Lancaster University
Management School

Economics Working Paper Series

2025/010

Who Likes It More? Using Response Times to Elicit Group Preferences in Surveys

Carlos Alós-Ferrer and Michele Garagnani

The Department of Economics
Lancaster University Management School
Lancaster LA1 4YX
UK

LUMS home page: http://www.lancaster.ac.uk/lums/

Who Likes It More? Using Response Times To Elicit Group Preferences in Surveys*

Carlos Alós-Ferrer† Michele Garagnani
Lancaster University University of Melbourne

This version: May 2025

Abstract

Surveys remain crucial tools for measuring societal preferences, but their reliability is limited by noise and bias in respondent data. We introduce a novel non-parametric method that leverages response times to reveal group preferences and rank preference strength across different populations. We validate the approach and apply it to key socio-economic questions using large representative surveys. The method complements traditional survey analysis techniques, providing clear indicators of when standard analyses may be inadequate and when response time data can yield additional insights. Importantly, our method also quantifies response biases, allowing researchers to adjust for systematic distortions in survey data.

JEL Classification: D11 · D81 · D83 · D87

Keywords: Survey Data · Revealed Preference · Response Times · Response Bias

1 Introduction

Survey data is extensively used to elicit societal preferences and attitudes, ranging from the support for redistributive policies to the willingness to pay for a new product. The advent of online survey platforms and the increased availability of extensive, well-maintained national-level panels have considerably increased the use of survey data in economics, political science, marketing, health research, and other disciplines.

One important question that surveys aim to answer is whether a given economic policy enjoys sufficient support in the population and is hence politically-sustainable.

*We are grateful to Stefano DellaVigna, Ido Erev, Ernst Fehr, Antonio Filippin, Holger Gerhardt, Fabio Maccheroni, Ulrike Malmendier, Luca Polonio, and seminar participants at the University of Basel, briq/University of Bonn, HEC-Lausanne, University of Milano-Bicocca, Paris School of Economics, the 2024 Psychoeconomics Workshop in Lancaster, and plenary sessions at the ESA Asia-Pacific conference in Osaka 2025 and the Newcastle Experimental Economics Workshop 2024.

†Corresponding author: c.alosferrer@lancaster.ac.uk. Lancaster University Management School, Lancaster, LA1 4YX, UK. Funded by the research budget of Carlos Alós-Ferrer while affiliated with the Center for Neuroeconomics at the University of Zurich (Switzerland). No external funding was received.

For example, risk, time, and social preferences determine the support for social security policies (e.g., health or unemployment protection), investments with delayed benefits (e.g., energy policy, infrastructures), and policies affecting future generations (e.g., climate protection) or non-citizens (e.g., migration), respectively (e.g., Falk et al., 2018; Bechtel and Liesch, 2020; Bechtel et al., 2020; Enke et al., 2022). A further, important question of interest is how groups differ in their attitudes, e.g. whether a redistributive policy would receive stronger support from the left or the right of the political spectrum, whether females would support a policy more than males, whether preferences differ across racial groups, or whether people with a higher socioeconomic status would demand a new service more than others (e.g., Fisman et al., 2006, 2008; Leeper et al., 2020; Giglio et al., 2021; Snowberg and Yariv, 2021; Ortoleva et al., 2021).

While survey data has been extensively used in economics (e.g. Bertrand and Mullainathan, 2001; Manski, 2005; Apesteguia and Ballester, 2023), it is notoriously inexact. First, even under ideal conditions, human choices are inherently noisy (e.g., Tversky, 1969; Camerer, 1989; Hey and Orme, 1994; Alós-Ferrer et al., 2021). The problem is exacerbated for survey data, due e.g. to increased measurement errors or possible lack of respondents' attention or understanding. A second and even more important problem is that surveys are often biased and responses might not reflect actual preferences. In particular, questions on sensitive topics are often subject to systematic survey misreporting due to *social desirability bias*, i.e. the misreporting of own preferences in surveys to be more aligned with attitudes perceived as socially acceptable (Zaller and Feldman, 1992; Coffman et al., 2017; Guriev and Treisman, 2020). For example, using data from 20 years of after-election phone surveys in Switzerland, Funk (2016) reports sizable differences between surveys and actual election results and identifies a "liberal bias" where initiatives perceived to be more prosocial or liberal receive a larger stated support in surveys than in reality. A different but related problem is *experimenter demand*, which affects economic experiments and surveys (Zizzo, 2010). Both survey bias and experimenter demand are important causes of concern which have motivated an extensive literature and increasingly-sophisticated methods to explore the robustness of conclusions derived from self-reported data obtained in surveys or experiments (e.g., Luce and Tukey, 1964; Hainmueller et al., 2015; de Quidt et al., 2018). Importantly, both biases affecting survey data can be expected to persist in repeated measurements (Belzil and Jagelka, 2024) and thus cannot be removed by applying standard techniques for dealing with measurement error such as the ORIV method popularized by Gillen et al. (2019).

In this contribution, we develop and validate a method for preference revelation with survey data which leverages a readily-available, additional source of information to improve over and complement existing analyses. Specifically, we rely on the joint use of choice frequencies and response times. The latter are nowadays easy to collect in online platforms and surveys, and indeed are typically already being passively collected (and then neglected) by standard software. Hence, the method we propose adds no cost to existing survey designs, and can be easily implemented whenever survey data

(or laboratory data) is collected electronically. Crucially, the method allows to reveal groups' preferences with minimal assumptions about the underlying behavioral model. It is non-parametric and hence agnostic with respect to specific utility shapes or functional forms for the distribution of behavioral noise. Overall, the method is easy-to-implement, costless, and complementary to existing procedures, which improves the quality of the inference that can be drawn from surveys.

The key insight behind our method derives from the cognitive sciences and is the link between response times and strength of preference (e.g. Dashiell, 1937; Moyer and Landauer, 1967; Laming, 1985; Shadlen and Kiani, 2013, among many others), which are receiving increasing attention in economics (e.g. Fudenberg et al., 2018; Baldassi et al., 2020), and which imply that the distribution of response times contains information on the underlying distribution of behavioral noise. Recently, Alós-Ferrer et al. (2021) built upon those regularities to provide individual preference revelation conditions using response times for settings where a single decision maker has repeatedly made the same decision.[1] In contrast, our method considers groups of many individuals but requires only one decision per individual, and is hence appropriate to use with electronically-collected survey data as well as laboratory experiments. That is, while Alós-Ferrer et al. (2021) considers the (laboratory) situation where one individual makes a fixed decision many times, here we consider the (survey) case where many respondents make a given decision, but each person makes that decision once. This setting makes our techniques immediately useful for a wider set of applications, including standard survey data.

Our main theoretical results assume a generalization of a utilitarian framework, which is equivalent to a standard population-level interpretation of random utility models as used in economics (McFadden, 1974, 2001; Anderson et al., 1992), marketing (e.g., Baltas and Doyle, 2001; Feng et al., 2022), political science (e.g., Nownes, 1992; Karp, 2009), and many other fields. That is, the preferences of a group are described by a distribution over utility functions, or, equivalently, a utility function and a distribution of behavioral noise, which can be seen as the individual deviations from the utility function (Block and Marschak, 1960). In this context, survey bias is intuitively easy to conceptualize and visualize. Imagine that preferences within a group mainly favor option x over option y, but social desirability bias creates a tendency to provide a y response. Within a random utility model, this will result in an *asymmetric* distribution, where the median differs from the mean. Hence, it is possible that more than half of the surveyed individuals give a y response (reflecting the median), while in reality the group preference (the mean) is in favor of x. Analogously, it is possible that the bias affects a group more than another (e.g., due to gender effects or cohort differences), biasing results on group differences. As pointed out by Alós-Ferrer et al. (2021) (there for the case of individuals), the use of choice frequencies alone might lead to unwarranted conclusions over preferences.

[1] We refer the reader to Alós-Ferrer et al. (2021) and Alós-Ferrer and Garagnani (2022a,b) and the references therein for additional evidence and discussion on *psychometric* and *chronometric effects*. Alós-Ferrer et al. (2024) extend the results in Alós-Ferrer et al. (2021) to analyze whether apparent transitivity violations are due to actual failures of transitivity or just to behavioral noise.

We develop the theoretical results necessary for our method, which build upon Alós-Ferrer et al. (2021). We obtain two main theoretical results, which derive conditions for when the data allow to reveal a group preference (Theorem 1), and when the preference of one group is revealed to be stronger than the preference of another group (Theorem 2), respectively. Contrary to Alós-Ferrer et al. (2021), however, we develop statistical tests which allow us to test when the (finite) data allows to conclude that the conditions for preference revelation are fulfilled at conventional levels of significance.

We apply our theoretical and statistical techniques to two online surveys. Our framework is of course based on a series of assumptions, e.g. that the chronometric effect is the main determinant of response time differences in the questions of interest in a survey. Notwithstanding the strength of the confidence on these assumptions that we can obtain from the cognitive science literature, or how many works in economics have already established these effects (Moffatt, 2005; Chabris et al., 2009; Alós-Ferrer and Garagnani, 2022a,b; Liu and Netzer, 2023, and others) it is always possible to object to theoretical assumptions or postulate the next dimension which might influence response times and affect the results (see Section 5 for additional comments on this point). Hence, our first (pre-registered) survey focused exclusively on the validation of the method. That is, we applied the techniques to a series of uncontroversial questions where there is an objectively correct answer (e.g., which line is longer between two, a classical psychophysics task), or where there is a clear understanding of the direction of the preferences (framing effects, donations to ranked charities). We show that our method clearly and robustly reveals and distinguishes groups' preferences.

Having done this, we then demonstrate and illustrate the practical value of the new method by collecting survey data and response times for a representative sample of the U.K. population (again pre-registered), including a wide range of important socio-economic issues, e.g. inequality, gender discrimination, and environmental policies. These questions, and the conclusions we obtain for each of them, are of course of independent interest. Our focus, however, is on showcasing the applicability and implications of the proposed method for a variety of important issues, and hence we designed the survey to explicitly cover different fields of application.

Current standard practice to analyze survey data focuses on statistical tests using choice proportions only. For example, to examine preference differences across groups, the analyst compares the proportion of choices in favor of one option between the two groups. If the two proportions are statistically different, according, say, to a test of proportions, it is concluded that one group prefers the option more than the other, and the researcher has no other instrument to further evaluate comparison across groups. This approach can only reveal differences in stated preferences, which might be heavily biased, and fails to reveal differences in actual preferences.

Our results for the second survey show that, in some occasions, our method is able to get more from the data than standard analyses, i.e. obtain significant results when tests that use only choice frequencies return non-significant results. This alone would already

justify collecting response times in survey and applying our technique, since the cost is negligible. Interestingly, however, we also show that our method often (for almost 40% of the questions we implemented) contradicts significant results obtained when looking at choice frequencies only, and several cases where this happens are clearly under suspicion of social desirability bias. That is, our method can also serve to signal possible false positives in preference differences across groups. This is possible because, in the presence of survey bias, the question of preference revelation differs from the question of whether a larger fraction of individuals choose a given response. For example, some women might oppose gender quotas, but might feel compelled to express an opinion in their favor. Due to the clear interest on this topic, we further develop two complementary approaches to quantify possible survey bias in these cases, one based on Floodlight analysis as often applied in marketing (Spiller et al., 2013; Hayes, 2018), and one based on swapping responses before applying a test of proportions.

We want to insist that the purpose of our second survey is a demonstration of the general applicability of the method. In particular, we purposefully refrain from building any additional elements into our analysis or trying to account for any additional dimensions of response times, least we be accused of adapting the method in a different way for every application. That is, we apply the exact same method and analyses to every question in our survey, without any consideration of whether we would be able to fine-tune the data. For example, some of our questions compare younger vs. older people. It is natural to assume that response times are affected by age, and hence we could have estimated an effect of age on response times and detrended our data before applying our method. It would also have been very simple to adjust response times at the individual level, e.g. by estimating individual fixed effects in a first analysis (Chabris et al., 2009) or subtracting or dividing by the response time of a control question (Liu and Netzer, 2023). We don't. Any such additional fine-tuning of the data would obviously just improve our results, but do so at the expense of clarity. Hence we establish the strength of our method without any such recourse. However, we of course would advise applied researchers interested on specific group comparisons to use all fine-tunings as appropriate for their research, and in particular in the case of groups differing in age or any other dimensions with known effects on response times or choice consistency.

We contribute to the literature on the identification of what we can reliably learn from surveys (e.g. Liu and Netzer, 2023; Belzil and Jagelka, 2024), i.e., 'use implicit information (such as response times or attention) to identify better individuals' actual views" (Apesteguia and Ballester, 2023). Belzil and Jagelka (2024) develop a framework to estimate preferences from observed behavior while accounting for individuals' effort and (cognitive) noise. Alós-Ferrer et al. (2021) introduced the general framework we use for individual decisions, which we expand to group decisions. Alós-Ferrer and Garagnani (2024b) further applied results of that work to show that response times can be used to improve out-of-sample prediction in individual decisions under risk. Other recent works have extended the techniques to show that preference revelation through response times

can help identify behavioral anomalies in decision making, e.g. the certainty effect (Alós-Ferrer and Garagnani, 2024a) or violations of transitivity (Alós-Ferrer et al., 2024). Liu and Netzer (2023) apply a different revelation concept, based on stochastic dominance but also relying on response times, to ordinal responses, e.g., happiness scales. Their revelation concept is less natural for our applications. However, we reproduce this concept and the corresponding revelation result within our framework in Section 6, and show that the concept is empirically less fruitful than ours. A further, important difference with Liu and Netzer (2023) is that their statistical tests are based on an approach where the null hypothesis corresponds to the fulfillment of necessary conditions for preference revelation. This is quite unfortunate, since it means that their approach, unlike ours, can never actually reveal preferences (since one can only reject the null), and the analyst interested in preference revelation is then put in the strange position of having to interpret a non-significance. We discuss these issues in Section 6.3.

The remainder of the paper is structured as follows. Section 2 describes our formal population framework and derives our theoretical results. Section 3 describes our empirical approach and statistical tests. Section 4 describes the validation of the method through our first survey. Section 5 presents the analysis of the data from the second survey, showcasing a number of situations of interest and, in particular, possible approaches to identify survey bias. Section 6 analyzes a stronger revelation concept based on stochastic dominance, shows it to be empirically less attractive, and relates it to the application to happiness surveys by Liu and Netzer (2023). Section 7 concludes. The Appendix contains the formal proofs of our results, the detailed transcript of the surveys, and question-by-question analyses of the survey data.

2 A Formal Framework to Use Response Times in Surveys

The framework we work with builds upon the standard additive random utility model (RUM), widely used in applied microeconomics, which is equivalent to stochastic preference models (McFadden, 1974, 2001; Anderson et al., 1992). As a model of stochastic choice for an individual agent, an additive RUM postulates that the agent is endowed with a utility function u over a feasible set, but is affected by random utility shocks. Thus, given a choice between two alternatives x and y, realized utilities are $u(x) + \varepsilon_x$ and $u(y) + \varepsilon_y$, respectively, where ε_x, ε_y are zero-mean random variables. Thus, a RUM generates choice probabilities, with the probability of x being chosen when y is also available given by

$$P(x, y) = \text{Prob}(u(x) + \varepsilon_x > u(y) + \varepsilon_y) = \text{Prob}(\varepsilon_x - \varepsilon_y > u(y) - u(x)).$$

where tie-breaking conventions are irrelevant for continuously-distributed errors. Under specific assumptions on the distributions of the error terms, one obtains particular models, as the celebrated logit choice (Luce, 1959) or the classical probit choice (Thurstone,

1927). This general setting has become one of the dominant approaches in economics to model the fact that choice is empirically (and overwhelmingly) observed to be stochastic.

Alós-Ferrer et al. (2021) introduced a more general class of RUM models where error terms are modeled directly for utility differences, i.e. the realized utility difference given a choice $\{x, y\}$ is $u(x) - u(y) + \varepsilon_{x,y}$ for a mean-zero random variable $\varepsilon_{x,y}$ and hence

$$P(x, y) = \text{Prob}(\varepsilon_{x,y} > u(y) - u(x)).$$

This class encompasses additive RUMs, but also trembling-hand models (Loomes et al., 2002) where decisions follow a fixed strict preference but pair-specific errors might occur.

Building upon insights in response times from psychology and neuroscience, Alós-Ferrer et al. (2021) provided sufficient conditions on the distributions of response times conditional on each possible choice (x or y for a given pair (x, y)) which ensure that any RUM that fits the data (in terms of choices and response times) reveals a preference for, say, x over y, in the sense that $u(x) > u(y)$ for the underlying u. Those results are powerful because they guarantee that an option is preferred to another for *any* utility function and *any* distribution of the error term that the analyst might consider, and hence the results are completely non-parametric and independent of functional forms.

An alternative interpretation of (additive) RUMs follows a utilitarian approach (e.g., d'Aspremont and Gevers, 2002). In this approach, the random utility terms are interpreted to model unobserved heterogeneity in a population of agents. That is, instead of considering a distribution of choices for a single agent with trial-to-trial variability, one assumes a distribution of agents, each endowed with a fixed utility drawn from a distribution, such that $u(x) - u(y)$ is the mean of the distribution of utilities for the particular choice $\{x, y\}$. Thus, u measures the *utilitarian welfare* and a revealed preference for x over y means that x is to be preferred, in (utilitarian) welfare terms, to y. Of course, the *caveat* of this approach (and of the utilitarian approach as a whole) is that it requires interpersonal comparability of the units in which utility is measured.

If applied to data of a given individual, the results in Alós-Ferrer et al. (2021) require multiple repetitions of the same choice. The utilitarian interpretation of a RUM opens the door to a different kind of applications. In those, a dataset contains the choices and response times of a population for a fixed choice, but only one choice per individual. This is particularly attractive for field and survey settings, where choice repetition might be difficult to implement, but the collection of large numbers of responses for short questionnaires entails little difficulty.

This section derives new results which allow for preference revelation in terms of relative strength of preference across different subgroups of a population. Specifically, our main result identifies a simple, joint condition on choice frequencies and the distributions of response times on the groups which, if fulfilled, allows to unambiguously rank the relative strength of preference of the groups. In other words, the condition guarantees that group A prefers x over y *more* than group B, for any utilitarian model which fits

the data. The result is nonparametric, because the conclusion follows independently of which utility functions and models of noise are used to fit the data.

In addition to the results on comparisons across groups, we also provide preference revelation results for groups, along the lines of Alós-Ferrer et al. (2021) but for survey data. That is, we identify a condition on the choice frequencies and response times of any fixed group which, if fulfilled, guarantees that an aggregate preference exists for that group, again independently of which utility and model of noise is used to fit the data.

The main result on the comparison of preferences across groups, however, is independent of whether a preference for x over y for the separate groups is actually revealed by the data or not. That is, it is perfectly possible that the data does not allow to conclude that either group prefers x over y, and yet the researcher can conclude that the first group favors x over y more strongly than the second group. This allows discussing group differences even when within-group preference revelation fails.

Further, it is not necessary that the comparison refers to the same choice pair. In their most general formulation, the results allow to conclude that a certain group prefers x over y *more* than another group prefers z over w, for any four alternatives x, y, z, w. This is interesting because, even for identical questions on surveys, it can always be argued that different groups might perceive the options differently. For example, a given level of health care has different consequences for men and women purely on biological grounds, or a policy question might be formulated in terms of instruments (e.g., taxes or subsidies) rather than actual outcomes (e.g., wealth distributions). While this difficulty is usually glossed over on pragmatic grounds, we are able to tackle it head on. The results can be immediately applied to (i) the comparison of preferences for a fixed pair (x, y) across two different groups, (ii) the comparison of preferences for group-tailored choice pairs, (x_1, y_1) for one group and (x_2, y_2) for another group; (iii) the comparison of preferences across two different pairs (x, y) and (z, w) for a fixed group, and (iv) the staircase-like comparison of preferences across (x, y) and (y, z) for a fixed group, to mention just the most obvious possibilities.

The conditions identified here are formulated in terms of weakenings of first-order stochastic dominance between appropriate, conditional distributions of response times, or obvious reformulations thereof. For each actual dataset, how weak the weakening is depends on actual choice proportions, so that, generally speaking, a larger percentage of choices in favor of an option requires less information from response times (i.e. the conditions become stronger), and vice versa.

2.1 The Formal Setting

Let X be a finite set of options, and denote by $C = \{(x, y) \mid x, y \in X, x \neq y\}$ the set of all binary choice problems, so (x, y) and (y, x) both represent the problem of choice between x and y. Let $D \subseteq C$ be the set of choice problems on which we have data in

the form of direct choices, assumed to be non-empty and symmetric, that is, $(x, y) \in D$ implies $(y, x) \in D$. For instance, D might be the set of (binary) questions in a survey.

A population-level dataset is modeled as follows.

Definition 1. A *stochastic choice function with response times* (SCF-RT) is a pair of functions (p, f) where p assigns to each $(x, y) \in D$ a frequency $p(x, y) > 0$, with the property that $p(x, y) + p(y, x) = 1$, and f assigns to each $(x, y) \in D$ a strictly positive density function $f(x, y)$ on $\mathbb{R}_+$. In the particular case where D contains only one pair, $D = \{(x, y), (y, x)\}$, we say that (p, f) is an SCF-RT for (x, y).

In our population setting, $p(x, y)$ is interpreted as the fraction of the population who chose x when offered the binary choice between x and y. The assumption that $p(x, y) > 0$ for all $(x, y) \in D$ implies that population choice is non-degenerate, that is, both alternatives enjoy at least some support. The density $f(x, y)$ describes the distribution of response times conditional on the subset of agents who chose x in the binary choice between x and y. The corresponding cumulative distribution function is denoted by $F(x, y)$. The following definition is taken from Alós-Ferrer et al. (2021).

Definition 2. A *random utility model with a chronometric function* (RUM-CF) is a triple $(u, \tilde{v}, r)$ where $u : X \to \mathbb{R}$ is a utility function and $\tilde{v} = (\tilde{v}(x, y))_{(x,y)\in C}$ is a collection of real-valued random variables, with each $\tilde{v}(x, y)$ having a density function $g(x, y)$ on $\mathbb{R}$, fulfilling the following properties:

(RUM.1) $\mathbb{E}[\tilde{v}(x, y)] = u(x) - u(y)$,

(RUM.2) $\tilde{v}(x, y) = -\tilde{v}(y, x)$, and

(RUM.3) the support of $\tilde{v}(x, y)$ is connected.

Further, $r : \mathbb{R}_{++} \to \mathbb{R}_+$ is a continuous function that is strictly decreasing in v whenever $r(v) > 0$, with $\lim_{v\to 0} r(v) = \infty$ and $\lim_{v\to\infty} r(v) = 0$.

The utility function u is interpreted as a representative utility for a population or group. The random variables $\tilde{v}(x, y)$ incorporate heterogeneity across individuals, modeled in terms of the distribution of random pairwise utility differences. That is, the density $g(x, y)$ describes the population distribution of the utility differences between x and y, with condition (RUM.1) requiring that the population mean identifies the utilitarian welfare difference. Alternatively, $g(x, y)$ can be seen as the density of the noise distribution, capturing choice inconsistencies and possible survey bias. In particular, it is important to note that this density might be asymmetric. Condition (RUM.2) reflects that the choice between x and y is the same as the choice between y and x, and condition (RUM.3) is a regularity condition stating that the distribution of a pair's utility differences has connected support, i.e. without gaps.

Last, r represents the chronometric function, which maps realized utility differences v into response times $r(|v|)$, such that larger absolute utility differences generate shorter

response times, that is, easier choices are faster. This "chronometric effect" is based on well-established empirical regularities from the cognitive sciences (see Alós-Ferrer et al., 2021, for details and references). Given a RUM-CF $(u, \tilde{v}, r)$ and a pair $(x, y) \in C$, the random variable describing the response times predicted by the model conditional on x being chosen over y is given by $\tilde{t}(x, y) = r(|\tilde{v}(x, y)|)$, conditional on $\tilde{v}(x, y) > 0$.

The results we seek will be in terms of preference revelation for *all* RUM-CFs which rationalize (explain) the data, as captured by the following definition.

Definition 3. A RUM-CF $(u, \tilde{v}, r)$ *rationalizes* an SCF-RT (p, f) if

(i) $p(x, y) = \text{Prob}[\tilde{v}(x, y) > 0]$ holds for all $(x, y) \in D$, and

(ii) $F(x, y)(t) = \text{Prob}[\tilde{t}(x, y) \leq t \mid \tilde{v}(x, y) > 0]$ holds for all $t > 0$ and all $(x, y) \in D$.

In other words, a RUM-CF (the model) rationalizes an SCF-RT (the data) if it reproduces both the choice frequencies and the conditional response time distributions in the latter. Obviously, fixing the set D, every RUM-CF generates an SCF-RT through the equations given in (i) and (ii) above, thus an alternative definition is that a RUM-CF rationalizes an SCF-RT if it coincides with the SCF-RT generated by the former. We say that an SCF-RT is *rationalizable* if there exists a RUM-CF that rationalizes it.[2]

2.2 Revealing Group Preferences

The next definition captures preference revelation in our framework.

Definition 4. A rationalizable SCF-RT *reveals a group preference for x over y* if all RUM-CFs that rationalize it satisfy $u(x) \geq u(y)$. It *reveals a strict group preference for x over y* if all RUM-CFs that rationalize it satisfy $u(x) > u(y)$.

Our results make use of the following technical concept, taken from Alós-Ferrer et al. (2021). Given two cumulative distribution functions G and H on $\mathbb{R}_+$ and a constant $q \geq 1$, we say that G *q-first-order stochastically dominates* H (also written G q-FOSD H) if

$$G(t) \leq q \cdot H(t) \text{ for all } t \geq 0.$$

If the inequality is strict for some t, then G *strictly q-first-order stochastically dominates* H (written G q-SFSD H). For $q = 1$, these concepts coincide with the standard notions of first-order stochastic dominance, but they are weaker when $q > 1$. Clearly, q-FOSD implies q'-FOSD whenever $q \leq q'$.

The following Theorem reformulates the main result of Alós-Ferrer et al. (2021) in welfare-utilitarian terms for a population. This result is obviously limited to preference revelation within a single, given group.

[2]The results in Alós-Ferrer et al. (2021) further study preference revelation within constrained subclasses of RUM-CFs, e.g. the class of all *symmetric* RUM-CFs (meaning that noise terms have symmetric densities). Here we consider only the unrestricted class of all RUM-CFs.

Theorem 1. *A rationalizable SCF-RT (p, f) reveals a group preference for x over y if $F(y,x)$ q-FOSD $F(x,y)$, and reveals a strict group preference if $F(y,x)$ q-SFSD $F(x,y)$, for $q = p(x,y)/p(y,x)$.*

2.3 Comparing Preferences Between Groups

We consider data from two groups, A and B, and two fixed choices, (x,y) and (z,w). However, we explicitly allow $A = B$ (and $(x,y) \neq (z,w)$), in which case the comparison is between two different choices made by the same group. We also explicitly allow $(x,y) = (z,w)$ when $A \neq B$, that is, comparing the same choice across two different groups, and $z = y$ even when $A = B$, that is, comparing two connected choices for the same group, as in staircase designs.

We now adapt all previous concepts to our setting, always taking A, B, and the pairs (x,y) and (z,w) as fixed and given. The data hence must deliver the choice frequencies for x and y (for group A) and for z and w (for group B), and the conditional response time distributions.

Definition 5. A *two-group stochastic choice function with response times* (2SCF-RT) for choices (x,y) and (z,w) is a quadruple $(p_A, f_A; p_B, f_B)$ such that (p_A, f_A) is an SCF-RT for the choice (x,y) and (p_B, f_B) is an SCF-RT for the choice (z,w).

The class of data-generating processes allows for different distributions of utility differences in the two groups, but fixes a chronometric function. Obviously, for applications where $A = B$, the latter assumption is void and thus inconsequential. For $A \neq B$, this assumption means that r is taken to reflect a (possibly neurophysiological) relation which is orthogonal to the nature of the groups (although the assumption could be weakened to some extent). We remind the reader, however, that r is not assumed to have any specific functional form, and that our preference-revelation results below are not only for all utilities and distributions of utilities which rationalize the data, but also for all r.

Definition 6. A *two-group random utility model with a chronometric function* (2RUM-CF) is a tuple $(u_A, \tilde{v}_A; u_B, \tilde{v}_B; r)$ such that $(u_A, \tilde{v}_A, r)$ and $(u_B, \tilde{v}_B; r)$ are RUM-CFs.

Rationalization is extended in a straightforward way.

Definition 7. A 2RUM-CF $(u_A, \tilde{v}_A; u_B, \tilde{v}_B; r)$ *rationalizes* a 2SCF-RT $(p_A, f_A; p_B, f_B)$ for choices (x,y) and (z,w) if $(u_A, \tilde{v}_A, r)$ rationalizes (p_A, f_A) (for (x,y)) and $(u_B, \tilde{v}_B; r)$ rationalizes (p_B, f_B) (for (z,w)).

The last definition captures our key concept, i.e. when group A prefers or likes x over y more than group B prefers z over w.

Definition 8. A rationalizable 2SCF-RT for choices (x,y) and (z,w) *reveals that group A prefers x over y more than group B prefers z over w* if all 2RUM-CFs $(u_A, \tilde{v}_A; u_B, \tilde{v}_B; r)$ that rationalize it satisfy

$$u_A(x) - u_A(y) \geq u_B(z) - u_B(w).$$

It *reveals that group A prefers x over y strictly more than group B prefers z over w* if all 2RUM-CFs that rationalize it fulfill the strict version of the inequality above.

We aim to reveal cardinal preference across groups. A first, trivial result in this direction follows from Theorem 1. Obviously, if group A prefers x over y but group B does *not* (strictly) prefer z over w, i.e. it rather prefers w over z, we can trivially conclude that group A prefers x over y more than group B prefers z over w. This straightforward observation can be translated in terms of revealed preference as follows.

Corollary 1. *If a rationalizable 2SCF-RT $(p_A, f_A; p_B, f_B)$ for choices (x, y) and (z, w) satisfies that*

(a) $F_A(y, x)$ q_A-FOSD $F_A(x, y)$, with $q_A = p_A(x, y)/p_A(y, x)$, and

(b) $F_B(z, w)$ $(1/q_B)$-FOSD $F_B(w, z)$, with $q_B = p_B(z, w)/p_B(w, z)$,

then it reveals that group A prefers x over y more than group B prefers z over w.

To state the following result, define for any SCF-RT (p, f) including data on a choice (a, b),

$$H(a, b)(t) = p(a, b)F(a, b)(t) - p(b, a)F(b, a)(t)$$

and note that the q-FOSD property is equivalent to $H(a, b)(t) \geq 0$ for all t. Given a rationalizable 2SCF-RT $(p_A, f_A; p_B, f_B)$, let $H_A(x, y)$ and $H_B(z, w)$ be defined analogously. Then, Corollary 1 states that a sufficient condition to reveal that group A prefers x over y more than group B prefers z over w is that $H_A(x, y)(t) \geq 0 \geq H_B(z, w)(t)$ for all t. Our main result identifies a weaker sufficient condition, namely that $H_A(x, y)(t) \geq H_B(z, w)(t)$ irrespective of whether the sign of either term is constant or not.

Theorem 2. *If a rationalizable 2SCF-RT $(p_A, f_A; p_B, f_B)$ for choices (x, y) and (z, w) satisfies that*

$$H_A(x, y)(t) \geq H_B(z, w)(t) \quad \text{for all } t > 0,$$

then it reveals that group A prefers x over y more than group B prefers z over w.

If the inequality is strict for some t, the revealed preference is also strict.

The proof is in the Appendix. The formalization in terms of H_A and H_B makes clear how Theorem 2 generalizes Corollary 1. Of course, as the condition becomes more involved, an intuitive interpretation also becomes more difficult. The condition $H(a, b)(t) \geq 0$ for all t can be interpreted as stating that revealed errors (choices of y when, ex post, a preference for x is revealed) are not much faster in a stochastic dominance sense than revealed correct responses, which very roughly could be taken as a weakening of a "slow errors" condition. Suppose, for concreteness, that $H_A(x, y)(t) \geq H_B(z, w)(t) \geq 0$, so that this interpretation holds for group A and group B (and their respective choices) separately. Intuitively, $H_A(x, y)(t) \geq H_B(z, w)(t)$ for all t means that the difference is farther away from zero, thus larger, for group A. That is, the

condition requires that the difference in response times (actually, the difference in the distributions) is larger for group A than for group B. There are, however, two caveats with this interpretation. The first is that the condition is distributional and cannot be summarized in terms of statistics as, e.g., the mean. The second is that the interpretation is only partial, as it does not take into account that the functions H_A, H_B incorporate the actual choice frequencies; that is, the actual condition refers to differences between cdf values adjusted by choice frequencies.

3 Empirical Approach

The method we propose is quite general, as it requires very few assumptions and can be applied to a wide array of datasets. The application is straightforward, as one only needs to check the conditions stated in Theorems 1 or 2 to see whether preferences are revealed. Further, as we will show in the empirical applications below, the method presents a substantial advance over other survey-analysis techniques. By relying both on choice frequencies and distributions of response times, the method is often able to uncover regularities which would remain hidden with standard methods and datasets lacking response-time information.

3.1 Estimation of the H functions

The theoretical conditions in Theorems 1 and 2 require knowledge of the distributions of choices and response times in a survey. Obviously, actual datasets entail finitely many observations, and are affected by measurement errors and other sources of noise beyond the behavioral noise captured in Theorems 1 and 2. Thus, we need to consider the statistical implementation of the theoretical conditions in the previous section.

The theoretical results are expressed in terms of inequalities involving the functions $H_i(\cdot)$, which in turn are defined in terms of the distributions of response times. Thus, the first step is to estimate the distribution of (log-transformed) response times. As in Alós-Ferrer et al. (2021), Alós-Ferrer and Garagnani (2024b), Alós-Ferrer et al. (2024), and Liu and Netzer (2023), for this purpose we use an Epanechnikov kernel with optimally chosen non-adaptive bandwidth. In our empirical applications below, kernel density estimates were performed in *Stata* using the *akdensity* function (Van Kerm, 2012).

In order to provide statistical tests for our preference-revelation conditions, the second step is to obtain 95%-confidence intervals for the estimated H_i functions. In particular, to obtain the confidence intervals of a given H_i, we apply a bootstrapping procedure to the computation of this function. We repeat the computation of each $H_i(x, y)$ a 1000 times while resampling participants (and hence choice frequencies) with replacement. That is, for every resampling iteration in the bootstrapping procedure, we recompute the kernel estimates of the distributions of response times ($F(x, y)(t)$) and the corresponding choice frequencies ($p(x, y)$). This delivers a confidence interval for every $H_i(x, y)(t)$.

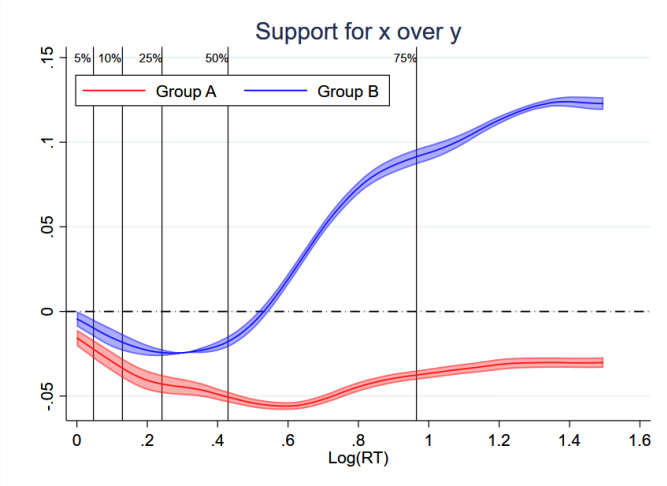

Figure 1: Sample graphical representation of the support for option x compared to y between group A and group B (hypothetical example). If a line does not cross the 0 on the vertical axis, then preferences are revealed (Theorem 1). If the lines representing the two groups do not cross, then a stronger preference for x for one group over the other is revealed (Theorem 2). Confidence intervals provide statistical tests for the statements.

Since there are only finitely many realized values of the response time t in every given dataset, the Epanechnikov kernel uses a specification of bins for the estimation procedure. However, it has been shown that kernel density estimates are asymptotically biased, a problem that can be corrected using an under-smoothing procedure (Horowitz, 2001). We rely on this procedure for our estimation, which in practice means that the width of the assumed bins for the confidence intervals are set to half the width of those used for the estimation of the means.[3] This gives us an unbiased estimation of $H_i(x, y)(t)$ for any sufficiently large empirical sample.

3.2 Statistical tests for Theorems 1 and 2

3.2.1 Graphical Representation

An intuitive way of thinking about the statistical procedures and the conditions in our Theorems 1 and 2 can be conveyed through a graphical representation. Figure 1 provides a (hypothetical) example of the representation we will rely on in the rest of the manuscript. It depicts data from a binary question with two possible responses, x and y. It plots the functions $H_A(x, y)(t)$ and $H_B(x, y)(t)$ for two groups of survey participants, A and B, as well as their confidence intervals derived as explained in Section 3.1. The horizontal scale corresponds to log-transformed response times. The vertical scale corresponds to the values of the H_i functions. Since these functions cannot take values below -1 or above $+1$, this scale is naturally restricted, but we will further restrict it to cover the actual data range for visibility purposes.

[3] In particular, a fixed bandwidth of 0.05 is used to compute the confidence interval, which is half of the one used to estimate the density.

To convey how the data is distributed, in all figures we also add vertical lines showing the percentiles of the (joint) distribution of response times. That is, the vertical line at 5% indicates that only 5% of the data is on the left-hand side of that line, and 95% is on its right-hand side. We provide lines for 5%, 10%, and the quartile boundaries (25%, 50%, and 75%). Our figures hence display the support for an alternative x over another alternative y in terms of the functions H_A and H_B, while allowing the reader to see how the actual responses were distributed. Unless otherwise mentioned, for ease of representation, all figures are also smoothed through a local-averaging procedure.[4]

3.2.2 Statistical Tests for Theorem 1

Theorem 1 tells us how to weight choice frequencies and response times together to verify whether preferences are revealed while accounting for behavioral noise. A group i is revealed to prefer x over y if $H_i(x,y)(t) \geq 0$ for all $t \geq 0$. If we know the true functions $H_i(.)$, or if we treat the estimated functions as if they were the true ones, the condition for preference revelation reduces to checking whether the estimated $H_i(x,y)(t)$ crosses the zero or not. To see this better, our figures always include a horizontal dashed line at the value zero. If it does, preferences are not revealed, and it is not warranted to draw any conclusion from the data. If it does not cross the zero, preferences are revealed, meaning that any model of noise and any assumed utility function fitting the data (within the universe of random utility models as generalized above) will deliver the same preference ordering between x and y. In Figure 1 (ignoring the confidence intervals for now), this means that we obtain preference revelation for Group A, and in particular that y is preferred to x for this group (as the graph of the function is below the horizontal line at zero). However, $H_B(x,y)$ crosses the zero, which means that the condition in Theorem 1 is not fulfilled and preferences are not revealed for Group B.

The functions $H_i(\cdot)$ that we plot, however, are estimated. The figures also include the corresponding confidence intervals (CIs). In agreement with the standard interpretation of CIs, the simplest statistical test of whether the condition in Theorem 1 is fulfilled (whether H_i crosses 0) is to check whether at any point t, the CI of $H_i(x,y)(t)$ includes the 0. If this is the case, the condition is not fulfilled. For example, in Figure 1, preferences for option x are still revealed for Group A according to this statistical criterion, but of course not for Group B. This is a very conservative verification criterion, which corresponds to testing whether $H_i(x,y)(t)$ is statistically different from 0 at all t. In practice, this procedure provides us with a set of p-values, one for each bin of the distribution of response times. We then can conservatively use the largest p-value to determine significance levels for whether a preference is revealed or not.

[4]Specifically, the log-transformed response times whose H values are to be represented are rounded down to five decimals, which graphically merges very-close bins together. This is done only for the graphical representation, and only if the (graphical) conclusions do not change, e.g. whether it becomes less clear that the confidence intervals overlap or not.

3.2.3 Statistical Tests for Theorem 2

Applied to a fixed binary question as above, Theorem 2 spells out conditions guaranteeing that a group is revealed to have a stronger preference for one option over another than a different group. In particular, Group A is revealed to prefer x over y more than Group B if $H_A(x,y)(t) \geq H_B(x,y)(t)$ for all $t \geq 0$ (and conversely for Group B having a stronger preference than Group A). Graphically, this requires that the functions $H_A(x,y)(t)$ and $H_B(x,y)(t)$ do not cross. In Figure 1, we see graphically that H_B is always above H_A. Thus, if we take the plotted functions to be the actual, theoretical ones, we reveal that Group B prefers x over y more strongly than Group A, even though no preference for x over y was revealed for Group B.

To obtain a statistical test, we examine the CIs for both H_A and H_B (see Figure 1) and check whether they intersect each other. That is, we ask whether the CIs of $H_A(x,y)(t)$ and $H_B(x,y)(t)$ overlap at any t. If they do, no difference in the preferences across groups is revealed. If they never overlap for any t, it is revealed that one group's preference is stronger than the other's. That is, any model of noise and any utility function fitting the data, within the universe of random utility models, will deliver the same ordering of utility differences. In Figure 1, we see that Group A has a stronger preference for option x over y than Group B according to this criterion. Again, following this procedure, we obtain a p-value for the test of whether $H_A(x,y)(t)$ and $H_B(x,y)(t)$ are different. Analogously to the case of Theorem 1, we conservatively use the largest p-value across all the bins of the distribution of response times to determine significance levels for group comparisons following Theorem 2.

3.2.4 Data Exclusions

Since response times cannot be negative, $F_i(a,b)(0) = F_i(b,a) = 0$ and $H_i(a,b)(0) = 0$ for any choice pair (a,b). Hence, with response times close enough to zero, the H_A, H_B functions fill be arbitrarily close to the horizontal line at zero and to each other. Empirically, with large enough datasets including fast-enough responses, the confidence intervals typically overlap and contain the value zero. However, very fast responses in surveys can be due to inattention, respondents speeding through the survey, or heuristics unrelated to the question being studied. It is hence reasonable to fix criteria excluding responses which are too fast.

Since questions in our surveys take longer than 1 second to read and answer, we first exclude from all analyses observations faster than 1 second. Thus, the horizontal scale in all figures starts at $\log 1 = 0$. This absolute criterion, however, results in very different numbers of exclusions for different questions. We hence complement it with a relative criterion, excluding the 5% fastest responses for the purposes of statistical tests. In the

figures, this means that the comparisons illustrating the tests correspond to the lines and confidence intervals on the right-hand side of the 5% vertical line.[5]

There is, of course, no strong theoretical basis for these conventions. To validate them, we hence adopt an empirical approach. Our first survey, carried out for the purposes of validating the method, contains questions reflecting well-established phenomena from the literature. Hence, we also use this survey to validate our data-exclusion conventions in the sense that we expect most of the postulated effects to obtain with those exclusions in place.

4 Survey 1: Validation of the Method

We ran two pre-registered, online surveys of the UK population. The purpose of Survey 1 ($N = 1,102$; see Appendix B for the derivation of the required sample sizes) was to validate the method, in preparation for applications studying subjective, complex issues (e.g., political preferences or moral issues; see Survey 2). This survey was pre-registered at the AEA RCT Registry (AEARCTR-0012668) and received ethical approval from the Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich, OEC IRB 2023-110. The survey lasted around five minutes and participants were compensated with a flat fee of GBP 0.5 plus task incentivisation (see details below).

Survey 1 was conceived as a demonstration that the techniques described in the previous sections are able to retrieve preferences using response times in simple, online situations and recover well-known differences. For this purpose, we studied a series of questions for which the results should be clear in advance. This included perceptual questions with objectively-correct answers (e.g., "which line is longer") reflecting well-known effects from the psychological literature, classical questions on well-established framing and related effects (Tversky and Kahneman, 1981) implemented through randomly-assigned groups, and preferential questions based on consistency with previously-elicited preferences (e.g., donations to most-favourite vs. least-favourite charities). The following subsections look at the results for reach kind of questions in the survey.

4.1 Line Lengths: Revealing Objective Preferences

Our approach is based on psychometric and chronometric regularities which are well-established in psychology and neuroscience. The simplest tasks establishing these effects arise from psychophysics, and involve the perceptual discrimination between two objects which differ along an objective scale: which of two lights is brighter, or which of two lines is longer. The psychophysics literature has established that error rates are larger and responses are slower when stimuli are more similar in the relevant scale (Dashiell, 1937;

[5]Specifically, we first exclude observations with response times faster than 1s, then compute the H functions, apply the bootsptrapping procedure, and generate the pictures, and finally exclude he fastest 5% of observations for the purposes of statistical testing.

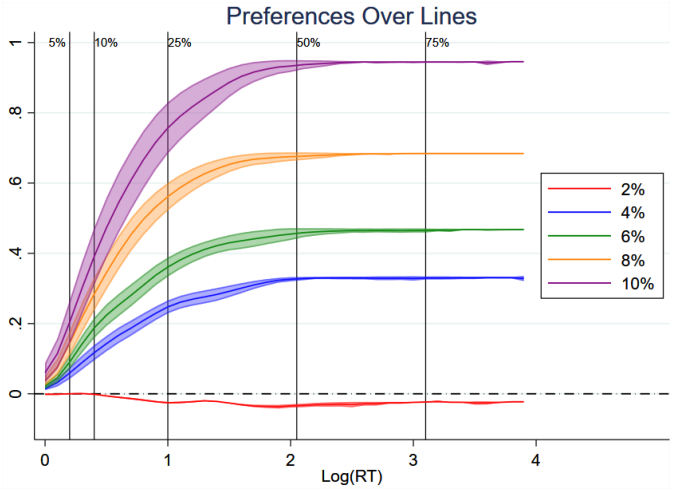

Figure 2: Revealed preferences over line lengths.

Moyer and Landauer, 1967; Laming, 1968, 1985; Klein, 2001; Wichmann and Hill, 2001). On the other hand, the literature on perception considers a *just noticeable difference* to capture the fact that, below a certain threshold, differences between two stimuli might not be detectable. Weber-Fechner's Law (Fechner, 1860) relates the magnitude of the just noticeable difference to the absolute intensity of the stimuli.

In the first part of Survey 1, participants were asked to differentiate the length of different lines on-screen. This task is representative of psychophysical discrimination tasks as above, but can be easily carried out online. Further, the task has been recently validated using monetary incentives (Duffy and Smith, 2025).

This is the first and simplest test of the validity of our method. It allows us to validate whether the proposed method is able to retrieve objective preferences using response times in online surveys. That is, as long as the lines are different enough, application of Theorem 1 should reveal "preferences" in the sense that the population of participants differentiates the longer line from the shorter one.

Specifically, the survey included 10 different questions where participants saw two lines of different lengths, one above the other, and were asked to identify the longest one (participants needed to click on the longest line). For each participant, one of these questions was randomly selected and paid at the end of the survey. If the participant gave the correct answer, he or she earned 1 GBP.

All lines were between 45 and 50 pt in length (of course we have no control on their actual length in the survey participants' devices, only on their relative size; however, the psychophysical relations expressed e.g. in Weber-Fechner's law refer to the *relative* difference between the stimuli). Figure 2 presents the data for five of the questions (see Figure C.1 for the other five), in the format explained in Section 3.2.1 above. For these five questions, the difference between lines was 2%, 4%, 6%, 8%, and 10% of the longest line in the pair, respectively (e.g., the 10% pair comprised lines of 45 and 50 pt).

Descriptively, the line comparisons clearly illustrate the psychometric regularity that stimuli are more clearly differentiated when they are further apart in the relevant scale, and how our method reflects this regularity. The $H(x, y)$ functions and their confidence intervals shift up progressively as the objective difference between the lines goes up. We can also observe the presence of a just noticeable difference. When the difference between the lines is only 2%, the percentage of correct responses is 48.82%, i.e. choices are in practice random. In contrast, when the difference is larger than 2%, error rates sharply decrease (66.61%, 74.05%, 84.21%, and 97.28% for differences of 4, 6, 8, and 10%, respectively). Error rates are also smaller when the stimuli difference is larger, reflecting standard psychometric effects.

As Figure 2 shows, application of Theorem 1 leads to clear preference revelation for all but the 2% comparisons. Specifically, after applying the exclusions discussed above (and, in particular, the 5% fastest participants), the H function for the 2% comparison is not significantly away from the zero ($p = 0.499$), while those for the 4, 6, 8, and 10% differences are very clearly above the horizontal zero line, as are their confidence intervals ($p < 0.001$ in all four cases).

Application of Theorem 2 further reveals stronger preferences for objective comparisons between lines for almost most comparisons. For example, the line for the 4% comparison is below the lines for the 8 and 10% comparisons, as is the case for the confidence intervals, but not for the 6% comparison. Application of the procedure described in Section 3.2.3 reveals significant differences for the first two comparisons ($p < 0.001$), but not for the latter ($p = 0.407$). In fact, we reveal significantly stronger preferences for almost all comparisons across line pairs. Specifically, the 4, 6, 8, and 10% comparisons reveal stronger preferences than for the 2% comparison (all 4 tests $p < 0.001$), the 8 and 10% comparisons reveal stronger preferences than for the 4% comparison ($p < 0.001$), and the 10% comparisons reveals stronger preferences than for the 6% comparison ($p < 0.001$). The only exceptions are the nearest comparisons, i.e., between the 8% pair and the 10% pair ($p = 0.301$), between the 6% pair and the 8% pair ($p = 0.211$), between the 4% pair and the 6% pair ($p = 0.407$).

In conclusion, the comparison of lines illustrates that the method recovers elementary psychophysical regularities. That is, it recovers stronger (objective) "preferences" for larger differences between the presented stimuli when we measure response times.

4.2 Framing: Revealing the Strength of Subjective Preferences

To further validate our method, we turn to classical manipulations which are known to shift choices, and hence the subjective preferences which they reveal, even under random assignment. Specifically, we presented participants with two classical framing questions taken from Tversky and Kahneman (1981) (see also Tversky and Kahneman, 1986). In each case, participants were randomly assigned to two groups, and each group was exposed to a different version (frame) of the same question. Even though the two frames

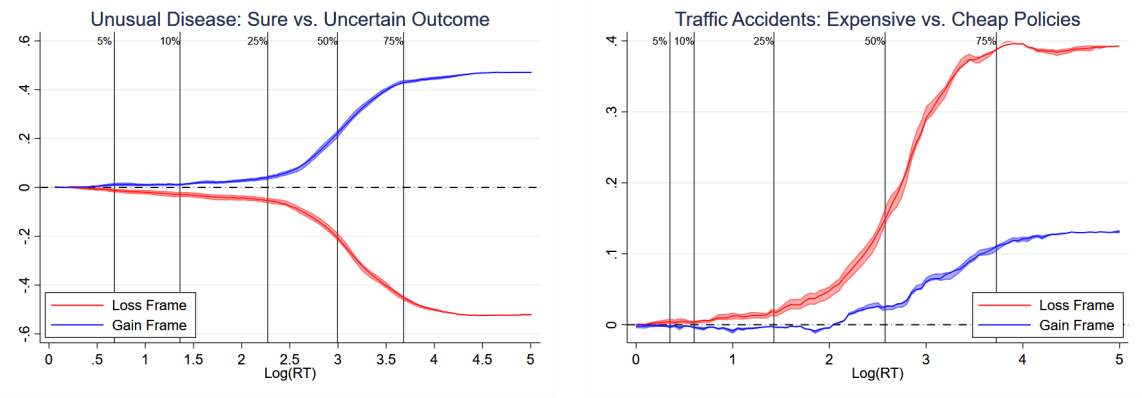

Figure 3: Revealed preferences for framing questions taken from Tversky and Kahneman (1981). Left panel: unusual disease. Right panel: policies to reduce traffic accidents.

of each question describe exactly the same decision situation, and hence there should not be differences in choice frequencies between frames for a population of rational decision makers, it is a widely-replicated observation that frames actually affect choices.

Our first framing question was the classical "unusual disease" question, where participants are asked to choose between policies to fight a hypothetical disease. One policy results in a sure outcome (a one-third reduction in casualties), and the other has probabilistic outcomes (a one-third possibility of full success with no casualties, but the policy is fully-ineffective with the remaining probability). In the loss frame, consequences are described in terms of deaths, and people tend to choose the policy with uncertain outcomes. In the gain frame, consequences are described in terms of lives saved, and people tend to choose the policy with sure outcomes.

The left panel of Figure 3 illustrates the result for this question. Application of Theorem 1 clearly reveals a distinct and opposite preference in the two framing conditions ($p < 0.001$ for both groups), with the group assigned to the loss frame (resp. gain frame) revealing a group preference for the uncertain outcome (resp. sure outcome). Application of Theorem 2 of course also reveals that people assigned to the loss frame have a stronger preference for the uncertain-outcome policy than people assigned to the gain frame ($p < 0.001$), but the result is actually a consequence of Corollary 1. This analysis serves the purpose to validate our method using a well-known manipulation. Of course, the result coincides with the conclusion of a test of proportions (uncertain-outcome policy chosen by 76.45% in the loss frame, vs. 26.18% in the gain frame; $z = 16.515$, $p < 0.001$).

The second framing question was essentially taken from Tversky et al. (1988). It describes policies to reduce casualties in traffic accidents. One policy results in less casualties than the other, but it is also more expensive. In the loss frame, consequences are described in terms of casualties, and people tend to choose the more-expensive policy.[6]

[6]Tversky et al. (1988) used this question to illustrate a preference-elicitation anomaly, confronting direct choice in the loss frame with a pricing procedure where participants matched costs to reveal indifference between the programs.

In the gain frame, in analogy to the first framing question, consequences are described in terms of lives saved. As expected, choice data reveals a framing effect, with 69.64% of participants choosing the expensive option in the loss frame vs. only 56.52% in the gain frame (test of proportions, $z = 4.5103, p < 0.001$).

The right panel of Figure 3 illustrates the result for this question. Application of Theorem 1 clearly reveals a group preference for the expensive policy in the loss-frame group, as in Tversky et al. (1988) ($p < 0.001$). The method, however, does not reveal a preference either way for the gain-frame group ($p = 0.471$), which is unsurprising given the choice frequencies above. However, application of Theorem 2 reveals that the loss-frame group has a stronger preference for the expensive option than the gain-frame group ($p < 0.001$). This example shows that the method recovers a framing effect even if the latter does not shift the preference of a group, but merely reduces its intensity.

As an illustration of the method, Survey 1 also included two further questions taken from Tversky and Kahneman (1981), related to sunk-cost effects (the "concert ticket question") and reference points (a fixed monetary discount from either a large or a low price). Results are discussed in Appendix C.3.

4.3 Charities and Organizations: Revealing Group Preferences

The final validation of our method in Survey 1 focused on donations to charities and political organizations. Specifically, we aimed to show that the method would be able to recover the preferences of a self-identified group from their incentivized donation decisions. For example, the method should reveal a stronger preference to donate to the charity that people liked best (compared to the one they liked worst), to the main political party they supported (compared to the other party from the two main ones in the UK), or to a dog or cat charity depending on personal pet preferences.

In a block of questions within the survey, participants were asked to rate 18 charities and two political parties according to how much they favored them, using 9-point Likert scales. We selected charities using the popularity ranking of yougov.uk for Q3-2023, to guarantee that our list would include both popular and unpopular ones. Popular charities (with approval rates in the 73%–86% range) included the British Heart Foundation, Macmillan Cancer Support, the World Wildlife Fund, Guide Dogs, and others. Less popular charities (with approval rates in the 24%–60% range) included UNICEF, the Fairtrade Foundation, Amnesty International, Cats Protection, and the Vegetarian Society, among others. The list also included the two main political UK parties, the Conservative Party ("Tories") and the Labour Party. Four additional questions in this block, after the ratings, asked for the most- and least-favorite organizations from the list, for whether the participant leaned to the left or the right politically, and whether the participant identified more as a cat or a dog person.

In a different block of questions within the survey, either before or after the ranking block described above (this order was randomized), participants were asked for their

willingness to make monetary donations to each of the 20 organizations. Specifically, for each organization, they were asked a binary question (for which responses times were measured), namely whether they would prefer keeping 4 GBP for themselves or splitting them with the organization (that is, donating 2 GBP and keeping 2 GBP for themselves). At the end of the survey, one of the 20 donation questions was randomly selected and truthfully implemented.

Our aim was to use the binary donation questions (and the response times in those) to recover group preferences. For example, people should be more willing to donate to their favourite charity than to their least-liked charity. This is an application of Theorem 2 when both groups are the same ($A = B$) but the questions are different (donate or not to your favourite charity vs. donate or not to your least-preferred charity). The results are illustrated in the top-left panel of Figure 4. Applying Theorem 1 we reveal a clear preference in favor of donating to the most-liked charity ($p < 0.001$; 81.12% decided to donate) and a clear preference against donating to the least-liked charity ($p < 0.001$; only 4.63% decided to donate). More interestingly, application of Theorem 2 confirms that our participants, considered as a single group, have a stronger preference toward donating to their (respective) favourite charities compared to donating to their least-liked, respective charities ($p < 0.001$). Of course, given that opposite preferences are revealed for both charities, this is also an immediate application of Corollary 1. The result is in agreement with a test of proportions ($z = 36.283$, $p < 0.001$); however, the point of this exercise is the validation of the method.

The top-right panel of Figure 4 depicts the results for the donation questions to the "own party," that is, it depicts the H function for donations to the Labour (resp. Conservative) party for people who declared to lean toward the left (resp. right) of the political spectrum. Application of Theorem 1 reveals that both groups prefer *not* to donate to their party (both $p < 0.001$). Application of Theorem 2 (with different groups), however, reveals that labourists have a stronger preference to donate to their own party than conservatives ($p = 0.042$). Again, this is in agreement with a test of proportions (15.64% of conservatives donate, vs. 21.24% of labourists; $z = -2.199$, $p = 0.028$).

The two bottom panels of Figure 4 complete the validation and recovery exercise for political parties. Analogously to the comparison of most- vs. least-liked charities, the two panels show that labourists have a stronger preference for donating to their own party than to the other party ($p < 0.001$), even though no preference for actually donating to their own party is revealed. For conservatives, however, the test is not significant ($p = 0.458$), and hence we do not reveal a stronger preference for donating to the own party. This is in contrast a test of proportions (conservatives' donation to own 15.64% vs. other 10.89% party, $z = 1.873$, $p = 0.030$). These comparisons also illustrate applications of Theorem 2 for a fixed group (conservatives and labourists, respectively. For conservatives, this is the first case we report where the proposed approach does not align with standard statistical tests, a case we will discuss more extensively when analyzing the data from Survey 2. We find similar results for donations to charities for

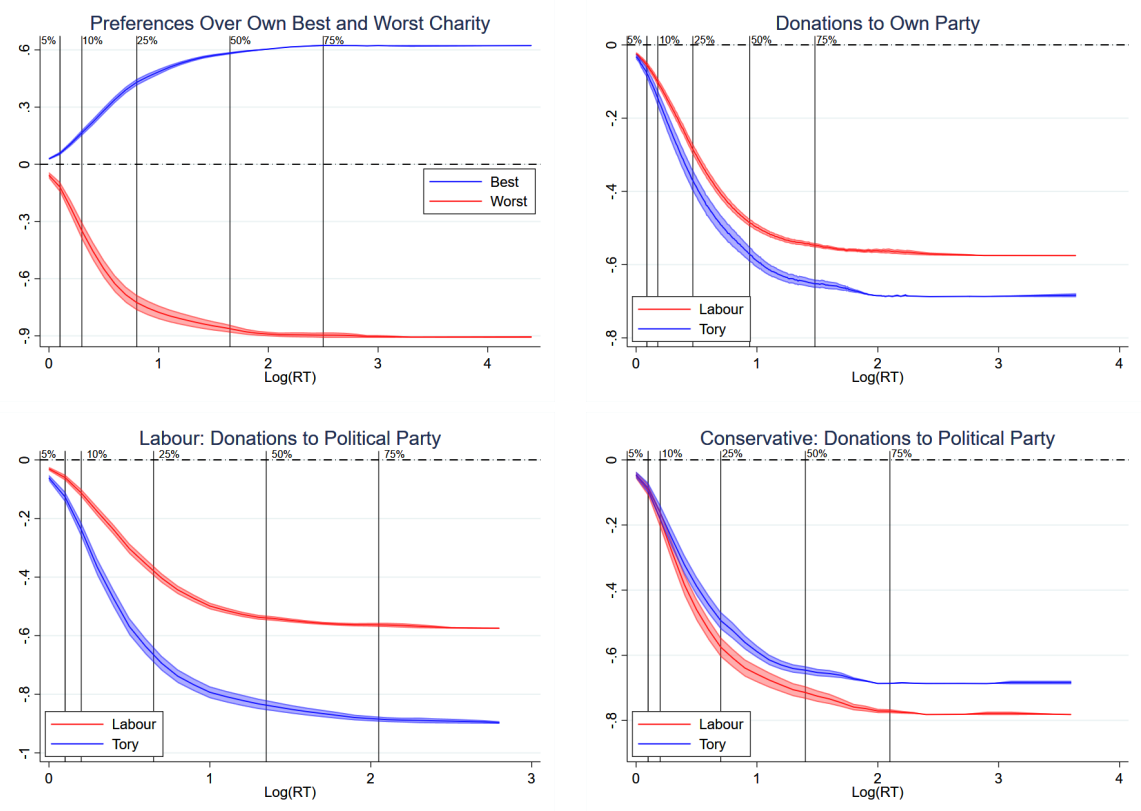

Figure 4: Donation questions (donate vs. keep). Top-Left: Preferences for donating to the own most-liked charity vs. the own least-liked charity. Top-right: comparison of the preference to donate to the own party for conservatives and labourists, respectively. Bottom panels: Preferences for donating to the own party vs. the other party, for labourists (left) and conservatives (right).

cats or dogs, conditional on participants self-declaring as cat or dog people, respectively. The results are reported in Appendix C.3.3.

For each of the 20 organizations, we also divided the sample of participants using a median split of the rating distribution and applied Theorem 1 separately to people above and below the median to reveal their preferences in favor or against donating for each of the organizations. We then applied Theorem 2 to reveal the relative strength of preference between the two groups. Appendix C.3.4 reports the results. We detected a significant difference in 13 of the 20 cases (all $p < 0.001$). In all 13 instances, people rating the corresponding organization above the median were revealed to have a stronger preference to donate to that organization than people rating it below the median.

5 Survey 2: Analysis of Political Preferences

We ran a second online survey of the UK population ($N = 1,202$; see Appendix B for the derivation of the required sample size). As the first one, this survey was pre-registered at the AEA RCT Registry (AEARCTR-0009022) and received ethical approval from

the Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich, OEC IRB 2022-012. Survey 2 involved a representative sample in terms of gender, age, and ethnicity. It lasted around five minutes and participants were compensated with a flat fee of GBP 0.5.

The purpose of Survey 2 was to apply the technique, as validated in Survey 1, to a wide array of questions of independent interest. For this purpose, participants provided binary answers (for which response times were recorded) to 22 questions on a variety of socio-economic topics, e.g. support for gender quotas, redistributive policies, mandatory vaccinations, and greenhouse gas policies, often taken from or inspired by the respective literature. The topics of the implemented questions are not central to the scope of the article, but their wide range and the relevance of the issues is used to showcase the applicability and implications of the proposed method.

To apply our method, for each question we divide participants into two groups according to a dimension relevant to the question, e.g. gender, age, political orientation (left vs. right), or income. The specific comparison to be used for each question was pre-registered. The complete list of questions and chosen group divisions is reported in Appendix D.1. We then apply Theorem 2 to examine whether a stronger preference for an option in one group compared to the other group is revealed.

We also apply Theorem 1 to examine when preferences for one option against another are revealed for each of the two groups, and when they are not. This allows us to show that, even in cases where preferences are not revealed for one of the groups, Theorem 2 can deliver a ranking of preferences between the groups. That is, there are situations where one cannot conclude that a group prefers an option over the other, but we can still say that a group prefers an option more than the other group.

In the following subsections, we showcase the most interesting empirical results from Survey 2. All other results are in Appendix D.2. For comparison and illustration purposes, we also conduct simple tests of proportions comparing the number of people supporting one of the two options across the groups. Our method relies on a nonparametric preference revelation technique using both choice frequencies and response times, while tests of proportions rely exclusively on choice frequencies. Hence, there is *a priori* no theoretical implication linking the significance of the two tests, and four different cases can occur. We structure the rest of the section according to these four cases.

Our method leads to different results than traditional approaches (test of proportions) for 9 of the 22 questions (40.91%). For 2 of the questions (9.09%), we obtain preference revelation across groups even though the corresponding test of proportions is not significant, i.e. we "get more from the data." This is particularly interesting since we stacked the odds in favor of simple tests of proportions finding significant differences, and hence against our method providing new evidence, by choosing and pre-registereing group dimensions that were likely to lead to polarized opinions.

Our results also differ from traditional approaches for 7 further questions (31.82%) where tests of proportions are significant but the conclusion might be unwarranted, as

there is no preference revelation across groups. These cases might flag potential survey bias. In the remaining cases, our method agrees with tests of proportions (positive concordance, 8 questions, 36.36%; negative concordance, 5 questions, 22.73%).

5.1 Getting More from the Data

Since Theorem 2 uses response times in addition to choice frequencies, it is intuitive that in some cases our technique might be able to "get more from the data," i.e. obtain significant results in cases where a test of proportions is not significant. Among our 22 pre-registered questions, two comparisons fall within this category.[7] That is, we find 9.09% of cases where a researcher might be misled to conclude that there is no difference by a non-significant test result, while our method reveals a preference difference.

One possible explanation for this difference might be that the preferences of two different groups are systematically different but relatively close. Hence, when choices are noisy it is likely that statistical tests accept the null hypothesis, delivering a *false negative* result, while our method (that uses more data than tests based on choice frequencies only) provides new information. That is, by using an additional source of information (response times), the analyst effectively reduces measurement errors and is able to identify a difference between the two groups.

However, for $N = 1,202$ (our survey size) and a significance level of $\alpha = 0.05$, with a small effect size (Cohen's $d \leq 0.2$), the expected proportion of false negatives for a test of proportions is below 0.1%. Thus, a more likely explanation is that the differences are not only due to the likelihood of false negatives. Rather, they might signal that the preferences of the groups are truly different, but many individuals of one group might have misrepresented their choices in the survey, possibly due to social desirability bias.

To better gauge these potential explanations, we now look at the two questions where discrepancies are found in more detail. The left-hand panel of Figure 5 depicts the data for a question on hypothetical redistributive policies, where the groups are defined according to a median-split on personal income ("poor" vs. ' rich"). Specifically, participants were asked whether they would support a redistributive policy which would increase their personal income by GBP 2,500, but decrease the average yearly income in their country by GBP 5,000.

This is, of course, a loaded question, as saying "yes" can be easily perceived to reveal a selfish disposition. Unsurprisingly, application of Theorem 1 reveals a preference against this policy both for the poor and for the rich (that is, the H functions and the confidence intervals are bounded away from and below zero; both $p < 0.001$). However, a test of proportions reveals no significant differences between the rich and the poor (Rich: 32.80% vs. Poor: 28.92%, $z = 1.356$, $p = 0.175$). That is, an analyst using choice data only would conclude that there is no difference in preferences across economic groups.

[7] Of course, we can find many other cases where we get more from the data using our method if we explore other, reasonable (but not pre-registered), combinations of groups and questions.

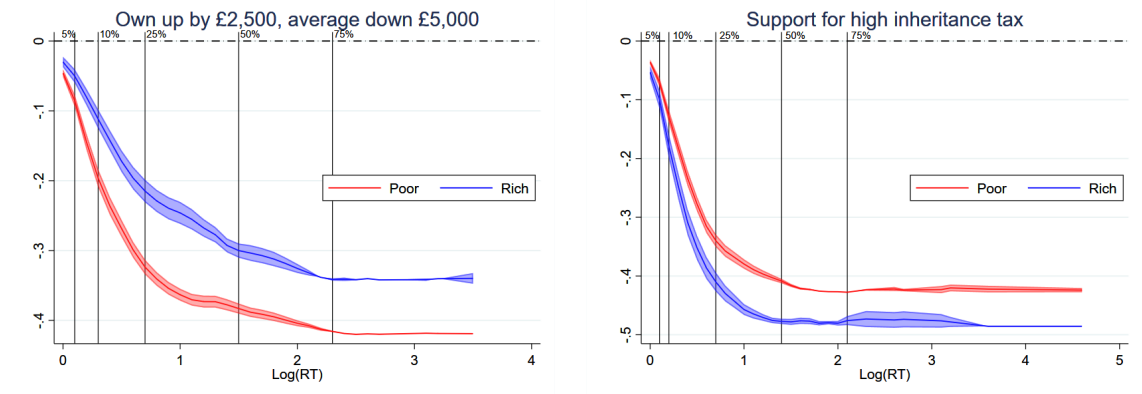

Figure 5: Survey 2. Left: Support for a policy which increases your personal income by GBP 2,500, but decreases the average income by GBP 5,000. Right: Support for high inheritance taxes. H functions plotted against log-transformed response times.

This conclusion is incorrect. As Figure 5(left) shows, H functions and their confidence intervals are clearly separated, with the one for the rich above the one for the poor. Thus, Theorem 2 reveals a stronger preference for the policy for the rich than for to the poor, and the effect is significant ($p = 0.024$). This means that, although neither group supports this policy, rich people do support it significantly more than poor people.

This discrepancy illustrates how our method might uncover differences in preference strength across groups when traditional analyses (based on choice frequencies only) fail to. Although we cannot discard the hypothesis that, in this case, this is due to a false negative in the traditional test, the content of the question suggests that this might be due to richer people mis-representing their choices to state (their perception of) a more socially-desirable response.

The right-hand panel of Figure 5 depicts a second question where the same discrepancy occurs, and our method allows the analyst to get more from the data. It illustrates the support for "high inheritance taxes," again split between poor and rich. Application of Theorem 1 reveals that both poor and rich participants have a preference against this policy (both $p < 0.001$). However, as in the previous example, a test of proportions fails to reveal any difference between the preferences of the groups (Rich: 25.81% vs. Poor: 28.67%, $z = -1.026$, $p = 0.305$). In contrast, application of Theorem 2 reveals a significant difference in preference strength between the two groups ($p = 0.025$). Specifically, as the figure illustrates, the rich oppose high inheritance taxes more than the poor. Again, the question is loaded, since favoring high-inheritance taxes can be perceived as pro-social and social desirability bias might have led some individuals above the median income to misrepresent their choices.

5.2 Unwarranted Conclusions and Possible Survey Bias

Among our 22 pre-registered questions, in 7 cases (31.82%) we found that a test of proportions would detect a significant group difference, but according to our method,

this is an unwarranted conclusion. These cases should be interpreted with care. Since Theorem 2 identifies a *sufficient condition* for preference revelation, it might simply be that preferences are indeed different across groups, but the condition is not fulfilled. On the other hand, and more worryingly, discrepancies in this case might signal that group preferences are not actually different, but the choices of one group are affected by social desirability bias, and hence the test of proportions simply reflects survey bias. If this is the case, a researcher might incorrectly conclude that there is a difference in preferences between the groups, when the data does not actually substantiate the claim.

5.2.1 Two Types of Discrepancies

To better gauge these discrepancies, we briefly discuss four of the cases here. The remaining three cases are presented in Appendix D.2.1. The top-left panel of Figure 6 depicts the support for the opinion that "women should earn less than men for the same job," split by gender. As the figure illustrates, application of Theorem 1 reveals that, unsurprisingly, both females and males have a preference against this opinion (i.e., the H functions do not cross zero, $p < 0.001$ for both). Understandably, the proportions of people who stated a support for such an opinion are rather low for both genders, but a test of proportions suggests a statistically significant result where the statement finds more support among males (Female: 0.65% vs. Male: 2.04%, $z = -2.108$, $p = 0.035$).

This (controversial) finding is unwarranted. As Figure 6(left) shows, the two H curves and their confidence intervals overlap across the entire range, and thus Theorem 2 does not reveal any group preference differences ($p = 0.500$). Given the content of the question, it is natural to speculate that social desirability bias might affect females more than males for this particular case, and this (and not an actual preference difference) is the difference captured by the test of proportions. Relying on Theorem 2 instead, an analyst is not entitled to conclude that people support this policy more depending on their gender. Thus, in cases as this, our method is an additional tool that flags the possibility that the statistical significance of simpler tests might be an artifact of survey bias, preventing the researcher from arriving at unwarranted conclusions.

The top-right panel of Figure 6 depicts another example of this phenomenon. It represents the support for the statement that the government should impose taxes on industry to discourage practices that contribute to global warming. This statement was taken from Bechtel et al. (2020), and we pre-registered an analysis differentiating groups according to a median-split of age (young vs. old; the median age was 45 years). Application of Theorem 1 reveals a preference in favor of such global-warming taxes for both groups (i.e., both H functions are above the zero, both $p < 0.001$). As in the last example, however, a test of proportions suggests a difference across the groups, namely that the young support such taxes more strongly than the old (Old: 83.65% vs. Young: 88.73%, $z = -2.557$, $p = 0.011$). This finding, however, is unwarranted according to Theorem 2, which reveals no difference in the strength of preference of the two groups

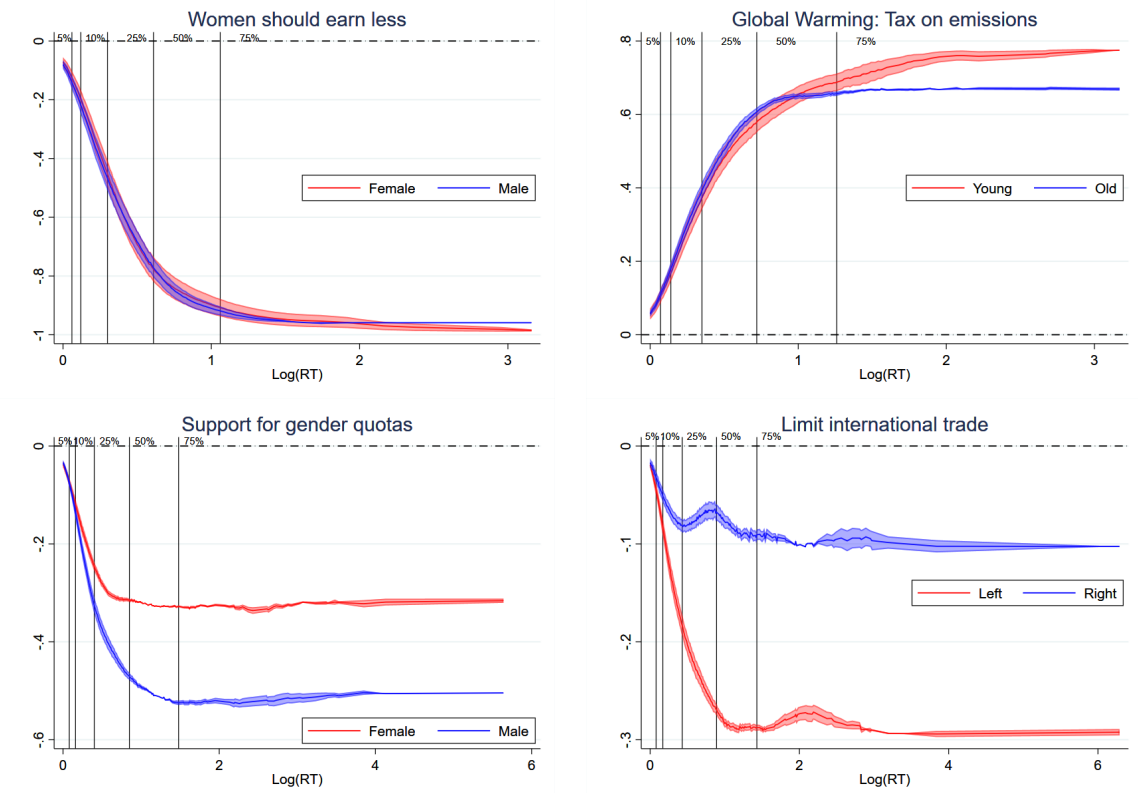

Figure 6: Survey 2. Four questions where an apparent difference in group preferences according to a significant test of proportions is flagged as possibly due to survey bias by Theorem 2. Top-left: support for the statement that women should earn less than men for the same job. Top-right: support for industry taxes to combat global warming. Bottom-left: support for gender quotas. Bottom-right: support for limiting international trade to protect national jobs.

($p = 0.467$; the two H functions intersect even after exclusion of the fastest 5% of responses). This can also be seen very clearly in the top-right panel of Figure 6, which shows how the confidence intervals overlap for most of the range.

The bottom-left panel of Figure 6 represents the support for gender quotas, conditional on the participants' gender. Again, application of Theorem 1 reveals preferences for both groups, namely that both males and females oppose gender quotas (i.e., both H functions are below the zero, both $p < 0.001$). As in the previous examples, a test of proportions suggests a group difference, in this case that males might oppose gender quotas more than females (Females: 36.15% in favor vs. Male: 24.70%, $z = 3.588$, $p < 0.001$). However, once again this (provocative) conclusion is unwarranted according to Theorem 2, which reveals no differences between the group preferences ($p = 0.500$). Examination of Figure 6, though, also suggests that the discrepancy might be qualitatively different in this case compared to the two previous ones, as the confidence intervals are clearly separated for most of the range, and only overlap for the fastest decisions.

This is very similar to the case depicted in the bottom-right of Figure 6, which represents the support for the politically-controversial statement that countries should

limit international trade to protect national jobs (taken from Sides and Citrin, 2007). Application of Theorem 1 reveals that both people on the right and on the left of the UK political spectrum *oppose* limiting international trade ($p < 0.001$ for both). A test of proportions again suggests a statistically highly significant result where limiting international trade finds more support among conservatives (Right: 44.88% vs. Left: 35.37%, $z = 3.274$, $p < 0.001$). As in the previous examples, application of Theorem 2 again suggests that this conclusion might be unwarranted, because the difference in group preferences is not significant ($p = 0.151$). As in the last example, and in contrast to the first two, Figure 6 shows that the confidence intervals only intersect for the fastest decisions, and are clearly separated for most of the range.

5.2.2 Quantifying Survey Bias (Floodlight Analysis)

The last four examples strongly suggest that discrepancies where traditional analyses find an apparent result but our method flags it as unwarranted might often be due to the presence of social desirability bias. Obviously, there is a strong social pressure in favor of closing the gender gap, creating a social desirability bias which might be even stronger in the case of women, creating a difference in choice frequencies which does not reflect genuine group preferences. Since gender quotas are often seen as a way to fight gender inequality the same argument applies to this question. Similarly, in the recent decades a strong social pressure to fight climate change has permeated society, and is often viewed as particularly supported by younger generations. Again, this might create a social desirability bias where younger people perceive a higher pressure to express socially-acceptable opinions on fighting global warming. In the last case, one could speculate that it is expected from conservatives to put a focus on protecting national jobs, possibly creating another bias.

Clearly, the strength of these intuitive arguments varies across questions. If discrepancies between our method and traditional methods might flag social desirability biases, then it is particularly important to quantify this possible bias. Close examination of Figure 6 suggests a first possibility. The two examples on the top of the figure are graphically very different from the two on the bottom of the figure, but this difference is not captured by our statistical tests. This is because the tests we use are extremely conservative in the sense that they are based on the maximum p-value across the range of the H functions. For the questions on the top of the figure, however, confidence intervals overlap across most of the range. For the ones on the bottom of the figure, they only overlap for fastest responses.

This difference can be quantified following an approach analogous to the *Floodlight Analysis* (Johnson-Neyman method) used commonly in the marketing literature (Spiller et al., 2013; Hayes, 2018, see more details below). For the question on limiting international trade, the political orientation difference according to Theorem 2 is significant (the H functions do not intersect; $p < 0.05$) if we exclude the 5.20% fastest responses

instead of only the 5% fastest. For the gender quotas question, the difference according to Theorem 2 would be significant ($p < 0.05$) if we dropped 11.73% of the fastest observations. However, for the statement that women should earn less, the gender difference according to Theorem 2 would only be significant ($p < 0.05$) if we excluded the 96.75% fastest responses. For the question on taxing industry to reduce global warming, a significant group difference according to Theorem 2 ($p < 0.05$) would only be achieved if we excluded 75.35% of the fastest choices.

Understandably, an applied researcher might be willing to accept excluding the fastest 5.20% responses as partial evidence in support of a political difference for limiting international trade. Some researchers might even be willing to accept excluding 11.73% of the observations as suggestive evidence of a gender difference in the support for gender quotas. However, we doubt that anybody would defend excluding 96.75% or 75.35% of the observations to accept a group difference. Hence, one could add the percentage of fastest observations to be excluded in order to obtain a significant difference as a criterion to gauge the strength of the discrepancy between methods, and in particular the likelihood that the discrepancy reflects an actual social desirability bias.

This proposal is conceptually similar to Floodlight Analysis, which is widely used in the marketing and consumer research literature as a quantification of ranges of significance for statistical tests (e.g., Spiller et al., 2013; Hayes, 2018; see also Alós-Ferrer and Garagnani, 2020 for an application in experimental economics). The idea of Floodlight Analysis is to perform a series of statistical tests for differences between groups (in our case, differences in preferences between groups) along different values that two variables can obtain to find when the two groups start to be significantly different (those are called the Johnson-Neyman points). However, while for Floodlight analysis the difference between groups is assumed to be monotonic (i.e., a significant difference at one point typically implies a significant and larger difference at any later point), in our method one could potentially obtain a preference revelation for an interior segment of the range of the response time distribution (see, for example, the left-hand panel of Figure 8). For this reason (and since biased decisions might in general be associated with fast responses), we propose to use just the percentage of fastest observations to be excluded in order to obtain a significant difference at the 5% level, i.e. the smallest response time such that excluding all faster decisions ensures that the H functions do not cross for the remainder of the range.

5.2.3 Quantifying Survey Bias (Response Swap)

A different approach to quantifying survey bias is to focus on a lower bound for the possible response bias in a given survey where our method and standard statistical techniques differ. Specifically, we propose to use the minimum percentage of independent choice observations which would make the two methods agree. Since in most survey applications the relevant observations are the responses of individual participants, this

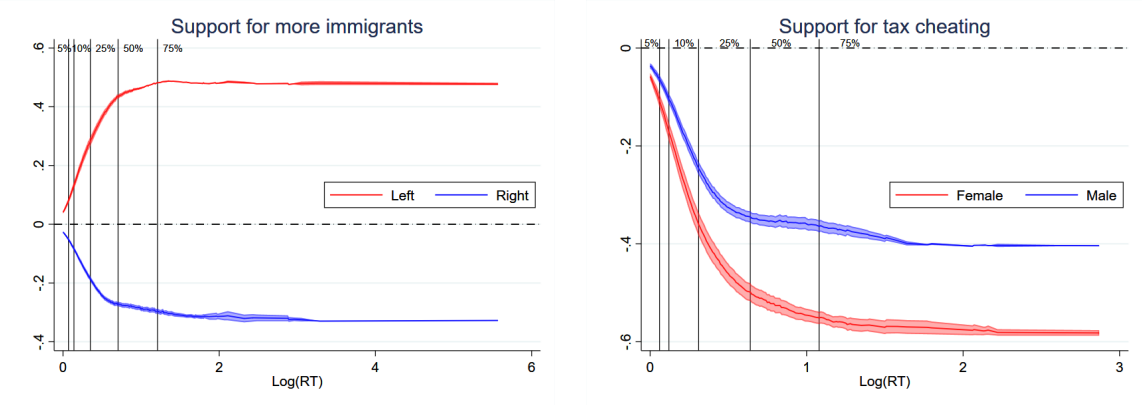

Figure 7: Survey 2: Two examples of positive concordance. Left: Support for immigration comparing political orientations. Right: willingness to cheat on taxes, compared across genders. H functions plotted against log-transformed response times.

is equivalent to the smallest percentage of participants whose answers would need to be changed for a test of proportions to arrive at the same conclusion as our method, e.g. to lose significance. This provides a conservative (lower-bound) estimation on the bias in responses to the question of interest.

For the cases presented in Figure 6 above we obtain a lower bound for the survey bias of 4.26% (gender quotas), 3.82% (limiting international trade), 0.17% (women's earnings), and 1.38% (global-warming taxes). Again, this suggests a difference between the two questions on the top of Figure 6 and the two on the bottom. As the Floodlight analyses in the previous section, this alternative criterion also suggests that the significant tests of proportions for the questions on the top of the figure are particularly likely to reflect survey bias, compared to the other two questions.

The two criteria we propose are aligned (Pearson correlation, $N = 7$, $\rho = -0.866$, $p = 0.012$), with a larger percentage of excluded fast responses or a lower percentage of response swaps flagging possible survey bias when using response frequencies only.

5.3 Positive Concordance

For the remaining 13 of the 22 questions in Survey 2, our method delivers the same message as traditional analyses using only choice frequencies. This can happen because both approaches deliver significant results (positive concordance), or because both tests are non-significant (negative concordance). Positive concordance occurred for eight of the questions in Survey 2. Figure 7 depicts two examples. The remaining six cases of positive concordance are reported in Appendix D.2.2.

The left-hand panel of Figure 7 depicts the support for accepting more immigrants in the U.K. Theorem 1 reveals preferences for people on both sides of the political spectrum (H functions do not cross zero, both $p < 0.001$), and Theorem 2 reveals that people on the left have a stronger preference for supporting additional immigration than those on the right ($p < 0.001$). Since preferences for the groups are revealed in opposite

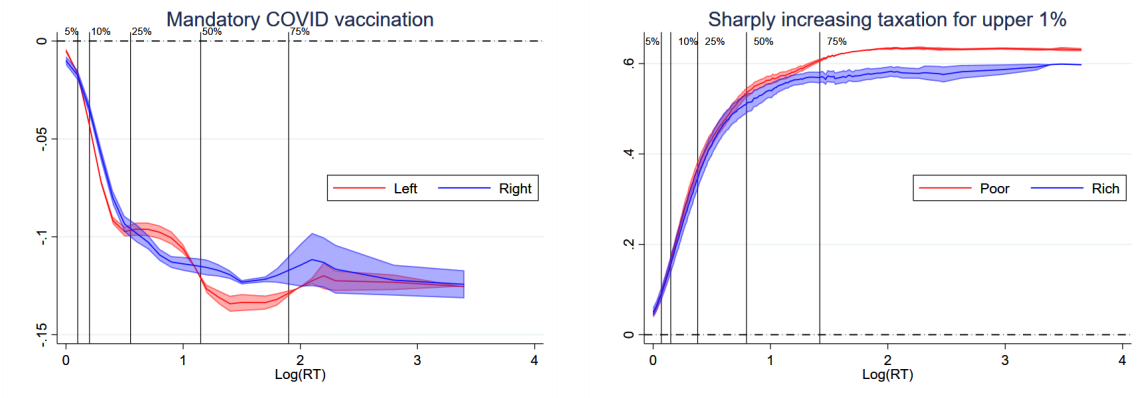

Figure 8: Survey 2: Two examples of negative concordance. Left: Support for a mandatory COVID vaccine. Right: Support for high taxes for the upper 1%. H functions plotted against log-transformed response times.

directions, the group difference also follows from Corollary 1. A test of proportions also detects a statistically significant difference between groups (Left: 73.85% vs. Right: 33.56%, $N = 1202$, $z = 13.692$, $p < 0.001$). Analogous results were obtained for a question on whether Brexit was or not "a good idea" (see Appendix D.2.2).

The right-hand panel of Figure 7 depicts the willingness to cheat on taxes ("Would you cheat on taxes if you had a chance?"). Application of Theorem 1 reveals a preference against cheating on taxes both for females and for males (H functions do not cross zero, both $p < 0.001$). However, 2 reveals a gender difference: males have a weaker preference against cheating on taxes than females ($p < 0.001$). This effect is also captured by a test of proportions (Male: 29.81% vs. Female: 20.65%, test of proportions, $N = 1202$, $z = 3.661$, $p < 0.001$).

5.4 Negative Concordance

Negative concordance occurred for five of the questions in Survey 2. Figure 8 depicts two examples. The remaining three cases are reported in Appendix D.2.3.

The left-hand panel of Figure 8 depicts the support for a mandatory COVID vaccine, splitting the groups on political orientation. Theorem 1 reveals preferences against mandatory vaccines for both groups (H functions do not cross zero, both $p < 0.001$). However, the two H functions intersect multiple times over their range, and hence Theorem 2 does not reveal a difference in preferences across groups ($p = 0.493$). This suggests that the issue of mandatory vaccinations is not related to political positions. A test of proportions is also not able to reject the null hypothesis (Left: 43.75% vs. Right: 43.76%, $N = 1202$, $z = 0.002$, $p = 0.998$).

The right-hand panel of Figure 8 depicts the support for a hypothetical policy that would sharply increase taxation for the upper 1% of the income range, splitting the groups on income (rich vs. poor). Theorem 1 reveals preferences in favor of this policy for both groups (both $p < 0.001$), but Theorem 2 does not reveal a difference in preferences

between the groups, as the confidence intervals overlap for a large part of the range ($p = 0.498$). A test of proportions is also non-significant (Rich: 79.84% vs. Poor: 81.45%, $N = 1202$, $z = -0.656$, $p = 0.512$).

6 Revealing Stochastically-Dominated Preferences

In this section, we discuss an alternative approach to preference revelation across groups in surveys relying on a more stringent criterion and relate it to the recent literature, especially Bond and Lang (2019) and Liu and Netzer (2023). This more stringent criterion for preference revelation requires that the distribution of preference differences in one group first-order stochastically dominates (FOSD) the distribution in the other group.

Requiring that the distribution of one variable for one group FOSD the one for another group is natural for some economic variables, but not for others. The focus on FOSD for group comparisons arises from variables as income. The reason is that if the distribution of income for group A FOSD the one for group B, the average utility of income (or welfare) for group A will be larger than the one of group B for any representative, increasing utility of income (or welfare function). Hence, one can unambiguously conclude that group A is better off than group B without actually knowing the representative utility or welfare function. Of course, income is an observable variable and no revelation technique is necessary. However, the approach has carried over to other unobservable variables, and in particular to the happiness literature. Bond and Lang (2019) argue that, since the cardinal scale of happiness is unclear, an unambiguous ranking of the average happiness of two groups can only be obtained if the happiness distributions are ordered by FOSD. For this reason, in the context of happiness studies, (Liu and Netzer, 2023) proposed to adapt the techniques from Alós-Ferrer et al. (2021) to examine FOSD relations in the distributions of happiness between groups.

For other latent variables, relying on FOSD is far less appealing. For the case of surveys with binary questions that we examine here, the latent variables are utilities, and 2 already reveals group differences for any utility function (and model of noise) which fits the data. There is no requirement to further integrate some additional function over utilities to obtain a statement. Theorem 2 already accomplishes the desired objective, which is an ordinal revelation of differences in group preferences. Thus, there is no theoretical reason to develop a strengthening of Theorem 2 to examine FOSD relations in distributions of utilities.

The only potential reason to favor a stronger criterion based on FOSD is a technical one. Theorem 2 derives a sufficient condition for the revelation of preference differences between groups. In contrast, it is straightforward to obtain a stronger analytical result (Theorem 3 below) for preference revelation based on an FOSD criterion, namely a full characterization in the form of necessary and sufficient conditions.

Empirically, however, this stronger criterion turns out to be less useful. As the application to our two surveys has demonstrated, Theorem 2 is empirically useful, as

the test of the sufficient condition is fulfilled often. However, as we will see, the stronger conditions which characterize the FOSD criterion are far more demanding than the condition in Theorem 2, and as a consequence they are fulfilled empirically less often.

6.1 Stochastic Dominance and Revelation of Group Differences

The following definition spells out the stronger concept based on stochastic dominance.

Definition 9. A rationalizable 2SCF-RT *reveals group A's preferences for x over y stochastically dominate group B's preferences for z over w* if, for each 2RUM-CF $(u_A, \tilde{v}_A; u_B, \tilde{v}_B; r)$ that rationalizes it, $G_A(x, y)$ FOSD $G_B(z, w)$, where $G_A(x, y)$ and $G_B(z, w)$ are the cumulative distribution functions of $\tilde{v}_A$ and $\tilde{v}_B$, respectively.

Of course, if a group difference in preferences is revealed in the FOSD sense, this implies that the means are ordered, $u_A(x) - u_A(y) \geq u_B(z) - u_B(w)$, and thus the revelation of stochastically dominated preferences (Definition 9) implies revealed preference (Definition 8). The following result derives necessary and sufficient conditions for uniform preference revelation.

Theorem 3. *A rationalizable 2SCF-RT $(p_A, f_A; p_B, f_B)$ reveals that group A's preferences for x over y stochastically dominate group B's preferences for z over w if and only if the following two conditions hold.*

(a) $F_A(y, x)$ $p_B(w, z)/p_A(y, x)$-FOSD $F_B(w, z)$, and

(b) $F_B(z, w)$ $p_A(x, y)/p_B(z, w)$-FOSD $F_A(x, y)$.

Further, it is easy to show that if the true data generating process does fulfill that $G_A(x, y)$ FOSD $G_B(z, w)$, then the data *must* fulfill the conditions in Theorem 3.

Corollary 2. *Consider a 2RUM-CF such that $G_A(x, y)$ FOSD $G_B(z, w)$. Then, the generated 2SCF-RT fulfills (a) and (b) in Theorem 3.*

The interpretation of the conditions in Theorem 3 is as follows. Ex post, it is revealed that group A's preferences for x over y stochastically dominate group B's preferences for z over w. Although this does not imply that a preference for x over y is revealed for group A, nor that a preference for z over w is revealed for group B, suppose for concreteness that this is the case. Hence, choices of y or w are revealed errors and choices of x or z are revealed correct answers. Thus, (a) states that revealed errors of group A are not much faster (in the same sense as for Theorems 1 and 2) than revealed errors of group B, and condition (b) states that revealed correct answers of group A are not much faster than revealed correct answers of group A.

6.2 Stochastically Dominated Preferences: Application to Survey 2

We apply the more stringent Theorem 3 to our data from Survey 2. Since the revelation of stochastically dominated preferences implies preference revelation, it follows that the condition of Theorem 3 must fail whenever Theorem 2 failed. Hence, we only need to consider the set of 10 questions where Theorem 2 revealed a preference difference across groups (i.e., either subsection 5.1 "Getting more from the data" or subsection 5.3 "Positive Concordance").

We apply a bootstrapping approach analogous to the ones described in Section 3.2.1. Figure 9 gives a graphical illustration for two examples. Following Theorem 3, we compute the functions

$$p_A(y,x)F_A(y,x) - p_B(w,z)F_B(w,z) \quad \text{and} \quad p_A(x,y)F_A(x,y) - p_B(z,w)F_B(z,w).$$

All our questions are of the form $(x,y) = (z,w) = (\text{Yes}, \text{No})$. Thus, the first function evaluates the response times of "No" answers across the groups, and the second does the same for "Yes" answers. The conditions in Theorem 3 requires the second function to be positive, and the first one to be negative, to reveal that group A's preferences for Yes over No stochastically dominate group B's preferences for Yes over No (or both negative, for the reverse group ordering). Thus, group A supports the statement ("Yes") more than group B in an FOSD sense. A bootstrapping procedure allows us to obtain confidence intervals and p-values testing these statements, as depicted in Figure 9 (since these are two conditions for each application, we report the maximum p-value). Please note that these functions and their interpretations differ from the ones derived from Theorems 1 and 2.

Unfortunately, Theorem 3 reveals itself to be less useful empirically than Theorem 2. From the 10 questions where Theorem 2 revealed a group difference, Theorem 3 fails to detect a significant effect in four occasions. Unfortunately, this includes both examples discussed in Section 5.1, where Theorem 2 revealed differences undetected by traditional tests. Thus, in Survey 2, Theorem 3 does not reveal any qualitative new effects. It also fails to detect significant effects in two cases where, as discussed in Section 5.3, Theorem 2 confirms traditional analyses.

Four examples of the application of Theorem 3 are presented in Figure 9. The remaining questions are reported in Section D.3. The top-left panel illustrates the analysis for the question on whether the U.K. should accept more immigrants (left vs. right). As shown in Figure 7 (Section 5.3), Theorem 2 revealed a group difference. Figure 9 shows that group differences are revealed in the FOSD sense: people on the left of the political spectrum support additional immigration more (again, in the FOSD sense) than those on the right ($p < 0.001$). The top-right panel side of the Figure illustrates the analysis for the hypothetical policy increasing the own income by GBP 2,500 at the cost of a decrease of GBP 5,000 in the average income (poor vs. rich). As shown in Figure 5 (Section 5.1), Theorem 2 revealed a group difference which was not detected by a group

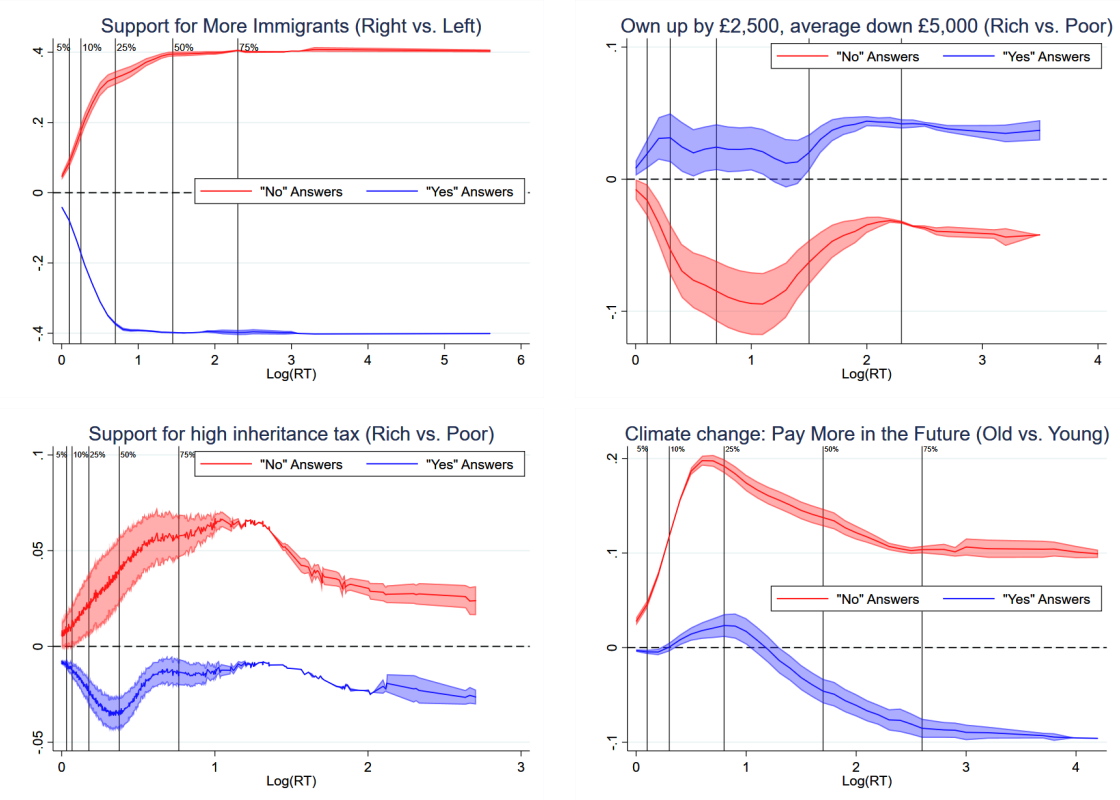

Figure 9: Revealing stochastically dominated preferences through Theorem 3. Top-left: Support for allowing more immigrants between political orientation. Top-right: Support for a policy increasing personal income by GBP 2,500 but decreasing the average income by 5,000. Bottom-left: Support for high inheritance taxes (not smoothed). Bottom-right: support for deferring costs of climate change policies to the future.

of proportions. In contrast, as illustrated in Figure 9, Theorem 3 fails to detect group differences in the FOSD sense ($p = 0.132$). The second case where Theorem 2 showed a group difference undetected by a test of proportions (Figure 5) was the question on the support for high inheritance taxes. As shown in the bottom-left panel of Figure 9, this effect is not detected by Theorem 3 at the conventional 5% significance level (although it would be marginally significant: $p = 0.071$; this figure is not smoothed, as smoothing creates an apparent effect contradicting the p-value). Last, the bottom-right panel of Figure 9 depicts the analysis for the question of whether climate-change policies should be financed by deferring costs to the future. Both Theorem 2 and a test of proportions detected a clear group difference (Figure D.11, Section D.2.2, Online Appendix). However, Theorem 3 fails to detect a significant difference ($p = 0.474$).

6.3 Comparison to Liu and Netzer (2023)

Liu and Netzer (2023) present an application of the techniques introduced in Alós-Ferrer et al. (2021) to happiness surveys, where responses are typically ordered and non-binary (scales from 1 to n). Their work answered a criticism by Bond and Lang (2019), who

pointed out that standard parametric analyses of self-reported happiness scales cannot reveal differences in happiness levels across groups, unless one is sure the distribution of the happiness latent variable for one group first-order stochastically dominates the other. With choice data (without response times), this requires the data to satisfy unrealistic conditions (e.g., with binary answers, everybody in one group to say "unhappy" while everybody in the other group says "happy"). Liu and Netzer (2023) pointed out that, if data includes response times, the revelation of FOSD conditions across groups might be possible. Unfortunately, for three or more categories response times add nothing, and the revelation conditions for intermediate categories remain unrealistic (Liu and Netzer, 2023, Proposition 2). Thus, response times are useful in the context of happiness surveys only if surveys are binary ("happy" vs. "unhappy"). In this context, the main result of Liu and Netzer (2023) (Proposition 2) is equivalent to Theorem 3 above, where $n = 2$ and "group A is detectably rank-order happier than group B" means that group A's preferences for "happy" over "unhappy" stochastically dominate those of group B.

Liu and Netzer (2023) carried out an MTurk survey on happiness and lookd at possible FOSD relations between group preferences. Their empirical approach, however, is diametrically opposed to ours. They point out that the conditions in their Proposition 2, as those of Theorem 1 in Alós-Ferrer et al. (2021) or those in Theorems 1–3 above, are functional conditions which formally parallel an FOSD relation between distribution functions.[8] The statistical literature has developed tests for FOSD relations between distribution functions, going back to McFadden (1989). The tests in this literature, however, formulate the null hypothesis as the existence of an FOSD relation (e.g., Barrett and Donald, 2003, p.74). This means that, if a test is significant, the functional relation is rejected. While this might be appropriate for studying relations between income distributions, it is not appropriate for our purposes. Liu and Netzer (2023) adapt a test from Barrett and Donald (2003), whose null hypothesis is again the existence of an FOSD relation, and apply it to their data, with the objective of testing the functional relations spelled out in Theorem 3. This means that, if their tests are significant, the conclusion is that the sufficient conditions are not fulfilled, and hence no preference revelation (in the dominance sense) can be established. However we are interested in establishing preference revelation. Ideally, data where preference revelation in the dominance sense is possible should result in the conditions in Theorem 3 being fulfilled, which then will lead to the tests of Liu and Netzer (2023) being *not* significant. The resulting empirical approach then would put us in the awkward situation of having to interpret non-significant statistical results. Liu and Netzer (2023, p. 3294) base their empirical conclusions on the "inability to reject the null hypothesis of first-order stochastic dominance," i.e. argue for a positive interpretation of absence of evidence. Further, their overall conclusion is that, if one is willing to accept this interpretation,

[8]The reader should not be confused between this statement and the search for FOSD relations between distribution of preferences of groups. These two links to the mathematical FOSD concept are unrelated.

said inability correlates with the significance of a parametric, ordered-probit regression analysis which does not rely on response times.

In contrast, the null hypothesis of our tests (Sections 3.2.2-3.2.3) is that the functions in the conditions of our theoretical results cross (the zero or each other). Hence, a significant result means that the functions are bounded away from each other, i.e. the conditions for preference revelation (in Theorems 1–3) are fulfilled. Thus, significant results in our statistical approach establish preference revelation, and non-significant results can be properly interpreted as absence of evidence for preference revelation.[9]

In spite of the discussion above, we applied the functional test of Barrett and Donald (2003) used by Liu and Netzer (2023) to the ten questions where our Theorem 2 revealed a group difference in preferences (but not necessarily stochastic dominance). As shown in Section 6.2, Theorem 3 fails to detect a stochastic dominance in four of these questions, including the two examples where Theorem 2 revealed differences undetected by traditional tests. The functional test of Liu and Netzer (2023) is also non-significant for these two questions ($p = 0.124$ and $p = 0.187$ for the questions on the left and the right of Figure 5, respectively). Thus, with the standard interpretation of a non-significance, nothing can be concluded. This is also the case for the question on tax cheating (right panel of Figure 7), where our statistical tests for Theorems 2 and 3 find group differences (even in the dominance sense), but the functional test used in Liu and Netzer (2023) is again non-significant ($p = 0.131$). For the other seven questions, the functional test is significant ($p = 0.036$ for the question on deferring the cost of climate-change policies to the future; $p < 0.001$ for the other questions), which however merely means that the test does not allow to conclude that there is preference revelation.

To better illustrate the difference, we also applied our approach to a preferential question in the data of Liu and Netzer (2023). Their MTurk survey on happiness also included the question "In general, how willing are you to take risks? [Rather willing vs. Rather unwilling]," which is a binary adaptation of a question from the Global Preference Survey introduced by Falk et al. (2018) (Liu and Netzer, 2023, Fig. 6). We re-analyzed eight group differences regarding this question.[10] The results are reported in Table 6.3. For the eight comparisons, tests of proportions suggest significant differences, but the approach of Liu and Netzer (2023) delivers non-significant results in seven occasions, the exception being the comparison between middle and high income. Thus, for that case their approach does not allow for preference revelation, and in all other cases points at absence of evidence. Application of our Theorem 2 and our statistical approach, however, reveals differences in group preferences for six of the eight group comparisons,

[9]The advantage of Theorem 3 over 2 is that the conditions are a characterization. However, this advantage disappears once one considers statistical tests. This is because significance establishes the sufficient conditions for preference revelation, but non-significance merely points at absence of evidence, and not at failure of these conditions.

[10]Liu and Netzer (2023) collected data for 8,000 participants but, due to problems with data collection, their final sample contains $N = 3,743$ subjects for the binary questions, and some of the group classification they use have very small group sizes. In particular, their "no education" group contains only 23 participants, and hence we drop the comparison including that group.

Group	Group 1	Group 2	Test of Prop.	Thm.2	Thm.3	LN23
Female vs. Male	0.590	0.705	0.001	0.004	0.497	> 0.100
High School vs. College Education	0.484	0.682	0.001	0.001	0.494	> 0.100
Don't Have Kids vs. Have Kids	0.506	0.734	0.001	0.001	0.497	> 0.100
Not Married vs. Married	0.514	0.717	0.001	0.001	0.498	> 0.100
Poor vs. Middle Income	0.581	0.670	0.001	0.001	0.499	> 0.100
Middle Income vs. Rich	0.696	0.637	0.001	0.015	0.496	< 0.050
Middle Age vs. Old	0.615	0.490	0.001	0.317	0.499	> 0.100
Young vs. Middle Age	0.685	0.615	0.001	0.116	0.496	> 0.100

Table 1: Application of our statistical test for Theorems 2–3 to the risk question in Liu and Netzer (2023). The second and third columns and the percentages of participants "rather willing" to take risks, and the fourth to last are p-values of a test of proportions, our tests for Theorems 2–3, and the test of Liu and Netzer (2023), respectively.

and is not significant for the remaining two. However, application of Theorem 3 does not allow to reveal stochastic dominance relations between any groups.

For example, the top-left-hand panel of Figure 10 shows that males reveal a stronger preference for taking risks than females (Theorem 2, $p = 0.004$). However, the top-right-hand panel shows that there is no dominance relation between the groups' preference distributions, as the curve for "no" answers (i.e., "rather unwilling") crosses the zero. Thus the test for Theorem 3 is not significant ($p = 0.497$). The test used in Liu and Netzer (2023) is also not significant, but that work argues that the "inability to reject the null hypothesis" should be interpreted as evidence for a dominance relation, which contradicts the figure. The two bottom panels illustrate the results for middle and high-income groups, the only comparison for which Liu and Netzer (2023) report a significant test ($p < 0.05$). Given the null hypothesis of that test, the conclusion is that no preference revelation in the dominance sense can be established from their result. In contrast, the bottom-left panel shows that application of our Theorem 2 reveals a difference in group preferences ($p = 0.015$), and the bottom-right panel shows that our statistical approach to Theorem 2, where the null hypothesis is derived from the natural confidence intervals, does not show evidence of a dominance relation ($p = 0.496$). Figures D.14 and D.15 in the Online Appendix (Section D.4) present the analogous pictures for the other six group comparisons in Table 6.3.

In conclusion, we believe that insisting on revealing stochastic dominance relations between group preferences is inappropriate for applications to survey data. The concept and the associated conditions are unrealistically strong, and hence can seldom be established empirically. Thus, Theorem 2, which reveals group differences in preferences without requiring dominance, is empirically more useful. Further, testing for the conditions in our results or others can be done through a simple and easy-to-interpret statistical approach based on confidence intervals, which leads to tests where rejection of the null hypothesis can be interpreted as establishing preference revelation. This is more useful than alternative approaches based on functional tests where preference revelation must be based on interpreting non-significant results.

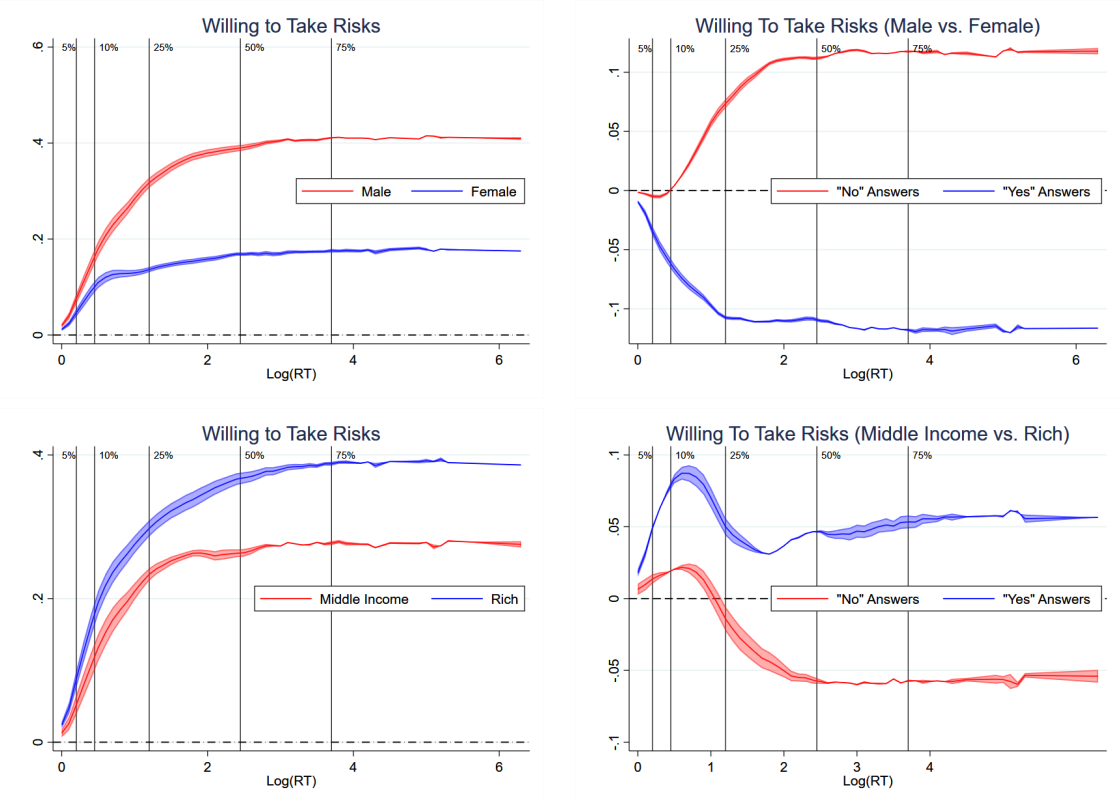

Figure 10: Application of Theorem 2 (left) and Theorem 3 (right) to the willingness to take risks between males and females (top) and between middle and high income (bottom) in Liu and Netzer (2023)'s data.

7 Conclusion

Surveys are an essential instrument to elicit societal preferences in a large variety of economic, political, and social issues, and are also regularly used to uncover differences in preferences across different socieconomic or demographic groups. However, survey data is extremely noisy, and survey bias is ubiquitous, strongly limiting the reliability and usefulness of standard analyses. In this work, we have presented a new way to analyze survey data to actually *reveal* preferences of groups and preference differences across groups. We do so by relying on response times, which are both inexpensive and easily-collected in the digital age. Our results are obtained by incorporating insights from psychology and neuroscience in standard economic models of noisy choice.

By integrating response-time data with choices, we offer a more nuanced understanding of preference structures without requiring extensive additional resources or burdening respondents. We provide ready-to-use techniques which can uncover group preferences and preference differences even when standard statistical tests are inconclusive. The reason, in addition to the fact that our techniques use more data than those tests, is that the question we answer is a different one. In the presence of survey bias, the question that the analyst should answer is not whether a majority of people state that they support a

certain proposition, but rather whether the data (and *all* the dimensions of data) allow to reveal an actual preference (or preference difference) between the alternatives.

In addition to deriving theoretical conditions for preference revelation, we have validated the approach and illustrated its usefulness in a representative, pre-registered surveys using a questions designed for validation (psychophysical tasks, framing, charity donations) and also a large variety of economically-relevant questions, ranging from support for vaccine mandates to redistributive policies, and from the financing of policies reducing greenhouse-gas emissions to whether one would cheat on taxes or on a partner.

The results confirm that the new techniques are useful and often deliver new insights. They are ready for immediate application to standard survey analysis and also a variety of immediate but relevant extensions and applications. For instance, the analysis can potentially identify which group is more receptive to a new product, more likely to endorse a new political candidate, or more likely to support a social change. They can also facilitate the analysis of staircase designs where preferences are elicited through sequences of interrelated questions (e.g. Falk et al., 2018). Overall, the techniques developed in this work have the potential to change how survey data is analyzed in economics, political science, marketing, health research, and many other fields.

References

Alós-Ferrer, C., E. Fehr, and M. Garagnani (2024). Identifying Nontransitive Preferences. Submitted (available at `alosferrer.github.io`).

Alós-Ferrer, C., E. Fehr, and N. Netzer (2021). Time Will Tell: Recovering Preferences when Choices are Noisy. *Journal of Political Economy 129*(6), 1828–1877.

Alós-Ferrer, C. and M. Garagnani (2020). The Cognitive Foundations of Cooperation. *Journal of Economic Behavior and Organization 175*, 71–85.

Alós-Ferrer, C. and M. Garagnani (2022a). Strength of Preference and Decisions Under Risk. *Journal of Risk and Uncertainty 64*(3), 309–329.

Alós-Ferrer, C. and M. Garagnani (2022b). The Gradual Nature of Economic Errors. *Journal of Economic Behavior and Organization 200*, 55–66.

Alós-Ferrer, C. and M. Garagnani (2024a). Common Ratio and Common Consequence Effects Arise from True Preferences. Working Paper.

Alós-Ferrer, C. and M. Garagnani (2024b). Improving Risky-Choice Predictions Using Response Times. *Journal of Political Economy: Microeconomics 2(2)*, 335–354.

Anderson, S. P., J.-F. Thisse, and A. De Palma (1992). *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press.

Apesteguia, J. and M. A. Ballester (2023). The Rationalizability of Survey Responses. Universitat Pompeu Fabra, Department of Economics and Business.

Baldassi, C., S. Cerreia-Vioglio, F. Maccheroni, and M. Marinacci (2020). A Behavioral Characterization of the Drift Diffusion Model and its Multi-Alternative Extension to Choice under Time Pressure. *Management Science 66*(11), 5075–5093.

Baltas, G. and P. Doyle (2001). Random Utility Models in Marketing Research: A Survey. *Journal of Business Research 51*(2), 115–125.

Barrett, G. F. and S. G. Donald (2003). Consistent Tests for Stochastic Dominance. *Econometrica 71*(1), 71–104.

Bechtel, M. M., F. Genovese, and K. F. Scheve (2019). Interests, Norms and Support for the Provision of Global Public Goods: The Case of Climate Cooperation. *British Journal of Political Science 49*(4), 1333–1355.

Bechtel, M. M. and R. Liesch (2020). Reforms and Redistribution: Disentangling the Egoistic and Sociotropic Origins of Voter Preferences. *Public Opinion Quarterly 84*(1), 1–23.

Bechtel, M. M. and K. F. Scheve (2013). Mass Support for Global Climate Agreements Depends on Institutional Design. *Proceedings of the National Academy of Sciences 110*(34), 13763–13768.

Bechtel, M. M., K. F. Scheve, and E. van Lieshout (2020). Constant Carbon Pricing Increases Support for Climate Action Compared to Ramping up Costs Over Time. *Nature Climate Change 10*(11), 1004–1009.

Belzil, C. and T. Jagelka (2024). Separating True Preferences from Noise and Endogenous Effort. Unpublished.

Bertrand, M. and S. Mullainathan (2001). Do People Mean What They Say? Implications for Subjective Survey Data. *American Economic Review (Papers & Proceedings) 91*(2), 67–72.

Block, H. D. and J. Marschak (1960). Random Orderings and Stochastic Theories of Responses. In I. Olkin (Ed.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pp. 97–132. Stanford: Stanford University Press.

Bond, T. N. and K. Lang (2019). The Sad Truth about Happiness Scales. *Journal of Political Economy 127*, 1629–1640.

Bursztyn, L., A. L. González, and D. Yanagizawa-Drott (2020). Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia. *American Economic Review 110*(10), 2997–3029.

Camerer, C. F. (1989). Does the Basketball Market Believe in the 'Hot Hand'? *American Economic Review 79*, 1257–1261.

Chabris, C. F., C. L. Morris, D. Taubinsky, D. Laibson, and J. P. Schuldt (2009). The Allocation of Time in Decision-Making. *Journal of the European Economic Association 7*(2-3), 628–637.

Coffman, K. B., L. C. Coffman, and K. M. M. Ericson (2017). The Size of the LGBT Population and the Magnitude of Antigay Sentiment Are Substantially Underestimated. *Management Science 63*(10), 3168–3186.

Dashiell, J. F. (1937). Affective Value-Distances as a Determinant of Aesthetic Judgment-Times. *American Journal of Psychology 50*, 57–67.

d'Aspremont, C. and L. Gevers (2002). Social Welfare Functionals and Interpersonal Comparability. In K. J. Arrow, A. K. Sen, and K. Suzumura (Eds.), *Handbook of Social Choice and Welfare*, pp. 450–541. Elsevier.

de Quidt, J., J. Haushofer, and C. Roth (2018). Measuring and Bounding Experimenter Demand. *American Economic Review 108*(11), 3266–3302.

Duffy, S. and J. Smith (2025). Stochastic Choice and Imperfect Judgments of Line Lengths: What is Hiding in the Noise? *Journal of Economic Psychology 106*, 102787.

Enke, B., R. Rodríguez-Padilla, and F. Zimmermann (2022). Moral Universalism: Measurement and Economic Relevance. *Management Science 68*(5), 3590–3603.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018). Global Evidence on Economic Preferences. *Quarterly Journal of Economics 133*(4), 1645–1692.

Fechner, T. G. (1860). *Elemente der Psychophysik [Elements of Psychophysics]*. Leipzig: Breitkopf & Härtel.

Fehr, E., T. Epper, and J. Senn (2022). Other-Regarding Preferences and Redistributive Politics. University of Zurich, Department of Economics, Working Paper.

Feng, Y., R. Caldentey, and C. T. Ryan (2022). Robust Learning of Consumer Preferences. *Operations Research 70*(2), 918–962.

Fisman, R., S. S. Iyengar, E. Kamenica, and I. Simonson (2006). Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment. *Quarterly Journal of Economics 121*(2), 673–697.

Fisman, R., S. S. Iyengar, E. Kamenica, and I. Simonson (2008). Racial Preferences in Dating. *Review of Economic Studies 75*(1), 117–132.

Frank, R. H. (2016). *Success and Luck: Good Fortune and the Myth of Meritocracy*. Princeton: Princeton University Press.

Fudenberg, D., P. Strack, and T. Strzalecki (2018). Speed, Accuracy, and the Optimal Timing of Choices. *American Economic Review 108*(12), 3651–3684.

Funk, P. (2016). How Accurate Are Surveyed Preferences for Public Policies? Evidence From a Unique Institutional Setup. *Review of Economics and Statistics 98*(3), 442–454.

Giglio, S., M. Maggiori, J. Stroebel, and S. Utkus (2021). Five Facts about Beliefs and Portfolios. *American Economic Review 111*(5), 1481–1522.

Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study. *Journal of Political Economy 127*(4), 1826–1863.

Goddard, M. and P. Smith (2001). Equity of Access to Health Care Services: Theory and Evidence from the UK. *Social Science & Medicine 53*(9), 1149–1162.

Guriev, S. and D. Treisman (2020). The Popularity of Authoritarian Leaders: A Cross-National Investigation. *World Politics 72*(4), 601–638.

Hainmueller, J., D. Hangartner, and T. Yamamoto (2015). Validating Vignette and Conjoint Survey Experiments Against Real-World Behavior. *Proceedings of the National Academy of Sciences 112*(8), 2395–2400.

Hayes, A. F. (2018). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach, 2nd Ed.* New York: Guilford Publications.

Hey, J. D. and C. Orme (1994). Investigating Generalizations of Expected Utility Theory Using Experimental Data. *Econometrica 62*(6), 1291–1326.

Horowitz, J. L. (2001). The bootstrap. In *Handbook of Econometrics*, Volume 5, pp. 3159–3228. Amsterdam: Elsevier.

Karp, J. A. (2009). Candidate Effects and Spill-over in Mixed Systems: Evidence from New Zealand. *Electoral Studies 28*(1), 41–50.

Klein, A. S. (2001). Measuring, Estimating, and Understanding the Psychometric Function: A Commentary. *Attention, Perception, & Psychophysics 63*(8), 1421–1455.

Laming, D. (1985). Some Principles of Sensory Analysis. *Psychological Review 92*(4), 462–485.

Laming, D. R. J. (1968). *Information Theory of Choice-Reaction Times.* Academic Press, New York, NY.

Leeper, T. J., S. B. Hobolt, and J. Tilley (2020). Measuring Subgroup Preferences in Conjoint Experiments. *Political Analysis 28*(2), 207–221.

Liu, S. and N. Netzer (2023). Happy Times: Measuring Happiness Using Response Times. *American Economic Review 113*(12), 3289–3322.

Lo, A. (2019). Demystifying the Integrated Tail Probability Expectation Formula. *The American Statistician 73*(4), 367–374.

Loomes, G., P. G. Moffatt, and R. Sugden (2002). A Microeconometric Test of Alternative Stochastic Theories of Risky Choice. *Journal of Risk and Uncertainty 24*(2), 103–130.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis.* New York: Wiley.

Luce, R. D. and J. W. Tukey (1964). Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement. *Journal of Mathematical Psychology 1*(1), 1–27.

Manski, C. F. (2005). Measuring Expectations. *Econometrica 72*(5), 1329–1376.

McFadden, D. (1989). Testing for Stochastic Dominance. In T. B. Fomby and T. K. Seo (Eds.), *Studies in the Economics of Uncertainty*, pp. 113–134. New York, NY: Springer New York.

McFadden, D. L. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. New York: Academic Press.

McFadden, D. L. (2001). Economic Choices. *American Economic Review 91*(3), 351–378.

Moffatt, P. G. (2005). Stochastic Choice and the Allocation of Cognitive Effort. *Experimental Economics 8*(4), 369–388.

Moyer, R. S. and T. K. Landauer (1967). Time Required for Judgements of Numerical Inequality. *Nature 215*(5109), 1519–1520.

Nownes, A. J. (1992). Primaries, General Elections, and Voter Turnout: A Multinomial Logit Model of the Decision to Vote. *American Politics Quarterly 20*(2), 205–226.

Ortoleva, P., E. Safonov, and L. Yariv (2021). Who Cares More? Allocation with Diverse Preference Intensities. National Bureau of Economic Research.

Shadlen, M. N. and R. Kiani (2013). Decision Making as a Window on Cognition. *Neuron 80*, 791–806.

Sides, J. and J. Citrin (2007). European Opinion About Immigration: The Role of Identities, Interests and Information. *British Journal of Political Science 37*(3), 477–504.

Snowberg, E. and L. Yariv (2021). Testing the Waters: Behavior Across Participant Pools. *American Economic Review 111*(2), 687–719.

Spiller, S. A., G. J. Fitzsimons, J. G. Lynch, and G. H. McClelland (2013). Spotlights, Floodlights, and the Magic Number Zero: Simple Effects Tests in Moderated Regression. *Journal of Marketing Research 50*(2), 277–288.

Thurstone, L. L. (1927). A Law of Comparative Judgement. *Psychological Review 34*, 273–286.

Tversky, A. (1969). Intransitivity of Preferences. *Psychological Review 76*, 31–48.

Tversky, A. and D. Kahneman (1981). The Framing of Decisions and the Psychology of Choice. *Science 211*(4481), 453–458.

Tversky, A. and D. Kahneman (1986). Rational Choice and the Framing of Decisions. *Journal of Business 59*(4), S251–S278.

Tversky, A., S. Sattath, and P. Slovic (1988). Contingent Weighting in Judgment and Choice. *Psychological Review 95*(3), 371–384.

Van Kerm, P. (2012). Kernel-Smoothed Cumulative Distribution Function Estimation with Akdensity. *The Stata Journal 12*(3), 543–548.

Wichmann, A. F. and N. J. Hill (2001). The Psychometric Function: I. Fitting, Sampling, and Goodness of Fit. *Attention, Perception, & Psychophysics 63*(8), 1293–1313.

Zaller, J. and S. Feldman (1992). A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences. *American Journal of Political Science 36*(3), 579–616.

Zizzo, D. J. (2010). Experimenter Demand Effects in Economic Experiments. *Experimental Economics 13*, 75–98.

ONLINE APPENDIX

Who Likes It More?
Using Response Times To Elicit Group Preferences in Surveys

C. Alós-Ferrer and M. Garagnani

Appendix A Proofs

Proof of Corollary 1. By Theorem 1, for any 2RUM-CF $(u_A, \tilde{v}_A; u_B, \tilde{v}_B; r)$ that rationalizes $(p_A, f_A; p_B, f_B)$,

$$u_A(x) \geq u_A(y) \quad \text{and} \quad u_B(w) \geq u_B(z)$$

and thus $u_A(x) - u_A(y) \geq 0 \geq u_B(z) - u_B(w)$. □

The following lemmata will be useful below.

Lemma 1. *If an SCF-RT (p, f) including data on a choice (a, b) is rationalized by a RUM-CF $(u, \tilde{v}, r)$, then*

$$H(a,b)(t) = G(a,b)(r^{-1}(t)) + G(a,b)(-r^{-1}(t)) - 1.$$

Proof of Lemma 1. Let $g(a,b)$ be the density function of $\tilde{v}(a,b)$ and $G(a,b)$ denote the corresponding cumulative distribution function. By Definitions 1, 2, and 3, $p(b,a) = G(a,b)(0)$, $p(a,b) = 1 - G(a,b)(0)$, $F(a,b) = (1 - G(a,b)(r^{-1}(t))/(1 - G(0))$, and $F(b,a)(t) = G(a,b)(-r^{-1}(t))/G(a,b)(0)$. Thus,

$$H(a,b)(t) = G(a,b)(-r^{-1}(t)) - (1 - G(a,b)(r^{-1}(t)) = G(a,b)(r^{-1}(t) + G(a,b)(-r^{-1}(t)) - 1.$$

□

Lemma 2. *For any 2RUM-CF $(u_A, \tilde{v}_A; u_B, \tilde{v}_B; r)$,*

$$E[\tilde{v}_A(x,y)] - E[\tilde{v}_B(z,w)] = \int_{-\infty}^{+\infty} [G_B(z,w)(v) - G_A(x,y)(v)]\, dv.$$

Proof of Lemma 2. By the integrated tail formula for expectations (Lo, 2019), if F is the cumulative distribution function of the real-valued random variable X,

$$E[X] = -\int_{-\infty}^{0} F(x)dx + \int_{0}^{+\infty} (1 - F(x))dx.$$

Thus,

$$E[\tilde{v}_A(x,y)] - E[\tilde{v}_B(z,w)] =$$
$$-\int_{-\infty}^{0} G_A(x,y)(v)dv + \int_{0}^{\infty} (1 - G_A(x,y)(v))dv + \int_{-\infty}^{0} G_B(z,w)(v)dv - \int_{0}^{\infty} (1 - G_B(z,w)(v))dv =$$
$$\int_{-\infty}^{+\infty} [G_B(z,w)(v) - G_A(x,y)(v)]\, dv.$$

□

Proof of Theorem 2. The proof is analogous to that of Theorem 1 in Alós-Ferrer et al. (2021), slightly condensed thanks to the use of the lemmata above. Consider any 2RUM-CF $(u_A, \tilde{v}_A; u_B, \tilde{v}_B; r)$ that rationalizes $(p_A, f_A; p_B, f_B)$.

By Lemma 2,

$$[u_A(x) - u_A(y)] - [u_B(z) - u_B(w)] = E[\tilde{v}_A(x,y)] - E[\tilde{v}_B(z,w)] =$$
$$= \int_{-\infty}^{+\infty} [G_B(z,w)(v) - G_A(x,y)(v)]\, dv =$$
$$\int_{-\infty}^{0} [G_B(z,w)(v) - G_A(x,y)(v)]\, dv + \int_{0}^{+\infty} [G_B(z,w)(v) - G_A(x,y)(v)]\, dv =$$
$$\int_{0}^{+\infty} [G_B(z,w)(-v) - G_A(x,y)(-v)]\, dv + \int_{0}^{+\infty} [G_B(z,w)(v) - G_A(x,y)(v)]\, dv =$$
$$\int_{0}^{+\infty} [G_B(z,w)(v) + G_B(z,w)(-v) - G_A(x,y)(v) - G_A(x,y)(-v)]\, dv$$

For any $v > 0$, let $t = r(v)$. By Lemma 1, $H_A(x,y)(t) \leq H_B(z,w)(t)$ can be rewritten as

$$G_B(v) + G_B(-v) \geq G_A(v) + G_A(-v)$$

for any v with $t = r(v) > 0$. The inequality follows for $v = 0$ by continuity. For any v with $r(v) = 0$, it follows because in that case $G_A(x,y)(v) = G_B(z,w) = 1$ and $G_A(x,y)(-v) = G_B(z,w)(-v) = 0$, as otherwise the corresponding RUM-CF would generate an atom at the response time of zero. We conclude that the inequality holds for all $v \geq 0$. Thus, the integral above is positive and the conclusion follows.

If $H_A(x,y)(t) > H_B(z,w)(t)$ for some t, it is strict for a nonempty interval by continuity, and it follows that the final inequality is also strict. □

Proof of Theorem 3. "If." Suppose (a) and (b) hold, and let $(u_A, \tilde{v}_A; u_B, \tilde{v}_B; r)$ be a 2RUM-CF that rationalizes the 2SCF-RT. By (a),

$$p_A(y,x)F_A(y,x)(t) \leq p_B(w,z)F_B(w,z)(t) \quad \text{for all } t > 0,$$

which, by Definition 7, implies

$$G_A(x,y)(-r^{-1}(t)) \leq G_B(z,w)(-r^{-1}(t)).$$

This implies that $G_A(x,y)(v) \leq G_B(z,w)(v)$ for any $v < 0$ such that $r(-v) = t > 0$. For $v = 0$, by continuity, the inequality also holds. If $r(v) = 0$ (as in the proof of Theorem 2), the inequality also holds since the absence of atoms at a response time of zero implies that $G_A(x,y)(v) = 0$ and $G_B(z,w)(v) = 1$.

An analogous argument shows that $G_A(x,y)(v) \leq G_B(z,w)(v)$ also for any $v \geq 0$. This completes the proof of the "if" direction.

"Only if."suppose that the 2SCF-RT reveals that group A prefers x over y uniformly more than group B prefers z over w. Let $(u_A, \tilde{v}_A; u_B, \tilde{v}_B; r)$ be a 2RUM-CF that rationalizes the 2SCF-RT. Hence, $G_A(x,y)(v) \leq G_B(z,w)(v)$ for all v.

Analogously to the "if" direction, evaluating this inequality for $-r^{-1}(t)$ proves (a), and evaluating it for $r^{-1}(t)$ proves (b). □

Proof of Corollary 2. The argument is identical to the "only if" direction in the proof of Theorem 3, with the only difference that the 2RUM-CF is fixed and it rationalizes the 2SCF-RT because the latter is taken to be generated by the former. □

Appendix B Additional Details on Survey Design

The sample size and power analysis is based on the tests of proportions for comparisons across groups. Given that in some cases participants might not be equally distributed between groups (e.g., political attitudes), we conservatively set the sample size to be able to allow for a 0.25 allocation ratio (80% vs. 20%). We further require to have enough power (0.8) to detect a 0.1 difference in the proportion of people supporting one option compared to the other between groups. The resulting sample size fulfilling these conditions is $N = 1008$. At the time the survey was conducted, a representative sample in Prolific required $N = 1200$, and hence we used this sample size.

During the survey, we implemented control questions for attention, recruiting participants who successfully passed the control questions until we reached the required sample size. Due to simultaneous online recruitment, the target was slightly exceeded ($N = 1202$).

Appendix C Survey 1: Additional Details and Analyses

C.1 List of Charities and Organizations To Be Rated

Charities are selected because they are either popular or unpopular, according to yougov.uk for Q3-2023. The popular charities we include had approval rates between 73% and 86% (this was the overall maximum) at the time of the design of the study (November 2023), and a "Fame" of 90% or above: British Heart Foundation; Macmillan Cancer Support; Samaritans; WWF; British Red Cross; Guide Dogs; RSPB (Royal Society for the Protection of Birds). We tried to include a variety of themes, e.g. avoiding having two different cancer charities on the list.

Less popular charities had approval rates between 24% and 60%: UNICEF; Fairtrade Foundation; Amnesty International; Cats Protection; Christian Aid; Greenpeace; Unite to End Violence Against Women; Black Lives Matter; PETA; Campaign for Nuclear Disarmament; Vegetarian Society.

We also include the two main political organizations in the UK, the Conservative Party and the Labour Party.

C.2 List of Framing and Related Questions

Q1-V1 You are the Health Minister of your country. Imagine that your country is preparing for the outbreak of an unusual disease, which is expected to kill 3000 people. Your team has investigated two alternative programs to combat the disease. Assume that the exact scientific estimates of the consequences of the programs are as follows. Which program would you choose?

- If Program A is adopted, 2000 people will die.
- If Program B is adopted, there is a 1/3 chance that nobody will die, and a 2/3 chance that 3000 people will die.

Q1-V2 You are the Health Minister of your country. Imagine that your country is preparing for the outbreak of an unusual disease, which is expected to kill 3000 people. Your team has investigated two alternative programs to combat the disease. Assume that the exact scientific estimates of the consequences of the programs are as follows. Which program would you choose?

- If Program A is adopted, 1000 people will be saved.
- If Program B is adopted, there is a 1/3 chance that 3000 people will be saved, and a 2/3 probability that nobody will be saved.

Q2-V1 You are the Minister of Transportation of your country. Imagine that about 1000 people are killed in the country in traffic accidents every year. Your team has investigated two alternative programs to reduce the number of casualties. Assume that the exact expected consequences and yearly costs of the programs are as follows. Which program would you choose?

- If Program A is adopted, there will be 800 casualties. The program will cost 55 million Pounds.
- If Program B is adopted, there will be 920 casualties. The program will cost 12 million Pounds.

Q2-V2 You are the Minister of Transportation of your country. Imagine that about 1000 people are killed in the country in traffic accidents every year. Your team has investigated two alternative programs to reduce the number of casualties. Assume that the exact expected consequences and yearly costs of the programs are as follows. Which program would you choose?

- If Program A is adopted, there will be 200 less casualties. The program will cost 55 million Pounds.
- If Program B is adopted, there will be 80 less casualties. The program will cost 12 million Pounds.

Q3-V1 Imagine that you have decided to see a concert where admission is 100 Pounds per ticket. As you enter the concert hall you discover that you have lost a 100 Pounds bill. Would you still pay 100 Pounds for a ticket for the concert? (Yes/No)

Q3-V2 Imagine that you have decided to see a concert and paid the admission price of 100 Pounds per ticket. As you enter the concert hall you discover that you have lost the ticket. The seat was not marked and the ticket cannot be recovered. Would you pay 100 Pounds for another ticket? (Yes/No)

Q4-V1 Imagine that you are about to purchase a jacket for 150 Pounds and a mobile phone for 1250 Pounds. The mobile phone salesman informs you that the model you wish to buy is on sale for 1200 Pounds at another branch of the store, located 20 minutes away. Would you do the trip? (Yes/No)

Q4-V2 Imagine that you are about to purchase a jacket for 150 Pounds and a mobile phone for 1250 Pounds. The clothing-store salesman informs you that the jacket you wish to buy is on sale for 100 Pounds at another branch of the store, located 20 minutes away. Would you do the trip? (Yes/No)

C.3 Additional Results (Survey 1)

C.3.1 Revealing Objective Preferences

The five line pairs reported in Figure 2 in the main text were (50pt,45pt), (50pt,46pt), (50pt,47pt), (50pt,48pt), and (49pt,48pt). The experiment also included five further line pairs with lengths (49pt,45pt), (49pt,46pt), (47pt, 45pt), (48pt, 46pt), and (47pt, 46pt),

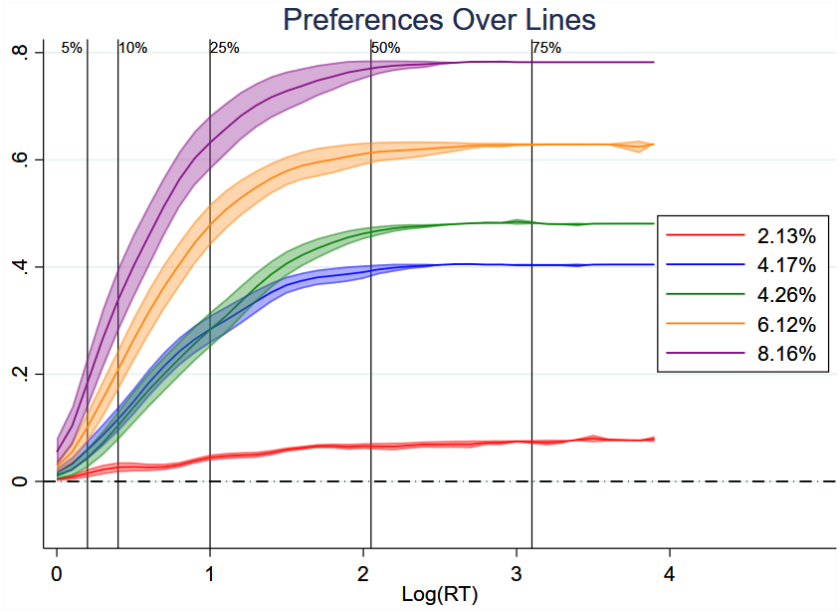

Figure C.1: Revealed preferences over line lengths.

corresponding to differences of 8.16%, 6.12%, 4.26%, 4.17%, and 2.13%, respectively. Figure C.1 depicts the results for these five pairs.

The proportions of correct answers are 89.11%, 73.41%, 70.24%, 66.61%, and 53.81%, from largest to smallest difference, in agreement with psychometric regularities. Application of Theorem 1 and the test described in section 3.2.2 leads to clear preference revelation for all comparisons. Specifically, after applying the exclusions discussed in Section 3.2.4 (and, in particular, the 5% fastest participants), the H functions for all five pairs are clearly above the horizontal zero line, as are their confidence intervals (all five tests $p < 0.001$), Strikingly, this includes even the 2.13% comparison, even though only 53.81% of answers were correct.

Applying Theorem 2 and the statistical test in Section 3.2.3 reveals stronger preferences for pairs with larger objective differences in all two-pair comparisons except one (nine tests, all $p < 0.001$). The only exception is the comparison between the extremely-similar pairs with 4.17% and 4.26% differences ($p = 0.500$) and between 4.26% and 6.12% ($p = 0.122$). That is, again, we clearly recover stronger (objective) preferences for larger differences between the presented stimuli even when we measure response times in a noisy environment as online surveys.

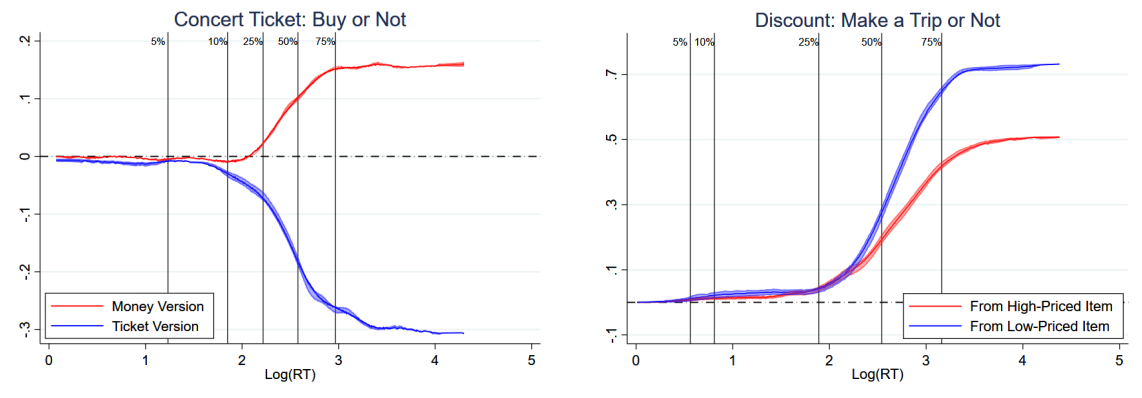

Figure C.2: Two further questions from Tversky and Kahneman (1981). Left: Lost-ticket question. Right: Discounting from high or low reference prices.

C.3.2 Revealing the Strength of Subjective Preferences

Figure C.2 reports the data from two other questions taken from Tversky and Kahneman (1981). In the first (left-hand panel), participants are asked whether or not they would buy a concert ticket after having lost either the monetary amount of the ticket (money version) or a previously-purchased concert ticket (ticket version). The result, which Tversky and Kahneman (1981) interpret as a sunk-cost effect, is that people tend to buy the ticket more often if they had lost its monetary equivalent than if they had lost a previously-purchased ticket.

Choice frequencies show the same effect as in Tversky and Kahneman (1981). 57.82% of participants exposed to the lost-money version of the question decided to buy the ticket, compared to only 34.96% of participants exposed to the lost-ticket (test of proportions, $z = 7.6069, p < 0.001$). Application of Theorem 1 reveals a group preference for not buying the ticket for the group of people assigned to the lost-ticket version ($p < 0.001$). However, as the picture illustrates, no preference is revealed for the group of people assigned to the lost-money version ($p = 0.457$). Application of Theorem 2, however, recovers the postulated effect. i.e. a stronger group preference for not buying the ticket for the lost-ticket group compared to the lost-money group ($p < 0.001$).

In the question reported in the right-hand panel of Figure C.2, participants are asked whether or not they would make a 20-minute trip to save 50 GBP. Participants were randomly assigned to two versions of the question, where the savings were a discount on either a high-priced or a low-priced item (50 GBP from either 1250 GBP or 150 GBP). Tversky and Kahneman (1981) reported a higher willingness to undertake the trip when the discount was from a low-priced item. Choice frequencies seem to confirm this effect, with 86.59% of people agreeing to make the trip for a discount from the low-priced item vs. only 75.45% for a discount from the high-priced item (test of proportions, $z = 4.7164$, $p < 0.001$). Application of Theorem 1 reveals a group preference for sacrificing 20 minutes to save 50 GBP in both groups (both $p < 0.001$). However, as Figure C.2(right) shows, Theorem 2 does not reveal a difference in preference strengths across groups ($p = 0.495$).[11]

[11] We remark that we scaled up the original question of Tversky and Kahneman (1981) by a factor of 10 to account for inflation. In hindsight, GBP 50 might have been an extremely high price to pay for 20 minutes for Prolific participants, given that Prolific requires a payment of 6 GBP per hour.

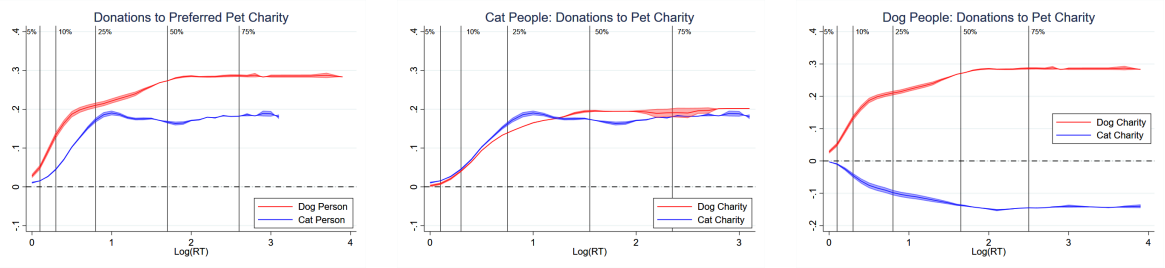

Figure C.3: Donation questions (donate vs. keep). Top-Left: Preferences for donating to a cat or dog charity, for participants self-declaring as cat or dog people, respectively. Center and right panels: Preferences for donating to the favored-pet charity vs. the other pet, for cat people (center) and dog people (right).

C.3.3 Revealing Group Preferences for Charity Donations

Figure C.3 depicts the results for the donation questions to the charities "Guide Dogs" and "Cats Protection," splitting the sample into two groups according to whether participants declared to be dog or cat people. The left-hand panel compares the preference to donate to the charity corresponding to the favourite pet. Application of Theorem 1 reveals that both dog- and cat-people exhibit a group preference to donate to charities for their favourite pets (both $p < 0.001$), but application of Theorem 2 reveals that dog people have a stronger preference to donate than cat people ($p < 0.001$). While this effect is quite clear through the test for Theorem 2, a test of proportions only reveals a marginally significant difference (cat people 59.19%, dog people 64.18%, $z = -1.6745$, $p = 0.094$).

The center and right panels of Figure C.3 complete the validation and recovery exercise for cat and dog charities, showing that cat people have a preference for donating to a cat charity (center panel, $p < 0.001$), but also a preference for donating to a dog charity ($p = 0.011$). Further, applying Theorem 2, the two preferences are not revealed to be different ($p = 0.463$). This agrees with a test of proportions (cat charity 59.19%, dog charity 60.09%, $z = -0.2730$, $p = 0.785$). However, for dog people (right panel), in addition to a preference for donating to a dog charity ($p < 0.001$), also a preference for *not* donating to a cat charity ($p < 0.001$) is revealed. Of course, application of Theorem 2 reveals a stronger preference for donating to the dog compared to the cat charity ($p < 0.001$), as also implied by Corollary 1 in this case. A test of proportions supports the same conclusions (cat charity 42.84%, dog charity 64.18%, test of proportions $z = -7.7493$, $p < 0.001$).

C.3.4 Revealing Group Preferences for Charity Donations Using Ratings

In Survey 1, participants ranked each of the 20 charities and organizations using a Likert scale (1 to 10). For each organization, we divided the sample of participants using a median split of the rating distribution. We then applied Theorem 1 separately to both groups (above and below the median) to reveal their preferences in favor or against donating for each of the organizations. Finally, we applied Theorem 2 to reveal the relative strength of preference between the two groups.

Table C.1 summarizes the obtained results. The first column lists the 20 organizations, and the second column reports the median rating (in descending order). For people above the median (columns 3–4), Theorem 1 reveals a preference in 14 of the 20 cases (all $p < 0.001$), 5 in favor of donating to the corresponding organization and 9 against. For people below the median (columns 5–6), Theorem 1 reveals a preference in 13 of the 20 cases (again all $p < 0.001$), 2 in favor of donating to the corresponding organization and 11 against. The last column reports the results of the tests using Theorem 2, which reveal a difference between the preferences of the groups in 13 of the 20 cases (all $p \leq 0.001$). In all of those cases, the revealed difference is that people who rate the organization above the median reveal a stronger preference to donate to that organization than people who rate it below the median. The H functions and confidence intervals for each of the 20 organizations are plotted in Figures C.4–C.8.

Table C.1: Revealed preferences for each of the 20 charities based on the median split of participants' ratings. *Median Rating*: median of the distribution of ratings, used to build the median-split. *Above (Below) Median*: Data por participants ranking the charity above (below) the median. % Donation: percentage of people who donated to the charity. p-value: test for Theorem 1 (for each group). *Comparison*: p-value of the test for Theorem 2.

Charity	Median	Above median		Below median		Comparison
	Rating	% Donations	p-value	% Donations	p-value	p-value
British Red Cross	8.41	64.22	$p = 0.457$	61.68	$p < 0.001$	$p = 0.488$
Unite to End Violence Against Women	8.01	66.52	$p = 0.356$	52.47	$p = 0.496$	$p < 0.001$
Campaign for Nuclear Disarmament	7.99	33.53	$p < 0.001$	30.51	$p < 0.001$	$p = 0.001$
British Heart Foundation	7.71	83.20	$p < 0.001$	61.74	$p = 0.496$	$p < 0.001$
Macmillan Cancer Support	7.66	79.85	$p < 0.001$	68.22	$p = 0.492$	$p < 0.001$
Conservative Party (UK)	7.54	9.28	$p < 0.001$	8.03	$p < 0.001$	$p = 0.330$
UNICEF	7.49	57.33	$p = 0.493$	55.50	$p = 0.491$	$p = 0.498$
Cats Protection	7.43	53.28	$p = 0.486$	46.82	$p < 0.001$	$p = 0.413$
Fairtrade Foundation	7.42	51.09	$p = 0.499$	38.49	$p < 0.001$	$p < 0.001$
PETA	7.33	38.26	$p < 0.001$	27.35	$p < 0.001$	$p < 0.001$
Royal Society for the Protection of Birds	7.17	49.30	$p < 0.001$	50.48	$p = 0.059$	$p = 0.499$
Vegetarian Society	7.16	23.85	$p < 0.001$	16.46	$p < 0.001$	$p < 0.001$
Greenpeace	7.13	43.65	$p < 0.001$	39.28	$p < 0.001$	$p = 0.491$
WWF	6.53	56.59	$p = 0.378$	55.17	$p = 0.488$	$p < 0.001$
Guide Dogs	6.40	65.60	$p < 0.001$	69.10	$p < 0.001$	$p = 0.496$
Labour Party (UK)	6.36	20.91	$p < 0.001$	15.01	$p < 0.001$	$p < 0.001$
Black Lives Matter	6.33	42.90	$p < 0.001$	31.94	$p < 0.001$	$p < 0.001$
Amnesty International	5.79	55.95	$p < 0.001$	38.13	$p < 0.001$	$p < 0.001$
Samaritans	4.90	64.82	$p < 0.001$	61.70	$p = 0.492$	$p < 0.001$
Christian Aid	2.98	41.12	$p < 0.001$	30.66	$p < 0.001$	$p < 0.001$

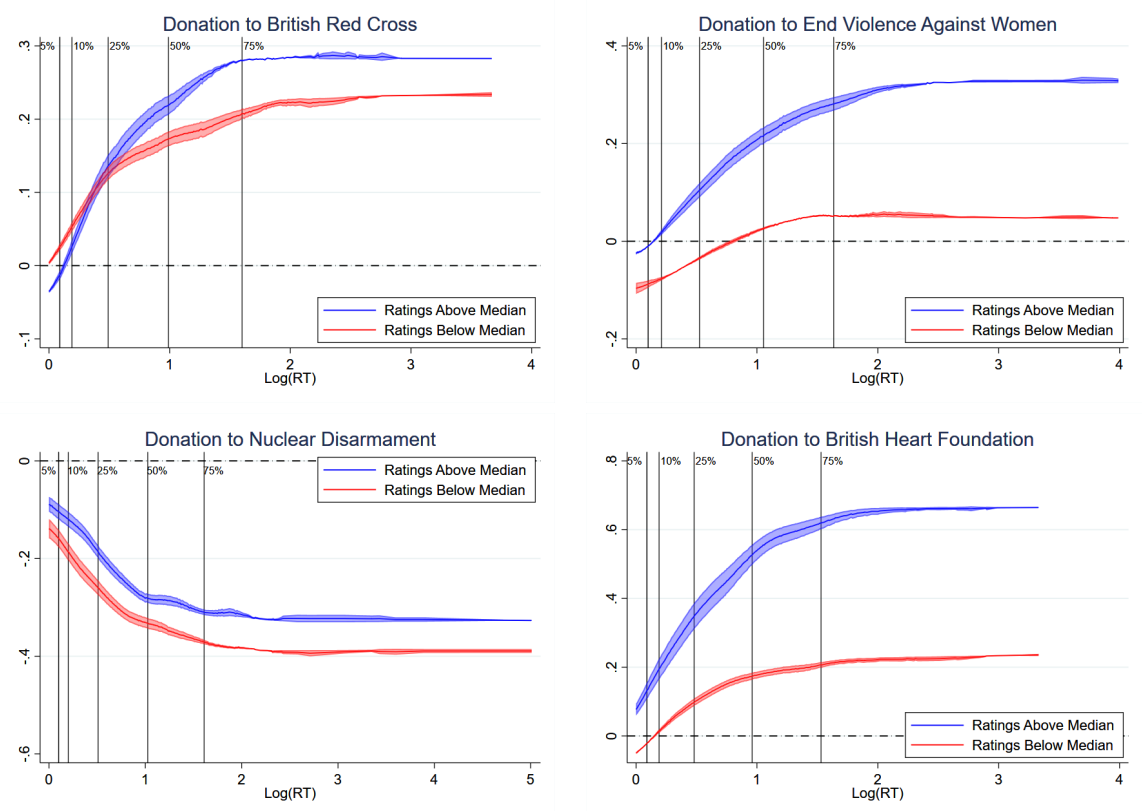

Figure C.4:

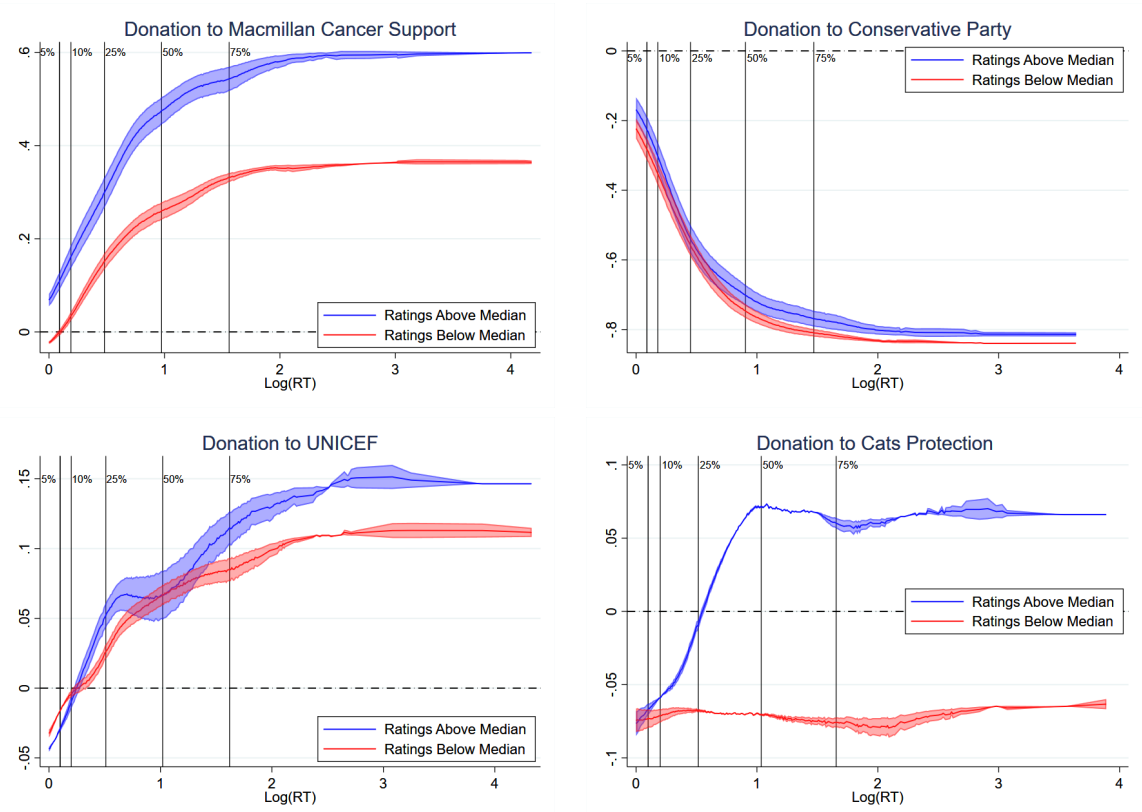

Figure C.5:

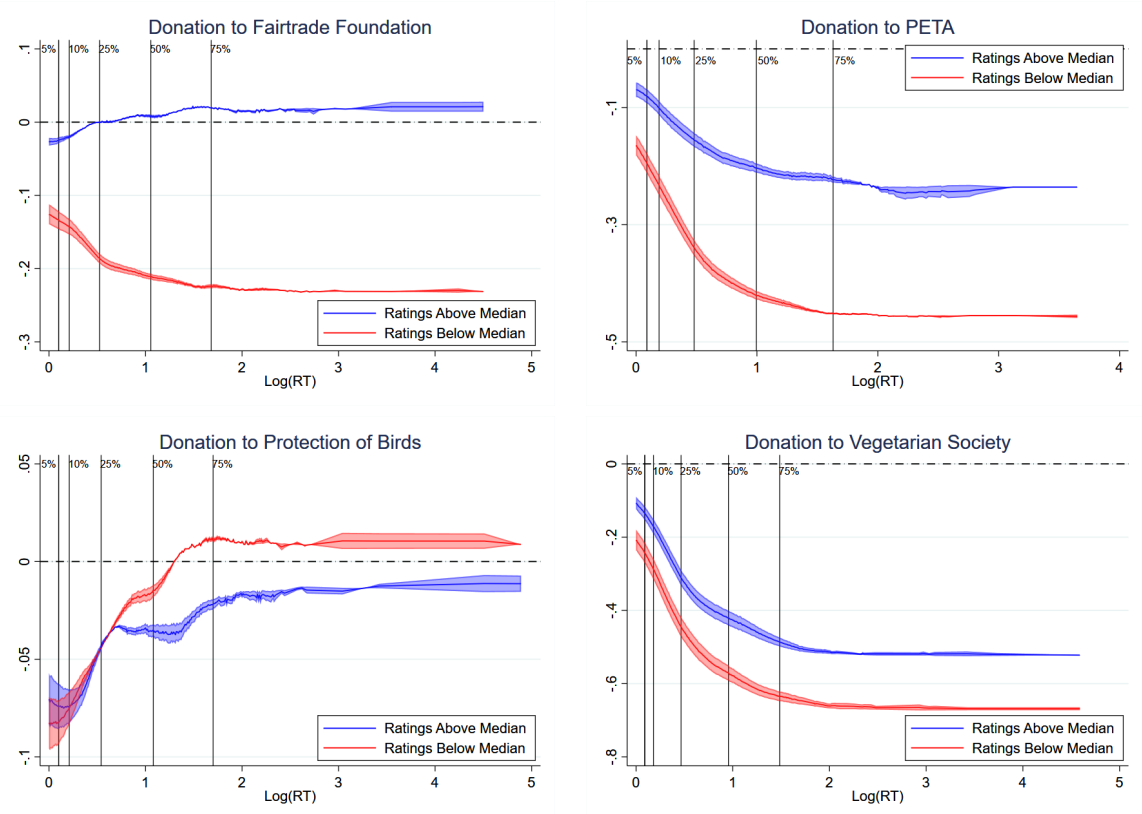

Figure C.6:

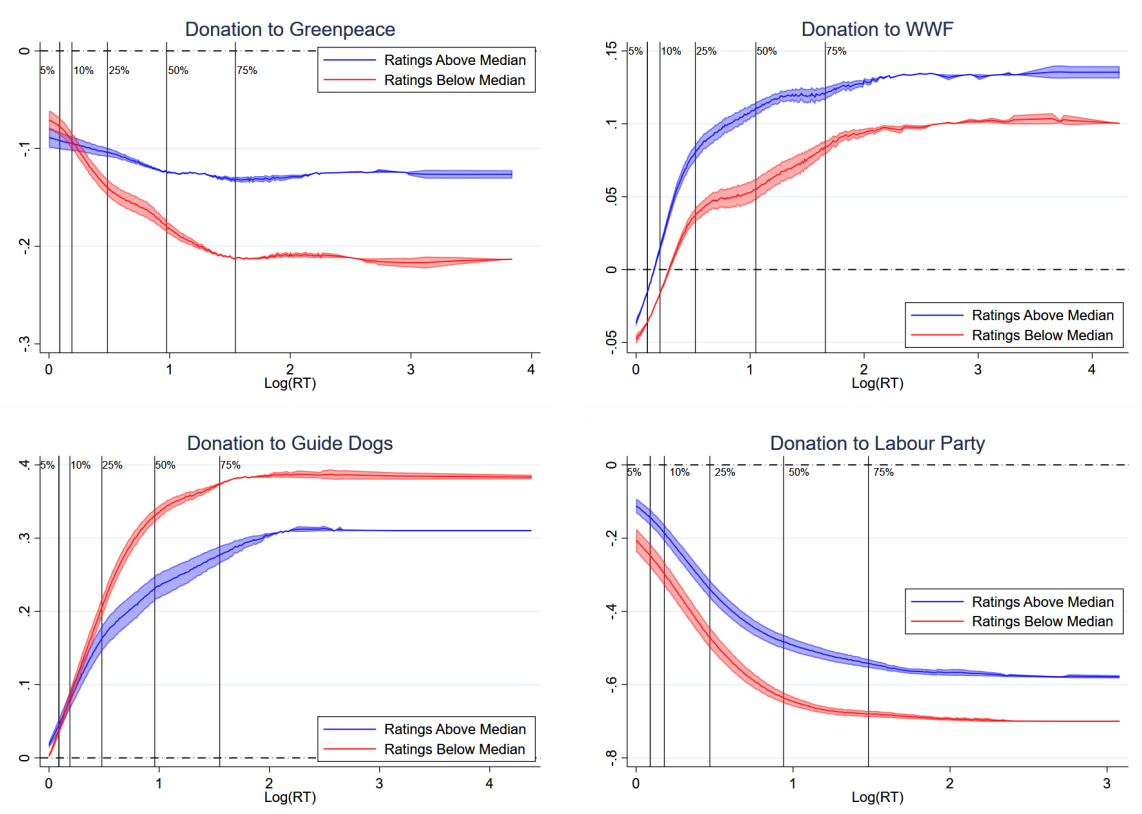

Figure C.7:

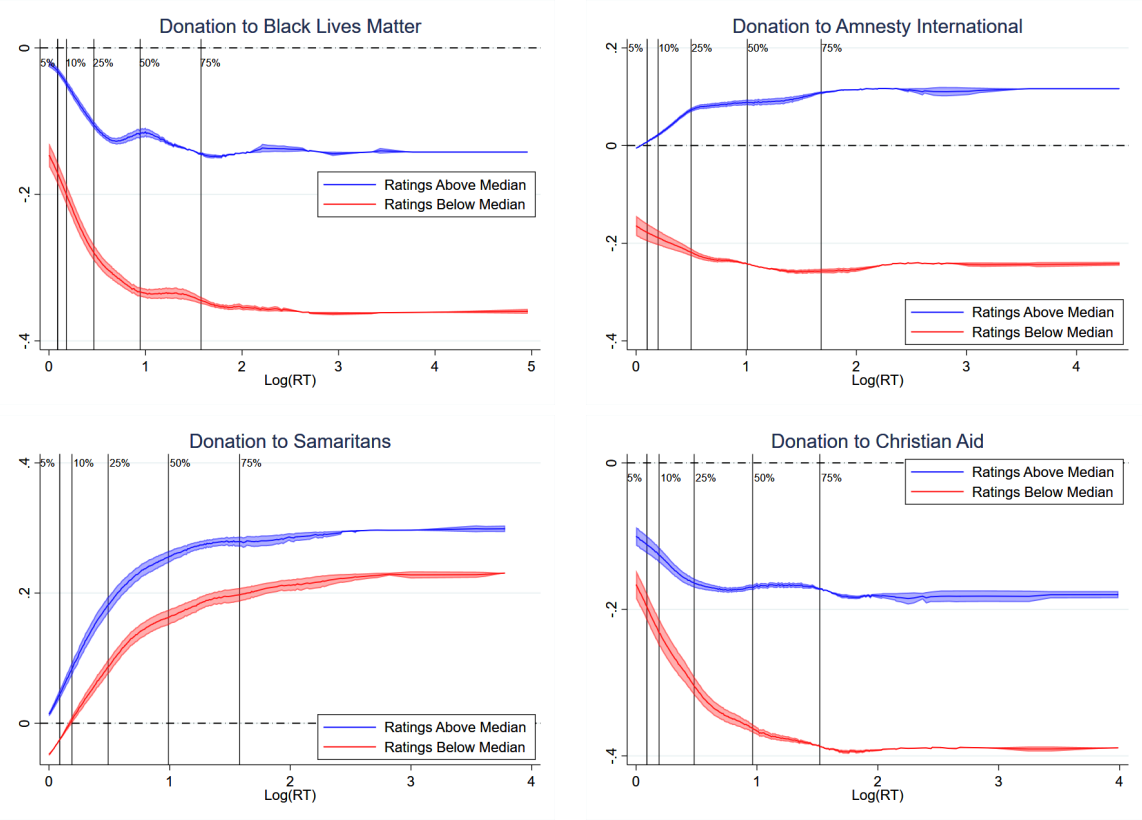
Donation to Black Lives Matter
5%
10%
25%
50%
75%
Ratings Above Median
Ratings Below Median
Log(RT)
Donation to Amnesty International
Donation to Samaritans
Donation to Christian Aid

Figure C.8:

Appendix D Survey 2: Additional Details and Analyses

D.1 List of Questions (Survey 2)

The actual order of questions in the survey was randomized.

- Group division based on gender (classification on the basis of gender reported in the prolific registration).

  The first three questions are inspired by Bursztyn et al. (2020).

  1. Are you in favor of gender quotas? [Yes/No]
  2. Do you think women should earn less than men for the same job? [Yes/No]
  3. Do you think mothers should stay at home with their kids instead of working? [Yes/No]
  4. Would you cheat on a partner if given the occasion (and she/he would never find out)? [Yes/No]
  5. Would you cheat on taxes if you had a chance? [Yes/No]

- Group division based on age (median split on the basis of age reported in the prolific registration).

  The first two questions are inspired by Bechtel and Scheve (2013) and Bechtel et al. (2019). The third question in this group is inspired by Bechtel et al. (2020).

  1. Do you think rich countries should pay more than poor countries to finance policies decreasing greenhouse gas emissions (independently of the individual history of emissions)? [Yes/No]
  2. Do you think policies to decrease greenhouse gas emissions should be financed by rich countries only, instead of by all countries proportionally to current emissions? [Yes, only rich countries / No, proportionally to current emissions]
  3. Do you think policies to decrease greenhouse gas emissions should impose increasing costs over time (countries pay more in the future than now) compared to constant costs? [Yes, increasing costs / No, constant costs over time]
  4. Should the government impose a tax on industry to discourage industry practices that contribute to global warming? [Yes/No]
  5. Are you in favor of legalising the use and consumption of Cannabis (marijuana)? [Yes/No]

- Group division based on political orientation (classification on the basis of question at the end, see below).

  The first question is inspired by Frank (2016) while the second and third questions are inspired by Sides and Citrin (2007). The fourth question is inspired by Goddard and Smith (2001).

  1. Do you think people get ahead mostly because of their own merits, or rather because of luck and help from others? [Yes, merit / No, luck and help]
  2. Should your country allow more immigrants to come and live in it? [Yes/No]
  3. Do you think your country should limit international trade to protect national jobs? [Yes/No]

4. Do you think people who don't work should have guaranteed access to health services? [Yes/No]
5. Do you think BREXIT was a good idea? [Yes/No]
6. Are you in favor of same-sex marriage? [Yes/No]
7. Do you think vaccination against COVID-19 should be mandatory? [Yes/No]

- Group division based on household income (on the basis of question at the end, see below).

  The first two questions are inspired by Fehr et al. (2022). The third to sixth questions are inspired by Bechtel and Liesch (2020).

  1. Are you in favor of sharply increasing taxation for people at the upper 1% of the income range? [Yes/No]
  2. Are you in favor of a high inheritance tax? [Yes/No]
  3. Would you support a policy which increases the average yearly income in your country by £5,000, but decreases your personal income by £2,500? [Yes/No]
  4. Would you support a policy which increases your personal income by £2,500, but decreases the average yearly income in your country by £5,000? [Yes/No]
  5. Would you support a policy which increases the average income of the lowest-income quarter of the population by £5,000, but decreases the average income of the rest of the population by £2,500? [Yes/No] *(Comparison for this question is lowest quarter vs. the rest)*
  6. Do you think that it is the responsibility of the government to reduce the differences in income between people with high income and those with low income? [Yes/No]

Final questions for defining groups.

1. What is your household's approximate annual income? [less than £18,000][£18,000; £29,900][£29,901; £62,000][more than £62,000]
2. What describes you best politically: leaning more toward the left or toward the right? [Left/Right]

D.2 Additional Results (Survey 2)

D.2.1 Unwarranted Conclusions: Additional Questions

For three additional questions in Survey 2, a test of proportions suggests a significant difference in group preferences which is unwarranted according to an application of Theorem 2, as in the four examples presented in Section 5.2 in the main text. The data for these additional questions is presented in Figure D.9.

The first question (Figure D.9, top-left) was whether the participant would support a policy increasing the average yearly income in the country by GBP 5,000 while decreasing the participant's personal income by GBP 2,500. Analysis was conducted according to a median split on income (poor vs. rich). Theorem 1 reveals a preference against this policy for both groups (both $p < 0.001$). A test of proportions suggest that the rich support this policy more than the poor (Rich: 25.00% vs. Poor: 18.67%, $z = 2.505$, $p = 0.012$), but an application of Theorem 2 reveals no group differences ($p = 0.499$). A

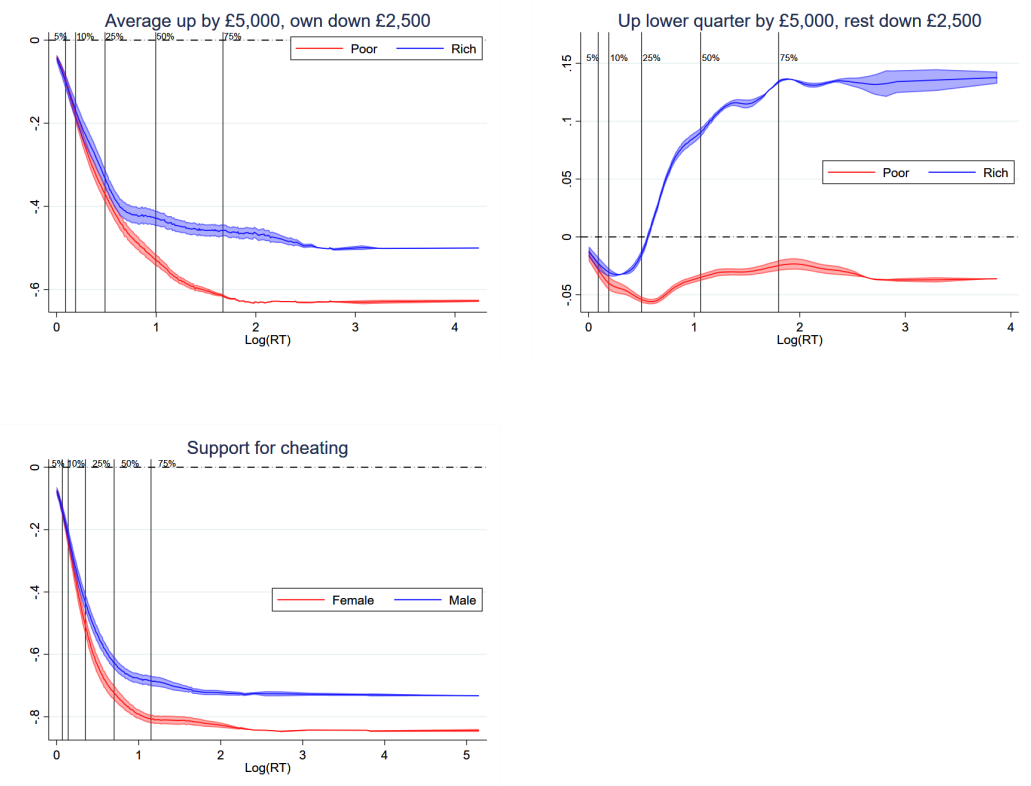

Figure D.9: Three further questions where an apparent difference in group preferences according to a significant test of proportions is flagged as possibly due to survey bias by Theorem 2. Top-left: support for a policy increasing the average yearly income by GBP 5,000 but decreasing the participant's income by GBP 2,500. Top-right: support for a policy increasing the average yearly income of the lowest-income quarter GBP 5,000 but decreasing the average income of the rest by GBP 2,500. Bottom: would you cheat on your partner if given the occasion (and she/he would never find out)?

Floodlight Analysis shows that a significant effect ($p < 0.05$) would be obtained only if dropping the 34.18% fastest choices. The significance of the test of proportions would be lost by swapping 1.33% of the answers.

The second question (Figure D.9, top-right) was whether the participant would support a policy increasing the average yearly income of the lowest-income quarter of the population by GBP 5,000, while decreasing the average income of the rest of the population by GBP 2,500. The analysis was again conducted according to a median split on income. Theorem 1 reveals a preference against this policy for the poor ($p < 0.001$), but no preference is revealed for the rich ($p = 0.459$; the H function crosses the zero). A test of proportions suggest that the rich support this policy more than the poor (Rich: 56.99% vs. Poor: 48.19%, $z = 2.820$, $p = 0.005$). An application of Theorem 2, however, reveals no significant group differences ($p = 0.197$). A Floodlight Analysis shows that a significant effect ($p < 0.05$) would be obtained if dropping the 10.05% fastest choices. The significance of the test of proportions would be lost only if swapping 4.77% of the answers.

The third question (Figure D.9, bottom) was whether the participant would cheat on a partner if given the occasion (and she/he would never find out). Theorem 1 reveals a preference against cheating for both genders (both $p < 0.001$). A test of proportions

suggest that males would be more willing to cheat than females (Male: 13.29% vs. Female: 7.80%, $z = 3.1020$, $p = 0.002$). However, an application of Theorem 2 reveals no significant gender differences ($p = 0.500$). A Floodlight Analysis shows that a significant effect ($p < 0.05$) would be obtained only if dropping the 21.99% fastest choices. The significance of the test of proportions would be lost if swapping 2.22% of the answers.

D.2.2 Positive Concordance

In addition to the two questions presented in Section 5.3, for six further questions our method and a test of proportions both delivered significant results (positive concordance). We briefly present the data for these six additional questions here.

Figure D.10 presents four questions where the pre-registered group split was on political orientation (left vs. right). The top-left panel of Figure depicts support for BREXIT ("Do you think BREXIT was a good idea?"), conditional on political orientation. Theorem 1 reveals a preference in favor for Tories and a preference against for labourists (both $p < 0.001$), and Theorem 2 (as well as Corollary 1) reveal a group difference ($p < 0.001$). The latter agrees with a test of proportions (Right: 53.97% vs. Left: 14.58%, $z = 14.4969$, $p < 0.001$).

The top-right panel presents the analysis for the question "Do you think people get ahead mostly because of their own merits, or rather because of luck and help from others?" As for the previous question, Theorem 1 reveals a opposed preferences, in favor for Tories and against for labourists (both $p < 0.001$), and Theorem 2 (as well as Corollary 1) reveal a group difference ($p < 0.001$). The latter agrees with a test of proportions (Right: 58.28% vs. Left: 37.98%, $z = 6.8127$, $p < 0.0001$).

The bottom-left panel depicts support for same-sex marriage, again conditional on political orientation. Theorem 1 reveals a preference in favor both for Tories and for labourists (both $p < 0.001$), but Theorem 2 reveal a stronger preference in favor for labourists compared to Tories ($p < 0.001$). The latter agrees with a test of proportions (Right: 67.57% vs. Left: 88.96%, $z = -9.1311$, $p < 0.001$).

The bottom-right panel presents support for guaranteeing health access to "people who don't work." As for the previous question, Theorem 1 reveals a preference in favor both for Tories and for labourists (both $p < 0.001$), but Theorem 2 reveal a stronger preference in favor for labourists compared to Tories ($p = 0.002$), which agrees with a test of proportions (Right: 75.96% vs. Left: 92.12%, $z = -7.8227$, $p < 0.001$).

Figure D.11 presents the last two questions where we observed positive concordance. For these two questions, the pre-registered group split was on age (young vs. old). The left-panel depicts the support for the legalization of use and consumption of Cannabis. Theorem 1 reveals a preference in favor for the young and a preference against for the old (both $p < 0.001$), and Theorem 2 (as well as Corollary 1) reveal a group difference ($p < 0.001$). The latter agrees with a test of proportions (Old: 44.75% vs. Young: 65.54%, $z = -7.2474$, $p < 0.001$).

Finally, the right-hand panel of Figure D.11 depicts support for policies aiming to decrease greenhouse gas emissions by imposing increasing costs over time (i.e., countries should pay more in the future than now), as opposed to costs constant over time. Application of Theorem 1 reveals a preference against such policies for the old ($p < 0.001$), but no preference is revealed for the young ($p = 0.496$), as clearly seen in the picture. Application of Theorem 1 reveals a group preference, with the old opposing such policies more than the young ($p < 0.001$). This agrees with a test of proportions (Young: 49.40% vs. Old: 39.84%, $z = 3.2558, p = 0.001$).

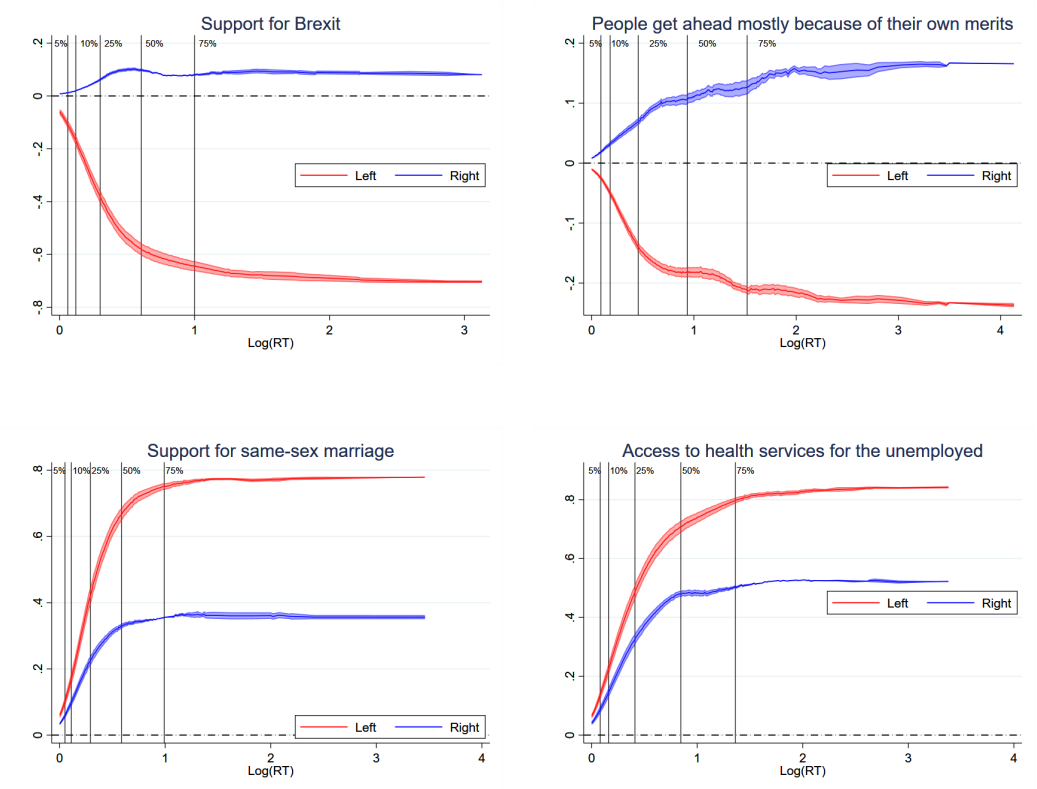

Figure D.10: Four further examples of positive concordance, based on political orientation (left vs. right). Top-left: Support for BREXIT. Top-right: Support for the view that people get ahead due to merit (as opposed to luck). Bottom-left: Support for same-sex marriage. Bottom-right: Support for health access for the unemployed.

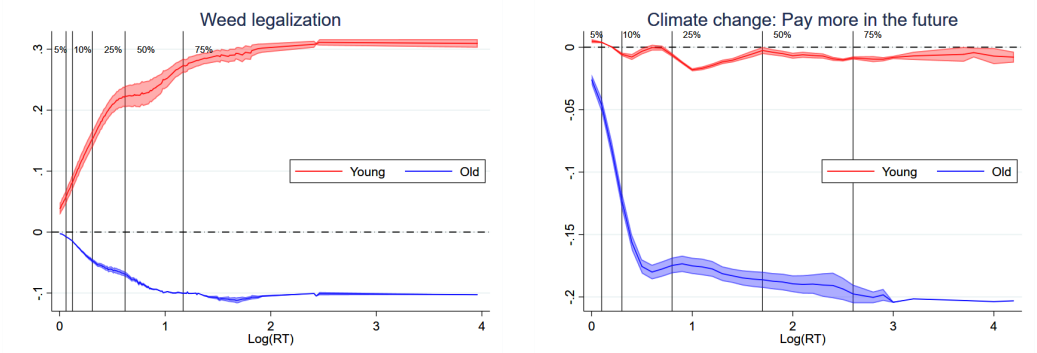

Figure D.11: Two further examples of positive concordance, based on age groups (young vs. old). Left: Support for weed legalization. Right: Support for paying more in the future to finance climate-change policies.

D.2.3 Negative Concordance

In addition to the two questions presented in Section 5.4, for three further questions our method and a test of proportions both delivered non-significant results (negative concordance). We briefly present the data for these questions here.

The top-left panel of Figure D.12 depicts the data for the question "Do you think rich countries should pay more than poor countries to finance policies decreasing greenhouse

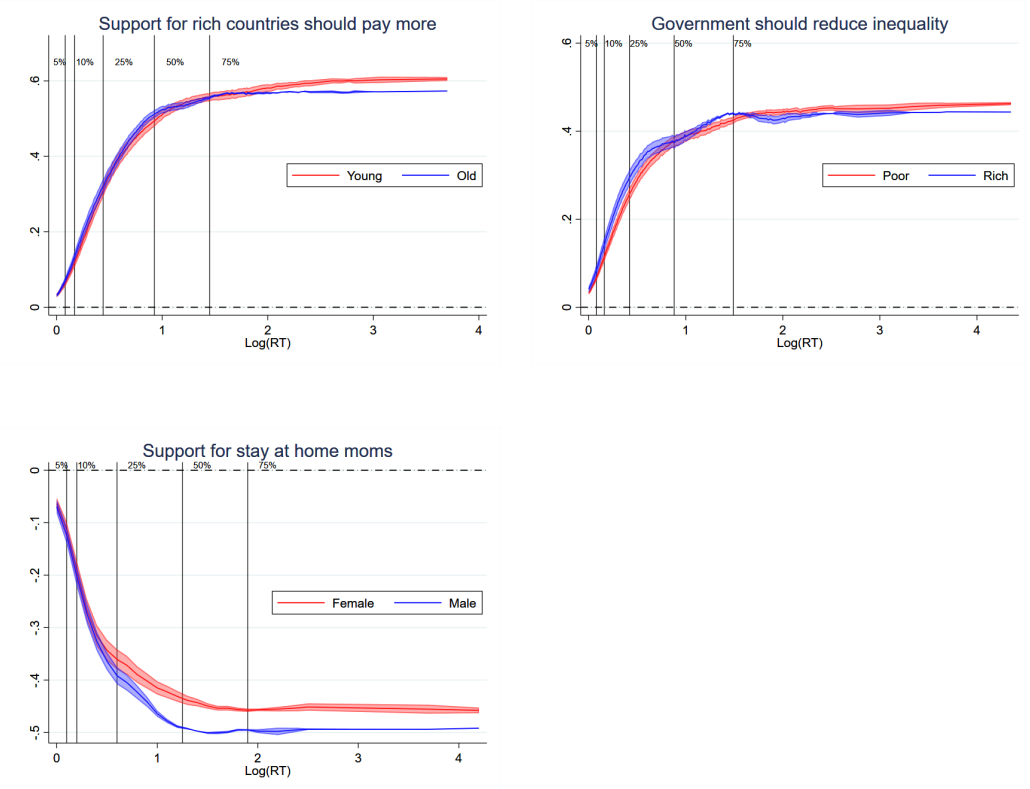

Figure D.12: Three further examples of negative concordance. Top-left: Support for rich countries paying more that poor countries to decrease greenhouse emissions. Top-right: support for the statement that reducing inequality is the responsibility of the government. Bottom: weed legalization. Right: Support for the statement that mothers should stay at home.

gas emissions (independently of the individual history of emissions)?" Applying Theorem 1 reveals a preference in favor of this policy for both young and old (median split on age; both $p < 0.001$), but Theorem 2 does not reveal a difference between the preferences of the groups ($p = 0.493$). This is also the result of a test of proportions (Old: 78.66% vs. Young: 80.35%, $z = -0.7286$, $p = 0.466$).

The top-right panel depicts the data for the question "Do you think that it is the responsibility of the government to reduce the differences in income between people with high income and those with low income?" As for the previous question, Theorem 1 reveals a preference in favor of this policy for both rich and poor (median split on income; both $p < 0.001$), but Theorem 2 does not reveal a difference between the preferences of the groups ($p = 0.498$). This is also the result of a test of proportions (Rich: 72.04% vs. Poor: 73.13%, $z = -0.3924$, $p = 0.695$).

Last, the bottom panel depicts the support for the statement that mothers should stay at home with their kids instead of working. Theorem 1 reveals a preference against this statement for both males and females (median split on gender; both $p < 0.001$), and Theorem 2 does not reveal a difference between the preferences of the groups ($p = 0.432$). Neither does a test of proportions (Female: 25.21% vs. Male: 26.99%, $z = -0.7018$, $p = 0.483$).

D.3 Stochastically Dominated Preferences: Survey 2

We applied Theorem 3 to the ten questions where Theorem 2 revealed a group difference. Section 6.2 (Figure 9) presented four of those questions. The remaining six are in Figure D.13. For all those six questions, both Theorem 2 and a test of proportions detected a significant group difference (Figure 7, Section 5.3 and Figures D.10–D.11, Section D.2.2 in the Online Appendix). The top-left panel corresponds to the question on same-sex marriage, for which Theorem 3 fails to detect a significant difference ($p = 0.464$). For the other five questions, 3 does detect the group difference (all $p < 0.001$).

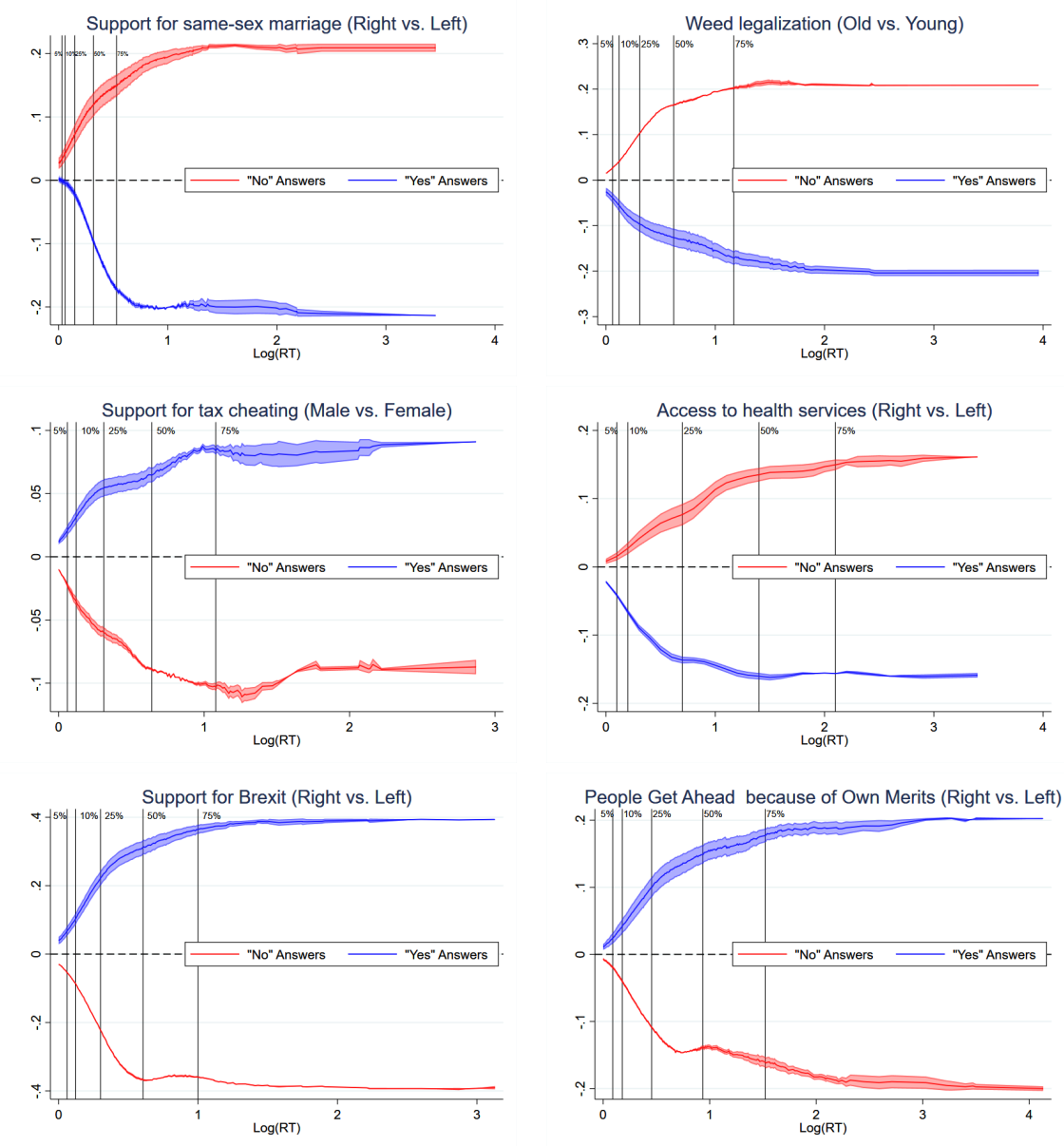

Figure D.13: Application of Theorem 3 to the six remaining questions where both Theorem 3 and a test of proportions detected a group difference.

D.4 Stochastically Dominated Preferences: Re-Analysis of the Risk Question in Liu and Netzer (2023)

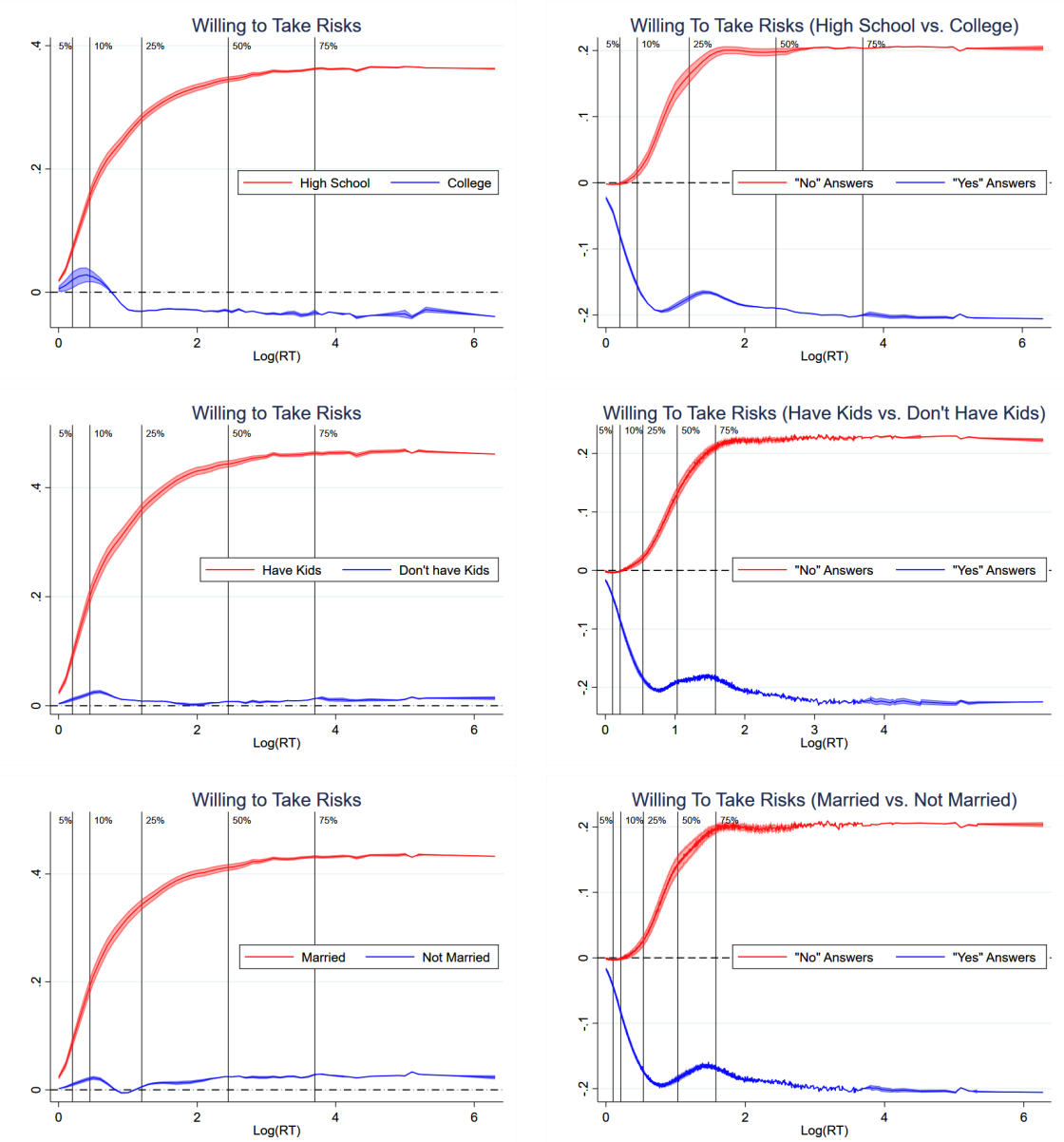

Figure D.14: Application of Theorems 2 (left column) and 3 to the willingness to take risks between different groups in Liu and Netzer (2023) data. Top: high school vs. college education. Middle: Don't have kids vs. have kids. Bottom: Not married vs. married.

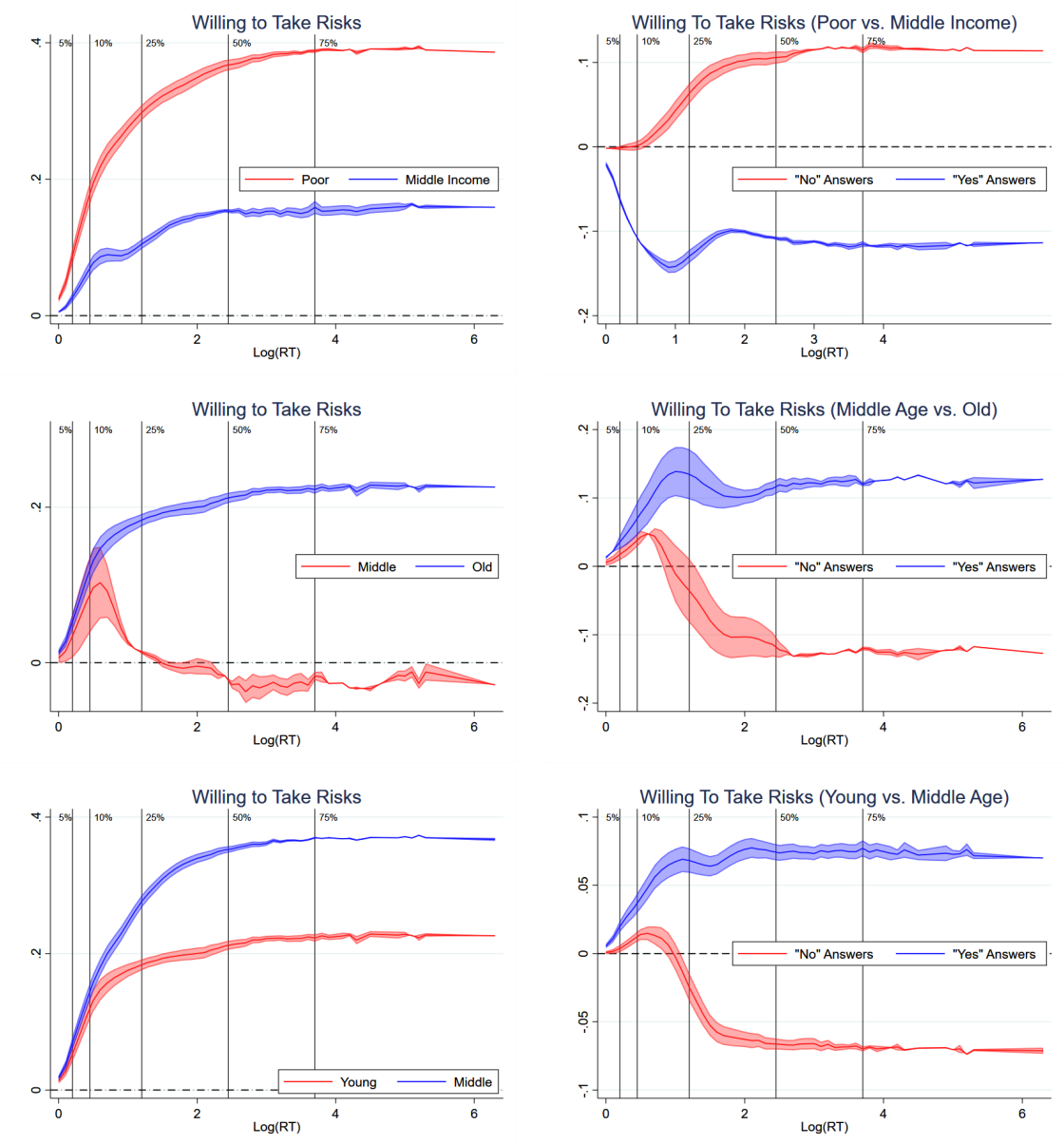

Figure D.15: Application of Theorems 2 (left column) and 3 to the willingness to take risks between different groups in Liu and Netzer (2023) data. Top: poor vs. middle income. Middle: middle age vs. old. Bottom: Young vs. middle age