & Lauderdale, B. E. (2022), 'Measuring attitudes toward public spending using a multivariate tax summary experiment', *American Journal of Political Science* **66**(1), 205–221.

Clayton, K., Horiuchi, Y., Kaufman, A. R., King, G. & Komisarchik, M. (Working Paper), 'Correcting measurement error bias in conjoint survey experiments'.

De la Cuesta, B., Egami, N. & Imai, K. (2022), 'Improving the external validity of conjoint analysis: The essential role of profile distribution', *Political Analysis* **30**(1), 19–45.

Douglas, B. D., Ewell, P. J. & Brauer, M. (2023), 'Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona', *Plos one* **18**(3), e0279720.

Duch, R., Loewen, P., Robinson, T. & Zakharov, A. (Working paper), 'Governing in the face of a global crisis: when do voters punish and reward incumbent governments?'.

Duch, R. M., Laroze, D., Reinprecht, C. & Robinson, T. S. (2022), 'Nativist policy: The comparative effects of trumpian politics on migration decisions', *Political Science Research and Methods* **10**(1), 171–187.

Golder, M. (2005), 'Democratic electoral systems around the world, 1946–2000', *Electoral Studies* **24**(1), 103–121.

Hainmueller, J., Hopkins, D. J. & Yamamoto, T. (2014), 'Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices Via Stated Preference Experiments', *Political Analysis* **22**(1), 1–30.

Kane, J. V. & Barabas, J. (2019), 'No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments', *American Journal of Political Science* **63**, 234–249.

Kaufman, A. R., King, G. & Komisarchik, M. (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.

Kirkland, P. A. & Coppock, A. (2018), 'Candidate Choice Without Party Labels: New Insights from Conjoint Survey Experiments', *Political Behavior* **40**, 571—591.

Leeper, T. J., Hobolt, S. B. & Tilley, J. (2020), 'Measuring subgroup preferences in conjoint experiments', *Political Analysis* **28**(2), 207–221.

Spencer, D. M. & Theodoridis, A. G. (2020), '"appearance of corruption": Linking public opinion and campaign finance reform', *Election Law Journal: Rules, Politics, and Policy* **19**(4), 510–523.

Vermeulen, B., Goos, P. & Vandebroek, M. (2011), 'Rank-order choice-based conjoint experiments: Efficiency and design', *Journal of Statistical Planning and Inference* **141**(8), 2519–2531.

Appendices

A1 Deviations from the pre-analysis plan

The empirical analysis was followed as detailed in our pre-registered, pre-analysis plan, except for the following minor deviations:

- In our discussion of H3 (efficiency), we stated that RCC should yield more "accurate" estimates. This should have read "precise". The analysis remains as stated in the plan.
- For the explicit test of H4 (latent preferences), our attribute importance score was calculated using the *sum* of AMCEs within each attribute. We amended this to the *mean* of AMCEs in this paper, to take into account that, in the candidate context, different attributes had different numbers of levels.
- Under H3 (efficiency) we stated we would estimate the adjusted R^2 and ROC AUC scores for the models. These results are now included in our discussion of H4.

A2 Further figures and tables

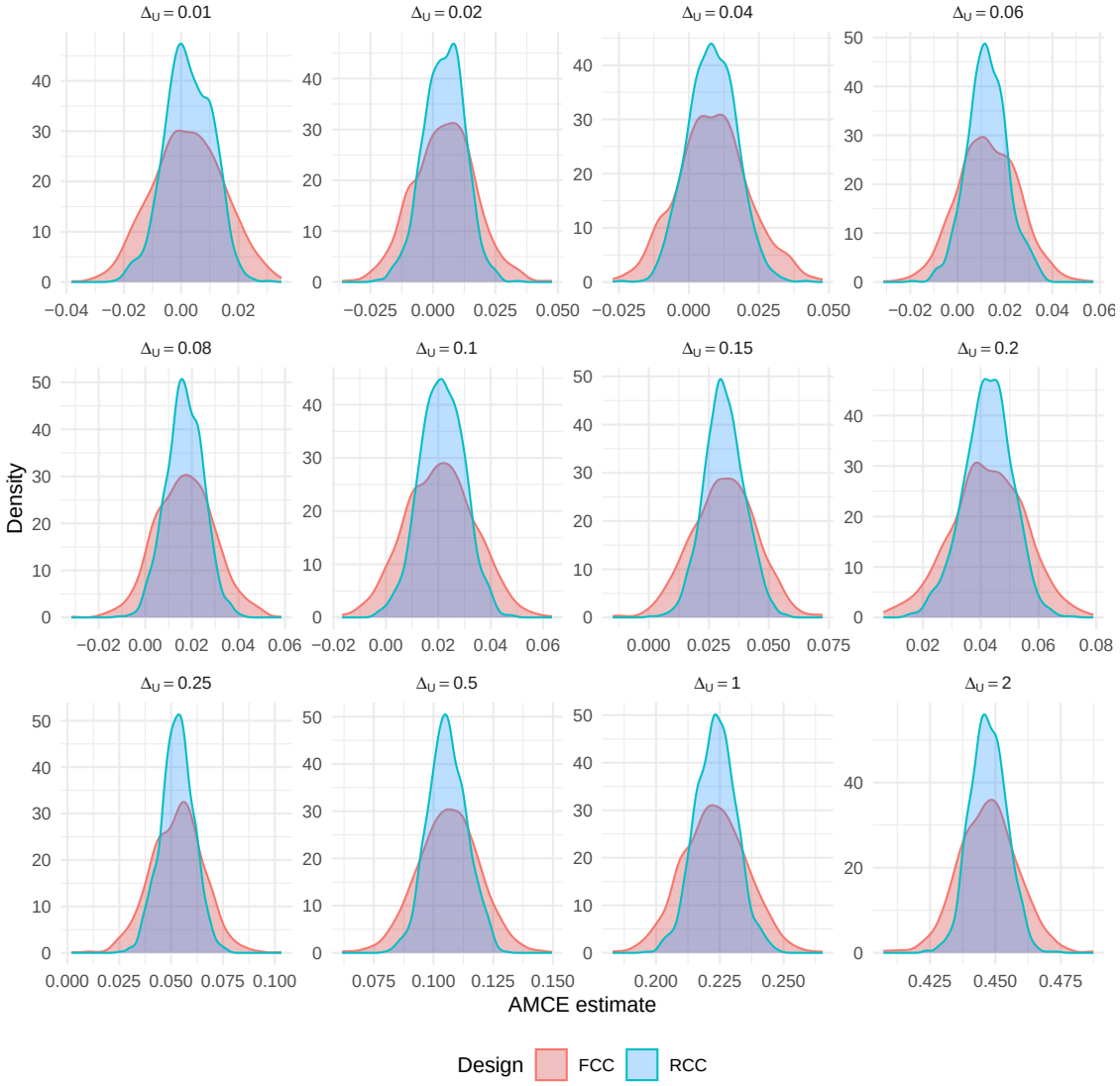

Figure A1: Density of AMCE estimates using RCC and FCC, respectively, over 1000 simulations. Each panes reflect the changes in utility (measured in standard deviations of the residual variance) for showing the same attribute-level.