Liu, Shuo; Netzer, Nick

Working Paper

Happy times: Measuring happiness using response times

Working Paper, No. 371

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Liu, Shuo; Netzer, Nick (2023) : Happy times: Measuring happiness using response times, Working Paper, No. 371, University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-194139

This Version is available at:
https://hdl.handle.net/10419/275653

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 371

Happy Times: Measuring Happiness Using Response Times

Shuo Liu and Nick Netzer

First version: December 2020
This version: March 2023

Happy Times: Measuring Happiness Using Response Times

Shuo Liu and Nick Netzer*

First version: December 2020
This version: March 2023

Abstract

Surveys that measure subjective states like happiness or preferences often generate discrete ordinal data. Ordered response models, which are commonly used to analyze such data, suffer from a fundamental identification problem. Their conclusions depend on unjustified assumptions about the distribution of a latent variable. In this paper, we propose using survey response times to solve that problem. Response times contain information about the distribution of the latent variable even among subjects who give the same survey response, through a chronometric effect. Using an online survey, we test and verify the existence of the chronometric effect. We then provide theoretical conditions under which group differences in happiness or other variables are detectable based on response time data without making distributional assumptions. In our survey, we find evidence supporting the assumptions of traditional ordered response models for some common survey questions but not for others.

Keywords: surveys, ordinal data, response times, non-parametric identification

JEL Classification: C14, D60, D91, I31

*Shuo Liu: Guanghua School of Management, Peking University, shuo.liu@gsm.pku.edu.cn. Nick Netzer: Department of Economics, University of Zurich, nick.netzer@econ.uzh.ch. This paper was previously circulated since 2020 under the title "Happy Times: Identification from Ordered Response Data". We are grateful for very helpful comments to Carlos Alós-Ferrer, Isaiah Andrews, Timothy Bond, Stefano DellaVigna, Yu Gao, Lukas Hensel, Damian Kozbur, Ian Krajbich, Xi Zhi Lim, Andrew Oswald, Rainer Winkelmann, and seminar participants at Central European University, HKBU-Kyoto-Osaka-NTU-Sinica, Jinan University, Nanjing University, Oxford University, Peking University, Renmin University of China, SFU & UBC Vancouver, Shanghai University of Finance and Economics, Shenzhen University, Tsinghua University, Wuhan University, Zhejiang University, the University of Zurich, the CESifo Area Conference on Behavioral Economics 2021, and the SSES Annual Congress 2021. Huaqing Huang and Jindi Huang provided excellent research assistance.

1 Introduction

Surveys have been an important tool in the social sciences for a long time (see Rossi et al., 1983, for a historical overview). Within economics, surveys have been used at least since Easterlin (1974) to measure happiness. The happiness literature has generated interesting insights, the most prominent one being Easterlin's paradox of a correlation between income and reported happiness within countries but not across countries or over time (but see also Stevenson and Wolfers, 2008, for contrary evidence). Recently, surveys have become popular as a tool for measuring economic preferences. For instance, Falk et al. (2018) have introduced the Global Preference Survey, which is conducted around the world and elicits individuals' preferences in different domains such as risk and time.

Surveys measuring subjective states like happiness or preferences usually generate discrete ordinal data (Likert, 1932). For example, the life happiness question in the General Social Survey (GSS, Davis and Smith, 1991) provides the three response categories "not too happy," "pretty happy," and "very happy." People responding "not too happy" are less happy than those responding "pretty happy," but there is no information by how much less. To analyze such survey data, researchers typically rely on ordered response models like ordered probit. These models assume that there is a cardinal latent variable (e.g., true happiness) which generates survey responses based on reporting thresholds (see Boes and Winkelmann, 2006). Using such models, the effect of observables on the outcome of interest can be estimated. For instance, one can compare average happiness between the rich and the poor.

In the context of happiness surveys, Bond and Lang (2019) have recently shown that almost none of the existing empirical findings are properly identified. The existing findings depend entirely on assumptions about the distribution of the latent variable in the ordered response model that is being used. Roughly speaking, since we cannot learn anything from the survey responses about the distribution of the latent variable within a given response category, making suitable assumptions about that distribution allows us to conclude almost anything. Bond and Lang (2019) indeed show that the distributions which are commonly employed in the literature (e.g., Gaussian) do not have to be twisted very much to reverse empirical findings. Plausible lognormal transformations that generate happiness distributions which resemble income or wealth distributions are sufficient to overturn standard results.[1] The observations made by Bond and Lang (2019) put at risk the entire happiness literature and threaten the emerging literature on preference surveys.

[1]Bond and Lang (2019) also discuss that the traditional models make strong assumptions in addition to specific happiness distributions, for instance that happiness is interpersonally comparable and that all survey participants employ the same reporting thresholds. The identification problem exists despite these additional assumptions.

In this paper, we argue that the use of survey response time data can help to solve the problem. Response time is the duration that a survey participant needs to answer a given question. To understand the logic of our argument, consider a happiness survey with just two response categories, "unhappy" and "happy." Suppose you answer this survey at a moment when you feel very happy. Most likely, you will find it easy to respond "happy" and you will do so quickly. Now suppose you answer the survey at a moment when you feel at best moderately satisfied. You may still end up responding "happy" but most likely it will take you longer to decide. The observable distribution of response times among the survey participants who respond to be happy then contains information about the unobservable distribution of happiness within that response category, and analogously for the "unhappy" category. Response time data can provide precisely the evidence that was missing for identification.

The idea that subjects respond faster when a stimulus is further away from an indecision threshold is not new. This *chronometric effect* has been documented in many studies in psychology, neuroscience and economics. In some of these studies, the stimulus is objective, such as the difference in brightness between two lights. Kellogg (1931) has first shown that subjects identify the brighter light faster if the difference in brightness becomes larger. The same is true in tasks where the larger of two objects has to be identified (Moyer and Bayer, 1976), or the direction of random dot motion (Palmer et al., 2005). Making the decision easier, by magnifying the stimulus away from the indecision threshold, shortens response times. In other studies, the stimulus is subjective, for example the utility difference between two options in an economic choice task. Moffatt (2005) has shown that choice between two lotteries becomes faster when the utility difference between the lotteries becomes larger. The same has been documented for intertemporal choices (Chabris et al., 2009; Konovalov and Krajbich, 2019) and choices between food items (Krajbich et al., 2010). Again, making the decision easier, by increasing the strength of preference away from the indifference point, shortens response times.[2] In the empirical part of our paper, we will later demonstrate that the chronometric effect exists in surveys as well.

In the theoretical part of the paper (Section 2), we integrate response times into a conventional ordered response model, in a way that reflects the chronometric effect. Following Bond and Lang (2019), we aim at comparing two groups (e.g., the rich and the poor) based on their responses in a survey (e.g., about happiness). The latent variable h (e.g., true happiness) follows continuous distributions in each group, but these distributions are unknown to the analyst. Individual responses are generated by reporting thresholds $\tau^1 < \tau^2 < \ldots < \tau^n$,

[2]There are many more studies documenting the chronometric effect in a variety of domains, which we cannot summarize here. See Alós-Ferrer et al. (2021) for a more detailed discussion of studies that find the chronometric effect in economic choices, and Clithero (2018b) for an excellent survey of the use of response times in economics.

which are also unknown but assumed to be the same for all survey participants. A participant with happiness $h \leq \tau^1$ responds in the lowest category 0, a participant with happiness $\tau^i < h \leq \tau^{i+1}$ responds in intermediate category i, and a participant with happiness $\tau^n < h$ responds in the highest category n. Bond and Lang (2019) have asked whether we can learn from survey response data that the happiness distribution in one group first-order stochastically dominates that in the other group. This may appear like a strong requirement, but note that first-order stochastic dominance is implicitly assumed in standard models like ordered probit. Bond and Lang (2019) show that detecting dominance is possible only under extremely stringent conditions. For instance, in a survey with two response categories, all participants in one group must respond to be happy and all participants in the other group must respond to be unhappy. If there are more than two categories, the condition is stronger than first-order stochastic dominance of the observed response distributions of the groups, and there still cannot be any responses in the lowest (highest) category from the group that is more happy (unhappy). The same conditions apply when merely asking about a ranking of average happiness between the groups, instead of first-order stochastic dominance.

Now suppose responses display the chronometric effect. As before, consider first a happiness survey with two response categories. The response time of a participant with happiness h is $c(|h-\tau^1|)$, where c is a strictly decreasing but unknown *chronometric function*, reflecting that the answer becomes easier and thus quicker for a participant when the distance $|h-\tau^1|$ between the stimulus h and the indecision threshold τ^1 becomes larger. We assume here that the chronometric function c is the same for all participants. This is analogous to the assumption of identical reporting thresholds for all participants in traditional ordered response models but, as we will show, it can be relaxed substantially. If the distribution of response times is observed in addition to the survey responses, the conditions for detecting first-order stochastic dominance of the happiness distributions become substantially weaker. Suppose the fraction of participants in group A who respond to be happy and do so at response time t or earlier, denoted $r_A^{happy}(t)$, is larger than the corresponding fraction in group B, denoted $r_B^{happy}(t)$. We can conclude that the fraction of participants with happiness $h \geq \tau^1 + c^{-1}(t)$ is larger in group A than in group B. If this holds for all t, then the participants who respond to be happy in group A are happier than in group B in the first-order stochastic dominance sense. Combined with the analogous argument for participants who respond to be unhappy, we ultimately obtain that $r_A^{happy}(t) \geq r_B^{happy}(t)$ and $r_A^{unhappy}(t) \leq r_B^{unhappy}(t)$ for all t is both necessary and sufficient for detection. These conditions are much weaker than the conditions in Bond and Lang (2019). For $t \to \infty$, they merely imply that the fraction of participants who respond to be happy must be higher in group A than in group B, and not that these fractions have to be one and zero. Our conditions are stricter than with traditional ordered

response models, because the inequalities have to hold for all response times.

We derive analogous conditions for detecting the ranking of average happiness of the groups, which turn out to be even weaker, thanks to the additional constraint on the set of admissible data-generating processes imposed by the need to explain response time data. We also show that our criterion always detects first-order stochastic dominance if it actually exists, for instance when ordered probit is the correct model. The true ranking of average happiness is always detected under some additional conditions.

When a survey has more than two response categories, chronometric effects are not straightforward in the intermediate categories. As the stimulus h varies within an interval $[\tau^i, \tau^{i+1}]$, it moves away from one indecision threshold but closer towards the other. Hence, any plausible specification of the chronometric effect generates response times that are not monotone in h between two interior reporting thresholds. As a consequence, response times from intermediate response categories are uninformative, and our detection condition coincides with that in Bond and Lang (2019) for these categories. We will show, however, that the condition can be relaxed when binary follow-up questions are included in the survey, provided that response times in the follow-up questions still exhibit the chronometric effect (which is less obvious than for an initial question and which we cannot confirm in our data). Alternatively, our results make a case for surveys with just two response categories. Due to their continuous and cardinal nature, recording response times may be more important than recording fine-grained responses.

The above arguments rest on the assumptions that the chronometric function is the same for all survey participants and that response times are measured accurately, which may not be satisfied in reality. Fortunately, these assumptions can be relaxed. We will show that our main results continue to hold if there is independent noise in response times, due to measurement error or idiosyncratic speed heterogeneity. For some of our results, we can even allow noise that is group-specific or that correlates with happiness, as long as monotonicity of the chronometric function is preserved. We will also present a robust condition for detecting the ranking of group averages despite allowing arbitrarily group-specific chronometric functions. Finally, since most surveys ask more than one question, we will formalize the idea of normalizing individual response times using the response time from a baseline question, to account for individual differences in the speed of reading, deciding, and submitting a response.

In summary, survey response times contain information that is lacking for identification of traditional ordered response models. Based on the well-established chronometric effect, the observable distribution of response times allows us to check whether latent distributions are so strongly skewed that standard results are reversed. In the words of Bond and

Lang (2019), response times may help analysts "justify their particular cardinalization or parametric assumption relative to other plausible alternatives" (p. 1639).

Our theoretical analysis is related to a recent paper by Alós-Ferrer et al. (2021), which studies the problem of eliciting preferences from choice data when choice is stochastic. While surprising at first glance, the problem with ordered response models is similar to the revealed preference problem in random utility models. In the latter, the utility difference between two choice options of an agent is an unobserved random variable which generates stochastic choices. Without assumptions on its distribution (e.g., logistic in a Luce model) it is not possible to deduce the agent's underlying deterministic utility function from observed choices. Alós-Ferrer et al. (2021) propose using response time data to solve that problem, exploiting the chronometric effect. Our methodology also relies on the chronometric effect, but our questions and results are different from Alós-Ferrer et al. (2021). Most importantly, revealed preference questions are questions about the properties of a single distribution (of the utility difference between the choice options). The detection questions considered in this paper are questions about the comparison of two distributions (of the latent variable in two groups).

In the empirical part of the paper (Section 3), we report results from an online survey with about 8,000 participants that we conducted on Amazon Mechanical Turk (MTurk). We asked several socio-demographic questions and several substantive questions about happiness, preferences, trust, and political attitudes. These questions were adopted from the GSS and from Falk et al. (2018). We implemented two versions of the survey, one with two answer categories and one with three answer categories. In both versions of the survey, each substantive question was accompanied by a follow-up question in which participants were asked to refine their previous answer. For example, a subject giving the highest possible response "rather happy" in the initial question about overall life happiness subsequently had the choice between "very happy" and "moderately happy" in the follow-up question, and a subject giving the intermediate response "neither happy nor unhappy" in the survey version with three categories subsequently had the choice between "tend more toward happy" and "tend more toward unhappy."

Conducting the survey online makes it easy to record response times, which we define as the time between the display of the question and the moment when the participant clicked on her answer. To account for individual heterogeneity in response speed, we follow our theoretical analysis and normalize the raw response times by subtracting (in logs) each subject's response time in the socio-demographic question about marital status, where there are arguably no uncertainties or varying intensities about the correct answer, and which was also answered quickest on average.

We first use responses from the follow-up questions to test for the existence of a chrono-

metric effect in surveys. We find that, among subjects who initially gave an identical answer, those who reveal a more extreme position in the follow-up question responded faster on average in the initial question. More specifically, we consider all subjects who responded in the same extreme category in an initial question (e.g. "rather happy") and partition them into two subgroups based on their response in the follow-up question. Those who give a more extreme response in the follow-up (e.g. "very happy") have larger values of the latent variable than those who give a more moderate response (e.g. "moderately happy"). The chronometric effect predicts that the former responded more quickly on average in the initial question than the latter. We find this prediction confirmed in our data, for both extreme response categories in all seven substantive questions and both versions of the survey. Among the 28 pairwise comparisons that we make, 25 are statistically significant at the 1% level. We further confirm in pooled regression analyses that giving a more extreme response in the follow-up question is associated with significantly quicker responses in the initial question, even when controlling for demographics or individual fixed or random effects. In other words, subjects for whom the latent variable is further away from the respective reporting threshold give a quicker response.

Having confirmed the chronometric effect, we proceed to check our detection conditions in the data. We compare different socio-demographic groups and, for each substantive question, check whether it is possible to detect a difference between the groups, for example whether rich participants are happier than poor participants, or whether the old are more risk-averse than the young. Our goal is not to make claims about causality, but rather to show how our techniques can be applied and to get a first impression whether the distributional assumptions made in traditional models will be confirmed or rejected. Our detection conditions can easily be visualized by plotting the evolution of response fractions over response time. To make statistical inference, we exploit the similarity between our novel detection conditions and the standard conditions for stochastic dominance, which leads to a bootstrap-based test adapted from the seminal work by Barrett and Donald (2003). Our paper is accompanied by Stata ado-files which implement these procedures for general surveys.

Using the binary version of the survey, our findings reveal systematic patterns across questions. For the questions about risk attitudes and about satisfaction with work and social life, we can detect the ranking of averages with high statistical confidence in all pairwise comparisons that we make, and very often we can detect first-order stochastic dominance. We interpret this as first cautious evidence that the latent variables for these questions follow distributions that satisfy assumptions in line with traditional ordered response models. On the other extreme, detection is often not possible with high confidence for the question about time preferences. This suggests that the distribution of discount rates may be less regular

than what is postulated by traditional ordered response models. The questions about overall life happiness, trust, and political attitudes are somewhere in between, with detection being possible in some pairwise comparisons but not in others.

In the trinary version of the survey, the part of the detection condition that applies to the intermediate response category is always violated. Since our response-time-based conditions coincide with the conventional ones for intermediate categories, we are unable to achieve any detection in the trinary survey even with response time data. Overall, our empirical findings indeed support the idea that surveys with just two response categories may be preferable to surveys with multiple categories, provided that response time data can be recorded.

The paper is organized as follows. Section 2 presents our theoretical results. Section 3 reports the empirical results from our survey. A more in-depth literature discussion can be found in Section 4. Section 5 concludes. Omitted proofs, the complete questionnaires of our survey experiment, and some additional empirical results are in the Online Appendix.

2 Theory

2.1 Ordered Response Model

Consider two groups $j = A, B$ of individuals. The distribution of happiness $\tilde{h}$ within group j is described by a cumulative distribution function $G_j : \mathbb{R} \to [0, 1]$, which is assumed to be continuous and to have a well-defined expected value

$$\mathbb{E}_{G_j}[\tilde{h}] = \int_{\mathbb{R}} h \, dG_j(h).$$

A data analyst does not observe individual happiness but observes only the individuals' survey responses on a finite ordered scale, with categories labelled $i = 0, \ldots, n$ for some $n \geq 1$. The latent variable $\tilde{h}$ generates responses through reporting thresholds $\tau = (\tau^1, \tau^2, \ldots, \tau^n) \in \mathbb{R}^n$, which satisfy $\tau^1 < \tau^2 < \ldots < \tau^n$ and for now are assumed to be the same for all individuals in both groups. We will discuss the case of heterogeneity in Section 2.3. An individual with happiness h responds in category i when $\tau^i < h \leq \tau^{i+1}$. This is applicable also to categories $i = 0, n$ with the convention that $\tau^0 = -\infty$ and $\tau^{n+1} = +\infty$. Hence, the fraction of individuals within group j who respond in category i is given by

$$r_j^i = G_j(\tau^{i+1}) - G_j(\tau^i). \tag{1}$$

This is again applicable also to $i = 0, n$ with the convention $G_j(-\infty) = 0$ and $G_j(+\infty) = 1$.

Given ordered response data $r_j = (r_j^0, r_j^1, \ldots, r_j^n) \in [0, 1]^{n+1}$ with $\sum_{i=0}^n r_j^i = 1$, the analyst

would like to learn about properties of the underlying distributions G_j. In particular, she is interested in comparing the happiness between the two groups. The following definition formalizes the idea of non-parametric detection of first-order stochastic dominance.

Definition 1. Given (r_A, r_B), group A is *detectably rank-order happier* than group B if

$$G_A(h) \leq G_B(h) \text{ for all } h \in \mathbb{R},$$

for all (G_A, G_B, τ) that satisfy (1) for $i = 0, \ldots, n$ and $j = A, B$.

Rank-order detection requires G_A to weakly first-order stochastically dominate G_B, written G_A FOSD G_B, for all pairs of happiness distributions and reporting thresholds that could have generated the observed survey data. This is a strong requirement, but note that first-order stochastic dominance is implicitly assumed in applications of, e.g., the classical ordered probit model. A conceptually weaker requirement is the following.

Definition 2. Given (r_A, r_B), group A is *detectably on-average happier* than group B if

$$\mathbb{E}_{G_A}[\tilde{h}] \geq \mathbb{E}_{G_B}[\tilde{h}],$$

for all (G_A, G_B, τ) that satisfy (1) for $i = 0, \ldots, n$ and $j = A, B$.

In words, on-average detection only requires the average happiness to be weakly larger in group A than in group B, but again for all pairs of happiness distributions and reporting thresholds that could have generated the data.

Recall the well-known fact that FOSD is equivalent to $\mathbb{E}_{G_A}[q(\tilde{h})] \geq \mathbb{E}_{G_B}[q(\tilde{h})]$ for all *weakly increasing* functions $q : \mathbb{R} \to \mathbb{R}$ (see e.g. Hanock and Levy, 1969). Hence, the definition of rank-order detection could be rephrased accordingly. As we show in Appendix A.1, the definition of on-average detection is equivalent to the requirement that $\mathbb{E}_{G_A}[q(\tilde{h})] \geq \mathbb{E}_{G_B}[q(\tilde{h})]$ for all *strictly increasing* functions $q : \mathbb{R} \to \mathbb{R}$. This is because we do not restrict the class of admissible distributions beyond the property of continuity, so that for any distributions (G_A, G_B) and any strictly increasing q, the induced distributions $(\hat{G}_A, \hat{G}_B)$ of happiness under transformation q are also admissible distributions that could have generated the same data. Hence, on-average detection implies a ranking of averages no matter which strictly increasing "cardinalization" (Bond and Lang, 2019, p. 1630) we choose to transform the scale of happiness.

We now state a first result about rank-order detection. This result is not new (see Bond and Lang, 2019, and the discussion therein) and we include a proof only for completeness and later reference.

Proposition 1. *Given (r_A, r_B), group A is detectably rank-order happier than group B if and only if*

(i) $r_A^0 = 0$,

(ii) $r_B^n = 0$, *and*

(iii) $\sum_{i=0}^{k} r_A^i \leq \sum_{i=0}^{k-1} r_B^i$ *for all* $k = 1, \ldots, n-1$.

Proof. If-statement. Let (G_A, G_B, τ) satisfy (1) for $i = 0, \ldots, n$ and $j = A, B$. It follows that $G_j(\tau^{i+1}) = G_j(\tau^i) + r_j^i$. Hence, for any $k = 0, \ldots, n$ and $h \in (\tau^k, \tau^{k+1}]$ we obtain

$$G_A(h) \leq G_A(\tau^{k+1}) = G_A(\tau^k) + r_A^k = G_A(\tau^{k-1}) + r_A^{k-1} + r_A^k = \ldots = \sum_{i=0}^{k} r_A^i, \text{ and}$$

$$G_B(h) \geq G_B(\tau^k) = G_B(\tau^{k-1}) + r_B^{k-1} = G_B(\tau^{k-2}) + r_B^{k-2} + r_B^{k-1} = \ldots = \sum_{i=0}^{k-1} r_B^i.$$

Conditions $(i) - (iii)$ thus imply $G_A(h) \leq G_B(h)$ for all $h \in \mathbb{R}$.

Only-if-statement. Suppose at least one of conditions $(i) - (iii)$ is violated. Suppose first that there exists $k^* = 1, \ldots, n-1$ for which $\sum_{i=0}^{k^*} r_A^i > \sum_{i=0}^{k^*-1} r_B^i$. Therefore, any (G_A, G_B, τ) that satisfies (1) for $i = 0, \ldots, n$ and $j = A, B$ must have $G_A(\tau^{k^*+1}) > G_B(\tau^{k^*})$. Starting from any such (G_A, G_B, τ), construct $(\hat{G}_A, \hat{G}_B, \tau)$ by setting $\hat{G}_j(h) = G_j(h)$ for all $h \notin (\tau^{k^*}, \tau^{k^*+1})$. For $h \in (\tau^{k^*}, \tau^{k^*+1})$, let $\hat{G}_A(h) = \hat{G}_A(\tau^{k^*+1})$ when $h \geq \tau^* := (\tau^{k^*} + \tau^{k^*+1})/2$, and $\hat{G}_B(h) = \hat{G}_B(\tau^{k^*})$ when $h \leq \tau^*$. Complete the construction of each $\hat{G}_j$ in an arbitrary increasing and continuous way. It follows that $(\hat{G}_A, \hat{G}_B, \tau)$ satisfies (1) for $i = 0, \ldots, n$ and $j = A, B$, and

$$\hat{G}_A(\tau^*) = \hat{G}_A(\tau^{k^*+1}) = G_A(\tau^{k^*+1}) = \sum_{i=0}^{k^*} r_A^i > \sum_{i=0}^{k^*-1} r_B^i = G_B(\tau^{k^*}) = \hat{G}_B(\tau^{k^*}) = \hat{G}_B(\tau^*),$$

so that $\hat{G}_A$ FOSD $\hat{G}_B$ is not true. The case where $r_A^0 > 0$ is immediate, because it is always possible to shift the probability mass $G_A(\tau^1) > 0$ in G_A to the left to obtain a contradiction to FOSD, and analogously when $r_B^n > 0$. □

Conditions (i) – (iii) apply for any number n of categories, whether small or large. They are essentially never satisfied in real-world data, as demonstrated by Bond and Lang (2019). We obtain a particularly striking corollary for the binary response case.

Corollary 1. *Given (r_A, r_B) for $n = 1$, group A is detectably rank-order happier than group B if and only if $r_A^0 = r_B^1 = 0$.*

One may wonder whether the issue can be solved by requiring only the detection of the ranking of the averages. Unfortunately, the on-average notion of detection is not more admissible than the rank-order notion.[3]

Proposition 2. *Given (r_A, r_B), group A is detectably on-average happier than group B if and only if group A is detectably rank-order happier than group B.*

Proof. If-statement. This follows because G_A FOSD G_B implies $\mathbb{E}_{G_A}[\tilde{h}] \geq \mathbb{E}_{G_B}[\tilde{h}]$.

Only-if-statement. Suppose group A is not detectably rank-order happier than group B, so there exists (G_A, G_B, τ) that satisfies (1) for $i = 0, \ldots, n$ and $j = A, B$ and

$$\mathbb{E}_{G_A}[q(\tilde{h})] < \mathbb{E}_{G_B}[q(\tilde{h})]$$

for some weakly increasing function $q : \mathbb{R} \to \mathbb{R}$. Now consider the function $\hat{q} : \mathbb{R} \to \mathbb{R}$ defined by $\hat{q}(h) = q(h) + \epsilon h$ for some $\epsilon > 0$, which is strictly increasing. We obtain

$$\mathbb{E}_{G_A}[\hat{q}(\tilde{h})] = \mathbb{E}_{G_A}[q(\tilde{h})] + \epsilon\,\mathbb{E}_{G_A}[\tilde{h}] < \mathbb{E}_{G_B}[q(\tilde{h})] + \epsilon\,\mathbb{E}_{G_B}[\tilde{h}] = \mathbb{E}_{G_B}[\hat{q}(\tilde{h})]$$

for sufficiently small $\epsilon > 0$. Hence group A is not detectably on-average happier than group B either (see Appendix A.1). □

To better understand this result, observe that any violation of the conditions (i) – (iii) in Proposition 1 makes it possible to construct a data-generating process for which the average happiness is higher in group B than in group A. This is obvious for the conditions concerning the extreme categories, but it can also be shown for the condition concerning the intermediate categories. Response data alone are not able to rule out such constructions.

2.2 Ordered Response Model with Response Times

Assume the analyst also measures the speed of the individuals' survey responses. Denote the smallest and largest possible response times by $\underline{t}$ and $\overline{t}$, respectively, with $\underline{t} < \overline{t}$ (where we allow for $\underline{t} = 0$ and $\overline{t} = +\infty$). Response times are related to the latent variable $\tilde{h}$ through a chronometric function $c : \mathbb{R}_{++} \to [\underline{t}, \overline{t})$. This function is assumed to be continuous, strictly decreasing in δ whenever $c(\delta) > \underline{t}$, and to satisfy $\lim_{\delta \to 0} c(\delta) = \overline{t}$ and $\lim_{\delta \to +\infty} c(\delta) = \underline{t}$. The chronometric function is for now assumed to be the same for all individuals in both groups, analogous to the assumption of identical reporting thresholds. We will demonstrate in Section 2.3 how this assumption can be relaxed.

[3]The literature seems to know that rank-order and on-average detection require the same conditions, but we are not aware of a formal clarification that this is true. We therefore state the result and give a short proof. Importantly, an analogous result no longer holds when we consider the case with response times.

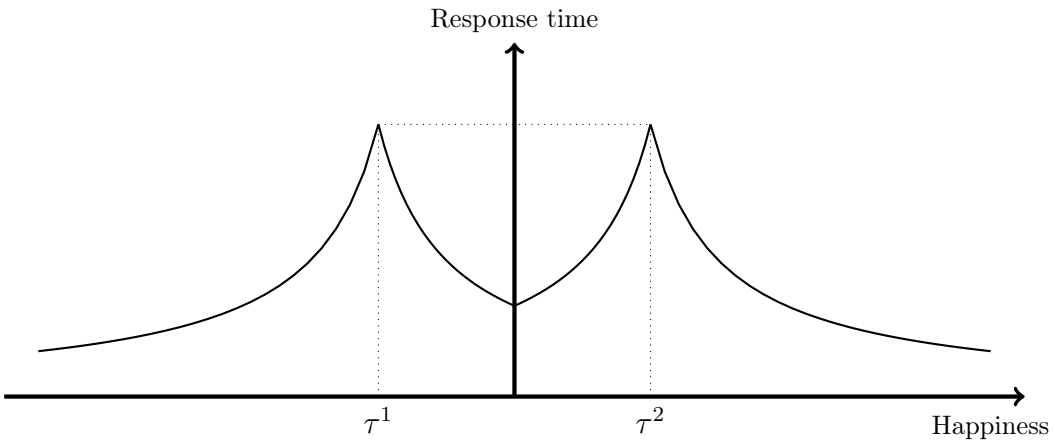


FIGURE 1: Example of response times with $n = 2$, $\tau^1 = -2, \tau^2 = 2$, and $c(\delta) = 1/(\delta + 1)$.

To understand how response times are generated, consider the binary case ($n = 1$) first. An individual with happiness $h < \tau^1$ responds in category $i = 0$ at response time $c(\tau^1 - h)$. This reflects the idea that a happiness level closer to the reporting threshold means that the individual finds it more difficult to determine whether the appropriate response category is $i = 0$ ("unhappy") or $i = 1$ ("happy"), resulting in a longer response time. Similarly, an individual with happiness $h > \tau^1$ responds in category $i = 1$ at response time $c(h - \tau^1)$. Note that this approach attaches cardinal meaning to happiness h, but since we do not restrict the set of distributions G_j and chronometric functions c, we implicitly allow for all possible cardinalizations (see the related arguments in the previous section).[4]

There are various ways how the chronometric effect could be modelled for intermediate response categories $i = 1, \ldots, n-1$ when $n \geq 2$. In the following, we adopt a simple symmetric specification where response time is driven by the distance between happiness and the closest reporting threshold. Our results are robust to various other specifications, which we will discuss later. Thus, an individual with happiness h exhibits a response time of $c(\min_i |h - \tau^i|)$. This formulation implicitly assumes $h \neq \tau^i$ for all $i = 1, \ldots, n$. Since $\tilde{h}$ follows a continuous distribution, we do not need to specify the response time of individuals with $h = \tau^i$, but we could set it to $\bar{t}$ (whenever finite). Figure 1 depicts an example of response times arising from a data-generating process that satisfies all our requirements.

In summary, among the individuals of group j who respond in category i, provided that

[4]This way of adding a chronometric function to an ordered response model is analogous to how Alós-Ferrer et al. (2021) add a chronometric function to a random utility model. They consider binary choice problems and assume that response time is monotonically driven by the absolute realized value of the utility difference between the two options.

they exist, the fraction responding at time $t \in (\underline{t}, \bar{t})$ or earlier is

$$F_j^i(t) = \frac{\max\{0, G_j(\tau^{i+1} - c^{-1}(t)) - G_j(\tau^i + c^{-1}(t))\}}{G_j(\tau^{i+1}) - G_j(\tau^i)}. \tag{2}$$

The maximum operator is required because too small response times t, for which $c^{-1}(t) > (\tau^{i+1} - \tau^i)/2$, cannot arise in category i, with our present specification.

Given data on responses $r_j = (r_j^0, r_j^1, \ldots, r_j^n)$ and response times $F_j = (F_j^0, F_j^1, \ldots, F_j^n)$, where each cumulative distribution function F_j^i is assumed to be continuous and to satisfy $F_j^i(\underline{t}) = 0$ and $F_j^i(\bar{t}) = 1$,[5] the analyst can ask the previous questions. At the risk of creating redundancy, we state these again as formal definitions.

Definition 3. Given (r_A, r_B, F_A, F_B), group A is *detectably rank-order happier* than group B if

$$G_A(h) \leq G_B(h) \text{ for all } h \in \mathbb{R},$$

for all (G_A, G_B, τ, c) that satisfy (1) and (2) for $i = 0, \ldots, n$, $j = A, B$, and all $t \in (\underline{t}, \bar{t})$.

Definition 4. Given (r_A, r_B, F_A, F_B), group A is *detectably on-average happier* than group B if

$$\mathbb{E}_{G_A}[\tilde{h}] \geq \mathbb{E}_{G_B}[\tilde{h}],$$

for all (G_A, G_B, τ, c) that satisfy (1) and (2) for $i = 0, \ldots, n$, $j = A, B$, and all $t \in (\underline{t}, \bar{t})$.

We start with a characterization of rank-order detection using response times, which is the first main result of our paper.

Proposition 3. *Given (r_A, r_B, F_A, F_B), group A is detectably rank-order happier than group B if and only if*

(i) $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq 0$ *for all* $t \in (\underline{t}, \bar{t})$,

(ii) $r_A^n F_A^n(t) - r_B^n F_B^n(t) \geq 0$ *for all* $t \in (\underline{t}, \bar{t})$, *and*

(iii) $\sum_{i=0}^{k} r_A^i \leq \sum_{i=0}^{k-1} r_B^i$ *for all* $k = 1, \ldots, n-1$.

Proof. If-statement. Let (G_A, G_B, τ, c) satisfy (1) and (2) for $i = 0, \ldots, n$, $j = A, B$, and all $t \in (\underline{t}, \bar{t})$. For $i = 0$ this implies

$$r_j^0 F_j^0(t) = G_j(\tau^1 - c^{-1}(t))$$

[5] If $r_j^i = 0$, we can specify F_j^i to be an arbitrary cumulative distribution function with these properties.

for all $t \in (\underline{t}, \bar{t})$. Thus, condition (i) implies $G_A(\tau^1 - c^{-1}(t)) \leq G_B(\tau^1 - c^{-1}(t))$ for all $t \in (\underline{t}, \bar{t})$. We claim that this implies $G_A(h) \leq G_B(h)$ for all $h \leq \tau^1$. This is immediate for any h for which there exists $t \in (\underline{t}, \bar{t})$ such that $h = \tau^1 - c^{-1}(t)$. For $h = \tau^1$ it follows from continuity of G_j. For any h with $c(\tau^1 - h) = \underline{t}$ it follows because $G_j(h) = 0$ in that case, as there is no atom at response time $\underline{t}$. By an analogous argument, condition (ii) implies $G_A(h) \leq G_B(h)$ for all $h > \tau^n$. The proof that condition (iii) implies $G_A(h) \leq G_B(h)$ for $\tau^1 < h \leq \tau^n$ is exactly like in the proof of Proposition 1.

Only-if-statement. Suppose at least one of conditions $(i) - (iii)$ is violated. Suppose first that $r_A^0 F_A^0(t^*) - r_B^0 F_B^0(t^*) > 0$ for some $t^* \in (\underline{t}, \bar{t})$. Any (G_A, G_B, τ, c) that satisfies (1) and (2) for $i = 0, \ldots, n$, $j = A, B$, and all $t \in (\underline{t}, \bar{t})$ must then have $G_A(\tau^1 - c^{-1}(t^*)) > G_B(\tau^1 - c^{-1}(t^*))$, so that G_A FOSD G_B is not true. An analogous argument applies when $r_A^n F_A^n(t^*) - r_B^n F_B^n(t^*) < 0$ for some $t^* \in (\underline{t}, \bar{t})$. Finally, suppose that there exists $k^* = 1, \ldots, n-1$ for which $\sum_{i=0}^{k^*} r_A^i > \sum_{i=0}^{k^*-1} r_B^i$. Starting from any (G_A, G_B, τ, c) that generates the data, we then construct $\hat{G}_j$ exactly like in the proof of Proposition 1. However, here we complete $\hat{G}_A$ for $h \in (\tau^{k^*}, \tau^*)$, where $\tau^* := (\tau^{k^*} + \tau^{k^*+1})/2$, in a specific way:

$$\hat{G}_A(\tau^{k^*} + z) = G_A(\tau^{k^*} + z) + G_A(\tau^{k^*+1}) - G_A(\tau^{k^*+1} - z)$$

for all $z \in (0, (\tau^{k^*+1} - \tau^{k^*})/2)$. It is easy to see that this construction yields a continuous and non-decreasing $\hat{G}_A$. It also follows that $\hat{G}_A$ generates $F_A^{k^*}$, because

$$\hat{G}_A(\tau^{k^*+1} - z) - \hat{G}_A(\tau^{k^*} + z) = G_A(\tau^{k^*+1} - z) - G_A(\tau^{k^*} + z)$$

for all $z \in (0, (\tau^{k^*+1} - \tau^{k^*})/2)$, and since G_A satisfies (2) for $i = k^*$ and all $t \in (\underline{t}, \bar{t})$, so does $\hat{G}_A$. Similarly, we can complete $\hat{G}_B$ for $h \in (\tau^*, \tau^{k^*+1})$ to generate the distribution $F_B^{k^*}$. It then follows that $(\hat{G}_A, \hat{G}_B, \tau, c)$ satisfies (1) and (2) for $i = 0, \ldots, n$, $j = A, B$, and all $t \in (\underline{t}, \bar{t})$, but $\hat{G}_A$ FOSD $\hat{G}_B$ is not true. □

Remarkably, the previous strong requirements $r_A^0 = 0$ and $r_B^n = 0$ in Proposition 1 are now replaced by weaker conditions (i) and (ii) that rely on response times. For $t \to \bar{t}$, these conditions imply $r_A^0 \leq r_B^0$ and $r_A^n \geq r_B^n$, which means that the fraction of responses in the lowest category must be smaller in group A than in group B, and conversely for the highest category. More generally, the conditions require that this must also hold when considering only those responses that took a response time of t or less, for all t. Intuitively, there must be fewer and slower "most unhappy" responses in group A than in group B, and conversely for the "most happy" responses. By contrast, condition (iii) is unaffected by the availability of response time data. Intuitively, since response times are not monotone

between two reporting thresholds, as illustrated in Figure 1, response times are uninformative in intermediate response categories. Nevertheless, we will argue later in Section 2.4 that condition (*iii*) can be weakened if we include simple binary follow-up questions in the survey.

The power of our weaker conditions becomes apparent when considering the case of binary survey responses, as we summarize in the following corollary.

Corollary 2. *Given (r_A, r_B, F_A, F_B) for $n = 1$, group A is detectably rank-order happier than group B if and only if $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq 0 \leq r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ for all $t \in (\underline{t}, \bar{t})$.*

In fact, if one is interested in rank-order detection, there is no need to consider surveys with more than two response categories, as the next result shows.

Proposition 4. *Suppose that the true happiness distribution of group A first-order stochastically dominates that of group B. For $n = 1$, the generated data (r_A, r_B, F_A, F_B) then satisfy that $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq 0 \leq r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ for all $t \in (\underline{t}, \bar{t})$.*

Proof. Suppose G_A FOSD G_B and consider a survey with $n = 1$. Let τ^1 and c be the reporting threshold and the chronometric function of the true data-generating process. Then it follows that

$$r_j^0 F_j^0(t) = G_j(\tau^1 - c^{-1}(t)) \text{ and } r_j^1 F_j^1(t) = 1 - G_j(\tau^1 + c^{-1}(t)),$$

for $j = A, B$ and all $t \in (\underline{t}, \bar{t})$. Since $G_A(h) \leq G_B(h)$ for all $h \in \mathbb{R}$, we obtain

$$r_A^0 F_A^0(t) - r_B^0 F_B^0(t) = G_A(\tau^1 - c^{-1}(t)) - G_B(\tau^1 - c^{-1}(t)) \leq 0$$

and

$$r_A^1 F_A^1(t) - r_B^1 F_B^1(t) = G_B(\tau^1 + c^{-1}(t)) - G_A(\tau^1 + c^{-1}(t)) \geq 0,$$

for all $t \in (\underline{t}, \bar{t})$. □

In words, whenever the true distributions of happiness of the two groups can be ranked according to first-order stochastic dominance, as assumed e.g. in the ordered probit model, our techniques applied to binary survey data will detect the dominance relation. Conversely, Proposition 4 also implies that if the conditions for rank-order detection are violated in a binary survey, then the true happiness distributions cannot exhibit a first-order stochastic dominance relation. This is stronger than the "only if" statement of Proposition 3 for the binary case. When our detection condition is violated, from Proposition 3 we learn that *at least one* possible data-generating process must violate first-order stochastic dominance,

while from Proposition 4 we learn that *all* possible data-generating processes must violate first-order stochastic dominance.[6]

Proposition 4 cannot be generalized to the case where $n > 1$. It is possible that the true distributions exhibit a first-order stochastic dominance relation but condition (iii) in Proposition 3 is not satisfied, for example if G_A and G_B coincide on $[\tau^1, \tau^n]$. In that sense, a binary survey is more informative than a multi-category one when response times are available.

The next proposition, which is the second main result of our paper, gives a weaker sufficient condition for on-average detection, which is implied by but does not imply the previous condition in Proposition 3. This shows that the on-average notion of detection is indeed weaker than the rank-order notion when response times are being used.[7]

Proposition 5. *Given (r_A, r_B, F_A, F_B), group A is detectably on-average happier than group B if*

(i) $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq r_A^n F_A^n(t) - r_B^n F_B^n(t)$ *for all* $t \in (\underline{t}, \bar{t})$, *and*

(ii) $\sum_{i=0}^{k} r_A^i \leq \sum_{i=0}^{k-1} r_B^i$ *for all* $k = 1, \ldots, n-1$.

Proof. Let (G_A, G_B, τ, c) satisfy (1) and (2) for $i = 0, ..., n$, $j = A, B$, and all $t \in (\underline{t}, \bar{t})$. Condition (i) implies that

$$G_A(\tau^1 - c^{-1}(t)) - G_B(\tau^1 - c^{-1}(t)) \leq [1 - G_A(\tau^n + c^{-1}(t))] - [1 - G_B(\tau^n + c^{-1}(t))],$$

for all $t \in (\underline{t}, \bar{t})$. Arguing like in the proof of Proposition 3, this implies that

$$G_B(\tau^n + h) + G_B(\tau^1 - h) - G_A(\tau^n + h) - G_A(\tau^1 - h) \geq 0 \tag{3}$$

for all $h \geq 0$. In addition, exactly like in the proof of Proposition 1, condition (ii) implies

$$G_A(h) \leq G_B(h) \tag{4}$$

[6]The data even allow us to pin down the percentiles for which the dominance relation does not hold. For instance, suppose that $r_A^0 F_A^0(t) > r_B^0 F_B^0(t)$ for some $t > 0$. Since any (G_A, G_B, τ, c) that could have generated the data must satisfy $r_j^0 F_j^0(t) = G_j(\tau^1 - c^{-1}(t))$, we can conclude that the $r_B^0 F_B^0(t)$-percentile of G_A must be strictly lower than that of G_B.

[7]Intuitively, the additional requirement (2) for response times implies that not all distributions $(\hat{G}_A, \hat{G}_B)$ which are obtained by monotonically transforming some distributions (G_A, G_B) that could have generated the data are themselves admissible data-generating processes. Therefore, in contrast to the case without response times, there are constraints on the ranking of expected values even when the conditions for rank-order detection are violated.

for all $h \in (\tau^1, \tau^n]$. Therefore, using the fact that

$$\mathbb{E}_G[\tilde{h}] = -\int_{-\infty}^{0} G(h)dh + \int_{0}^{+\infty} [1 - G(h)]dh,$$

we have

$$\begin{aligned}
&\mathbb{E}_{G_A}[\tilde{h}] - \mathbb{E}_{G_B}[\tilde{h}] \\
&= \int_{-\infty}^{0} [G_B(h) - G_A(h)]dh + \int_{0}^{+\infty} [1 - G_A(h) - 1 + G_B(h)]dh \\
&= \int_{-\infty}^{+\infty} [G_B(h) - G_A(h)]\, dh \\
&= \sum_{k=0}^{n} \int_{\tau^k}^{\tau^{k+1}} [G_B(h) - G_A(h)]\, dh \\
&\geq \int_{\tau^n}^{+\infty} [G_B(h) - G_A(h)]\, dh + \int_{-\infty}^{\tau^1} [G_B(h) - G_A(h)]\, dh \\
&= \int_{0}^{+\infty} [G_B(\tau^n + h) - G_A(\tau^n + h)]\, dh + \int_{-\infty}^{0} \left[G_B(\tau^1 + h) - G_A(\tau^1 + h)\right] dh \\
&= \int_{0}^{+\infty} [G_B(\tau^n + h) - G_A(\tau^n + h)]\, dh + \int_{0}^{+\infty} \left[G_B(\tau^1 - h) - G_A(\tau^1 - h)\right] dh \\
&= \int_{0}^{+\infty} \left[G_B(\tau^n + h) + G_B(\tau^1 - h) - G_A(\tau^n + h) - G_A(\tau^1 - h)\right] dh \\
&\geq 0,
\end{aligned}$$

where the first inequality follows from (4) and the second inequality follows from (3). □

There is no difference between these new and the previous conditions when it comes to the intermediate response categories. Hence, the difference between rank-order and on-average detection is most evident for the case of binary surveys.

Corollary 3. *Given (r_A, r_B, F_A, F_B) for $n = 1$, group A is detectably on-average happier than group B if $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ for all $t \in (\underline{t}, \bar{t})$.*

For $t \to \bar{t}$, this again just implies $r_A^0 \leq r_B^0$ and $r_A^1 \geq r_B^1$. More generally, it requires that the response difference between groups A and B is larger in category $i = 1$ than in category $i = 0$, but considering the responses that took place before time t, for all t. In contrast to the condition for rank-order detection, some fast "most unhappy" responses in group A relative to group B can be compensated by even more and faster "most happy" responses.

Unlike for rank-order detection, if one is interested in on-average detection, it can be useful to consider surveys with more than two response categories. Appendix A.2 contains

an example illustrating this point. The example rests on an asymmetric distribution of happiness within the happier group A, which has a small number of very unhappy subjects who consequently respond to be unhappy very fast in a binary survey. Since group A is happier than group B on average, these unhappy few are more than compensated by many happy subjects within the group, but with an asymmetric distribution this compensation can involve mostly subjects with moderate values of happiness rather than with very large values. Our sufficient detection condition can then be violated for fast response times, for example because these fast response times do not even exist among the subjects who report to be happy. In a survey with three response categories, by contrast, the threshold for reporting in the lowest category will plausibly be lower than in the binary survey, which slows down the unhappy responses in group A and may restore our detection condition in the data.

If one is willing to make the assumption that the happiness distribution of each group is symmetric around its mean, and that in a binary survey the reporting threshold lies between the two groups' average happiness levels, then the generated data will always satisfy the condition in Corollary 3.

Proposition 6. *Suppose that the true average happiness of group A is larger than that of group B, and that the happiness distribution of each group is symmetric around its mean. For $n = 1$ and $\tau^1 \in [\mathbb{E}_{G_B}[\tilde{h}], \mathbb{E}_{G_A}[\tilde{h}]]$, the generated data (r_A, r_B, F_A, F_B) then satisfy that $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ for all $t \in (\underline{t}, \bar{t})$.*

Proof. Denote $\mu_j = \mathbb{E}_{G_j}[\tilde{h}]$ for $j = A, B$. Suppose $\mu_B \leq \mu_A$ and consider a survey with $n = 1$. Let $\tau^1 \in [\mu_B, \mu_A]$ be the reporting threshold and c the chronometric function of the true data-generating process. Then, for all $t \in (\underline{t}, \bar{t})$,

$$\begin{aligned}
r_A^0 F_A^0(t) - r_B^0 F_B^0(t) &= G_A(\tau^1 - c^{-1}(t)) - G_B(\tau^1 - c^{-1}(t)) \\
&= \big[1 - G_A(2\mu_A - \tau^1 + c^{-1}(t))\big] - \big[1 - G_B(2\mu_B - \tau^1 + c^{-1}(t))\big] \\
&= G_B(2\mu_B - \tau^1 + c^{-1}(t)) - G_A(2\mu_A - \tau^1 + c^{-1}(t)) \\
&\leq G_B(\tau^1 + c^{-1}(t)) - G_A(\tau^1 + c^{-1}(t)) \\
&= r_A^1 F_A^1(t) - r_B^1 F_B^1(t),
\end{aligned}$$

where the second equality follows from the symmetry assumption, and the inequality follows from the fact that $\tau^1 \in [\mu_B, \mu_A]$. □

The requirement of symmetry is strong, even though conventional models also make that assumption. A useful aspect of Proposition 6 is that it provides a partial test of symmetry. If the condition for on-average detection is violated in a binary survey, then the true happiness

distributions cannot be symmetric, or the reporting threshold cannot lie between the groups' average happiness levels.[8]

2.3 Robustness

2.3.1 Category Heterogeneity

We now discuss to which extent some of the assumptions made previously can be relaxed. We start with the assumption that the chronometric function is identical in all response categories and applied symmetrically in intermediate categories.

The specific formulation of the chronometric function in intermediate categories is not essential for the conclusion that response times are uninformative in these categories. All previous results remain unchanged as long as response time is continuous between any two interior reporting thresholds and approaches $\bar{t}$ as h approaches any of the thresholds. For instance, the chronometric function could differ across the different intermediate response categories or not follow a simple symmetric specification.

Importantly, conditions (i) and (ii) in Proposition 3 also do not make comparisons across response categories. Hence our necessary and sufficient conditions for rank-order detection continue to hold even if we allow for arbitrary category-specific chronometric functions c^i. This is important when absolute happiness levels directly affect response times, with e.g. more unhappy people being slower (Studer and Winkelmann, 2014). Our results apply as long as such effects do not reverse the monotone chronometric relation within the extreme categories.

By contrast, condition (i) in Proposition 5 does involve a comparison of response times across the two extreme response categories, and hence category-specific chronometric functions are not admissible when on-average detection is concerned.

2.3.2 Group Heterogeneity

We next discuss how group differences in chronometric effects can be incorporated into the analysis. We focus on the case of binary surveys throughout.

Our previous results are not directly robust to group-specific chronometric functions c_j, because the analyst could then always attribute fast choices to either large/small happiness or to generally fast speed of a group. We propose two solutions to this problem. The first is

[8]It is indeed possible to construct an example with symmetric distributions and an extreme reporting threshold which violates the condition for on-average detection. This means that none of the assumptions in Proposition 6 is redundant.

a robust sufficient condition for on-average detection which makes no comparison of response times across groups.

Proposition 7. *Consider* (r_A, r_B, F_A, F_B) *for* $n = 1$ *and allow chronometric functions to be group-specific. Then, group* A *is detectably on-average-happier than group* B *if*

$$r_B^1 F_B^1(t) - r_B^0 F_B^0(t) \leq 0 \leq r_A^1 F_A^1(t) - r_A^0 F_A^0(t)$$

for all $t \in (\underline{t}, \bar{t})$.

This condition is stronger than the on-average detection condition in Corollary 3, but is neither implied by nor implies the condition for rank-order detection in Corollary 2. The result is an application of Theorem 1 in Alós-Ferrer et al. (2021), which implies that, under the stated condition, the average happiness of group A must be larger than the reporting threshold and the average happiness of group B must be smaller than the reporting threshold. This conclusion holds even if the chronometric functions used by the two groups are different. A ranking of the averages obtains from the conventional assumption that both groups use the same reporting threshold.

It follows immediately that the result continues to hold if we allow for group-specific reporting thresholds but assume that the analyst knows that $\tau_A^1 \geq \tau_B^1$, i.e., that the happier group is more reluctant to report high happiness, for example due to adaptation in reporting behavior analogous to actual hedonic adaptation.[9]

Our second solution works with normalized response times that account for individual heterogeneity. Suppose each individual is described by a happiness level h and a speed parameter $\eta > 0$. The response time in the happiness question is $t = c(|\tau^1 - h|) \cdot \eta$, where c is a common chronometric function like before. The distribution of $\tilde{h}$ and $\tilde{\eta}$ in group j is described by a joint cumulative distribution function $G_j(h, \eta)$. We denote by $G_j(h)$ the cumulative distribution function of the marginal distribution of happiness and assume that it satisfies our previous conditions. Our following detection results refer to this marginal distribution. Since $\tilde{\eta}$ can be distributed differently in the two groups, possibly correlated with happiness, this extended model allows for systematic group-differences in response speed.

Now suppose there is an additional baseline question before or after the happiness question, and the response time in the baseline question is $t^b = \phi \cdot \eta$ for some common $\phi > 0$,

[9]Kaiser (2022) investigates heterogeneity and adaptation of reporting thresholds, relying on comparisons between individuals' reported happiness and their memories of how happiness changed over time. His results indicate that adaptation typically goes in the direction that we can allow here, with happier individuals being more reluctant to report high happiness. However, he also shows that heterogeneity in reporting thresholds is typically small and not strong enough to compromise the conventional analysis based on identical reporting thresholds.

unknown to the analyst. This could be a demographic question about e.g. marital status (possibly, but not necessarily, defining membership to groups $j = A, B$) or a question asking for agreement to participate in the survey. The important assumption is that there are no individually varying "intensities" of responses in the baseline question.

The data generated by the extended survey are given by the responses $r_j = (r_j^0, r_j^1)$ to the happiness question for both groups, and cumulative distribution functions $F_j = (F_j^0, F_j^1)$, where $F_j^i(t, t^b)$ describes the joint distribution of response times in both questions among individuals in group j who responded to the happiness question in category i.[10] We now normalize each individual's response time in the happiness question by dividing by the response time in the baseline question, to obtain $\hat{t} = t/t^b$. The distributions of these normalized response times can be obtained from the raw data and are described by cumulative distribution functions $\hat{F}_j^i$ with $\hat{F}_j^i(0) = 0$, which we assume to be continuous.

Proposition 8. *Consider two-question data (r_A, r_B, F_A, F_B) for $n = 1$ and allow speed to be individual-specific. Then, group A is detectably rank-order happier than group B if and only if*

$$r_A^0 \hat{F}_A^0(t) - r_B^0 \hat{F}_B^0(t) \leq 0 \leq r_A^1 \hat{F}_A^1(t) - r_B^1 \hat{F}_B^1(t),$$

and detectably on-average happier if

$$r_A^0 \hat{F}_A^0(t) - r_B^0 \hat{F}_B^0(t) \leq r_A^1 \hat{F}_A^1(t) - r_B^1 \hat{F}_B^1(t),$$

for all $t > 0$.

While systematic speed differences may invalidate our sufficient detection conditions for the raw data, the conditions remain valid when applying them to the normalized data instead.

We just note here that the if-statements can be generalized to group-specific reporting threshold under the previous assumption that $\tau_A^1 \geq \tau_B^1$. We also just note that Propositions 4 and 6 hold for normalized response times as well: if there is first-order stochastic dominance in the (marginal) distributions of happiness, then our condition applied to the normalized

[10] We ignore the response in the baseline question, other than that it may define group membership. Note that the analysis of multiple questions with interdependencies (via η) could in principle give rise to questions of "rationalizability" (see Alós-Ferrer et al., 2021). Namely, is there always a data-generating process that explains any given extended dataset (r_A, r_B, F_A, F_B)? To see why the answer is yes here, fix an arbitrary $\phi > 0$, any chronometric function c satisfying our assumptions with $\underline{t} = 0$ and $\bar{t} = \infty$, and set $\tau^1 = 0$. Define the cumulative distribution functions $H_j^i(h, \eta) = F_j^i(c(-h)/\phi, \phi \cdot \eta)$ for $h < 0$ and $\eta > 0$. Then construct G_j by having $h < 0$ with probability r_j^0, in which case (h, η) follows distribution H_j^0, and having $h > 0$ with probability r_j^1, in which case $(-h, \eta)$ follows distribution H_j^1. This constructed process generates the given data (and the marginal distributions of happiness are continuous if the marginal distributions of response times in the happiness questions are continuous).

data will detect it, even when there are individual-specific differences in response speed, and the same is true for on-average detection under the additional assumptions required by Proposition 6.

2.3.3 Measurement Error

In this subsection, we discuss to what extent our results are robust to measurement error. We continue to consider binary surveys and response times generated by $c(|\tau^1 - h|) \cdot \tilde{\eta}$, but $\tilde{\eta}$ now is a random variable capturing noise in the measurement of response times by the analyst, rather than individual-specific speed differences. For example, the "time to response" that the analyst measures in an online survey may be inflated when the subject was distracted for that specific question. Normalization will not take care of such measurement error because the error is not the same across questions for an individual.

While our interpretation is one of measurement error in this subsection, there are other sources of noise that $\tilde{\eta}$ could capture. For example, if a baseline question is used for normalization, and there is noise or unaccounted variation of intensities in that baseline question, this gives rise to analogous errors in the normalized response times.

Stochasticity makes it difficult to obtain sufficient detection conditions, because minor violations of first-order stochastic dominance of the happiness distributions may be smoothed out by the noise and not detectable in the data. However, we can generalize Proposition 4 and show that our techniques, even though "misspecified" because they were derived for a model without measurement error, continue to detect a dominance relation whenever it exists, and a violation of our detection condition still falsifies the hypothesis of first-order stochastic dominance. For that result we need to assume that $\tilde{\eta}$ is distributed independently of happiness and i.i.d. in both groups. It is not necessary to make any other assumptions about the distribution. For example, its mean can be larger than one, so measurement error can be systematic.

Proposition 9. *Suppose that the true happiness distribution of group A first-order stochastically dominates that of group B, and that there is i.i.d. measurement error. For $n = 1$, the generated data (r_A, r_B, F_A, F_B) then satisfy that $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq 0 \leq r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ for all $t > 0$.*

The reason for this robustness is that i.i.d. measurement error affects both groups' recorded response times equally and thus does not distort our detection condition.

With regard to on-average detection, we can go one step further and allow for systematic group differences in measurement error. Formally, the random variable $\tilde{\eta}$ can follow different distributions in the two groups, but still assuming independence from happiness and across

groups. For example, individuals with different educational backgrounds may pay attention differentially in the question of interest. Our techniques continue to detect the correct ranking of happiness averages and to serve as a partial test of symmetry.

Proposition 10. *Suppose that the true average happiness of group A is larger than that of group B, that the happiness distribution of each group is symmetric around its mean, and that there is group-specific measurement error. For $n = 1$ and $\tau^1 \in [\mathbb{E}_{G_B}[\tilde{h}], \mathbb{E}_{G_A}[\tilde{h}]]$, the generated data (r_A, r_B, F_A, F_B) then satisfy that $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ for all $t > 0$.*

The reason for the robustness to group-specific measurement error is that our on-average detection criterion trades-off the behavior of a group across the two response categories, so that group differences wash out because they affect both the LHS and the RHS of the required inequality.

We conclude this subsection by just noting that Propositions 9 and 10 could be further generalized to allow for i.i.d. noise in the reporting threshold.

2.3.4 Limits

Based on the robustness results from the previous subsections, we can delineate the limits of our analysis. These limits are reached when factors influence response times that are either systematically different between the groups or correlated with happiness, and which cannot be addressed by normalization. For example, the groups may differ systematically in the attention that they bring to a complex question like happiness but not to a simple question like marital status, so normalization cannot address the issue. Rank-order detection will fail in that case. On-average detection may still work (see Proposition 10), but here the limit is reached when the devoted attention is additionally correlated with happiness.

2.4 Extension

We pointed out before that response times in intermediate response categories are not helpful for identification, as they are not monotone in the latent variable. We now show that intermediate categories become more informative if we add binary follow-up questions.

Whenever an individual responds in an intermediate category $i = 1, \ldots, n-1$ in the initial question, then she can be asked a follow-up question which requires her to indicate whether she felt closer to category $i-1$ or to category $i+1$. Let $\underset{\sim}{r}_j^i$ be the fraction of individuals within group j who respond in category i in the initial question and subsequently report to have been closer to category $i-1$. The response time distribution of these individuals in

the follow-up question is denoted by $\underline{F}_j^i$. Similarly, denote by $\overline{r}_j^i$ the fraction of individuals within group j who initially respond in category i and subsequently indicate to have been closer to category $i+1$, and let $\overline{F}_j^i$ be the corresponding distribution of response times in the follow-up question. Our extended data now consist of responses $r_j = (r_j^0, (r_j^i, \underline{r}_j^i, \overline{r}_j^i)_i, r_j^n)$, where $r_j^i = \underline{r}_j^i + \overline{r}_j^i$, and response time distributions $F_j = (F_j^0, (F_j^i, \underline{F}_j^i, \overline{F}_j^i)_i, F_j^n)$, assumed to have the previous properties.

We augment the data-generating process by adding, for each initial intermediate response $i = 1, \ldots, n-1$, a reporting threshold $\check{\tau}^i \in (\tau^i, \tau^{i+1})$ for the follow-up question, and a chronometric function $\check{c}^i$ with the previous properties. The follow-up responses are then given by $\underline{r}_j^i = G_j(\check{\tau}^i) - G_j(\tau^i)$ and $\overline{r}_j^i = G_j(\tau^{i+1}) - G_j(\check{\tau}^i)$. Provided that the respective responses exist, the response time distributions are determined by

$$\underline{F}_j^i(t) = \frac{\max\{0, G_j(\check{\tau}^i - (\check{c}^i)^{-1}(t)) - G_j(\tau^i)\}}{G_j(\check{\tau}^i) - G_j(\tau^i)}$$

and

$$\overline{F}_j^i(t) = \frac{\max\{0, G_j(\tau^{i+1}) - G_j(\check{\tau}^i + (\check{c}^i)^{-1}(t))\}}{G_j(\tau^{i+1}) - G_j(\check{\tau}^i)},$$

for all $t \in (\underline{t}, \overline{t})$. It is now straightforward to extend the definition of rank-order detection accordingly, and we obtain the following result.

Proposition 11. *Consider a survey with binary follow-up questions. Given (r_A, r_B, F_A, F_B), group A is detectably rank-order happier than group B if and only if*

(i) $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq 0$ *for all* $t \in (\underline{t}, \overline{t})$,

(ii) $r_A^n F_A^n(t) - r_B^n F_B^n(t) \geq 0$ *for all* $t \in (\underline{t}, \overline{t})$,

and, for all $k = 1, ..., n-1$,

(iii) $\sum_{i=0}^{k-1} r_A^i + \underline{r}_A^k \underline{F}_A^k(t) \leq \sum_{i=0}^{k-1} r_B^i + \underline{r}_B^k \underline{F}_B^k(t)$ *for all* $t \in (\underline{t}, \overline{t})$, *and*

(iv) $\sum_{i=k+1}^{n} r_A^i + \overline{r}_A^k \overline{F}_A^k(t) \geq \sum_{i=k+1}^{n} r_B^i + \overline{r}_B^k \overline{F}_B^k(t)$ *for all* $t \in (\underline{t}, \overline{t})$.

The introduction of follow-up questions does not affect the extreme categories: conditions (i) and (ii) remain the same as in Proposition 3. The novelty in Proposition 11 are conditions (iii) and (iv), which are weaker than the condition for intermediate response categories in Proposition 3. Intuitively, the binary structure of the follow-up questions allows us to fully exploit monotonicity of response times. Hence, we can use the data generated from these

additional questions to examine dominance relations between the happiness distributions of the two groups also within intermediate response categories.

An analogous extension for on-average detection using follow-up questions does not exist without very strong assumptions. The reason is that the follow-up reporting threshold $\check{\tau}^k$ is not necessarily exactly in the middle of the interval (τ^k, τ^{k+1}) of possible happiness levels among individuals that get the respective follow-up question. Thus, some small response times could arise for one of the follow-up answers but never for the other, which invalidates the averaging across response categories that we used earlier.

We conclude with a note of caution. Our analysis assumed that the chronometric effect exists in the follow-up questions as well, i.e., responses are faster when the latent variable is more distant from the reporting threshold in the follow-up question. This is less obvious than for the initial question. For example, subjects may have made up their mind about their happiness already when the initial question was posed, and response time in the follow-up question may therefore be driven by other considerations.

3 Empirical Application

3.1 Survey Description

In this section, we connect our theoretical framework to actual survey data. The goal of our empirical investigation is twofold. First, we want to verify the key assumption of our model: the presence of chronometric effects in surveys. Second, we want to show how our response time techniques can be implemented in practice. To this end, we designed and conducted a survey experiment on the online platform MTurk, which has become increasingly popular among behavioral scientists in economics (e.g. Kuziemko et al., 2015; DellaVigna and Pope, 2018), marketing (e.g. Goodman and Paolacci, 2017), and psychology (e.g. Paolacci and Chandler, 2014). Conducting the survey on an online platform like MTurk has the advantage of allowing accurate records of the response times of participants.

Our survey was programmed using the software Qualtrics and was conducted in April-May 2022 through the ETHZ Decision Science Laboratory.[11] The survey consisted of two parts. The first part included 6 standard socio-demographic questions concerning gender, age, education, marital status, co-residence with children, and family income. These questions are commonly asked in large-scale surveys like the GSS, which is the primary source for US evidence on a broad set of social science issues (Davis and Smith, 1991). In the second

[11]The first discussion paper version of this paper (Liu and Netzer, 2020) contains data from another survey conducted on MTurk already in 2019. This older survey had a smaller number of participants, no question about income, and it did not contain follow-up questions. We are not using those data here.

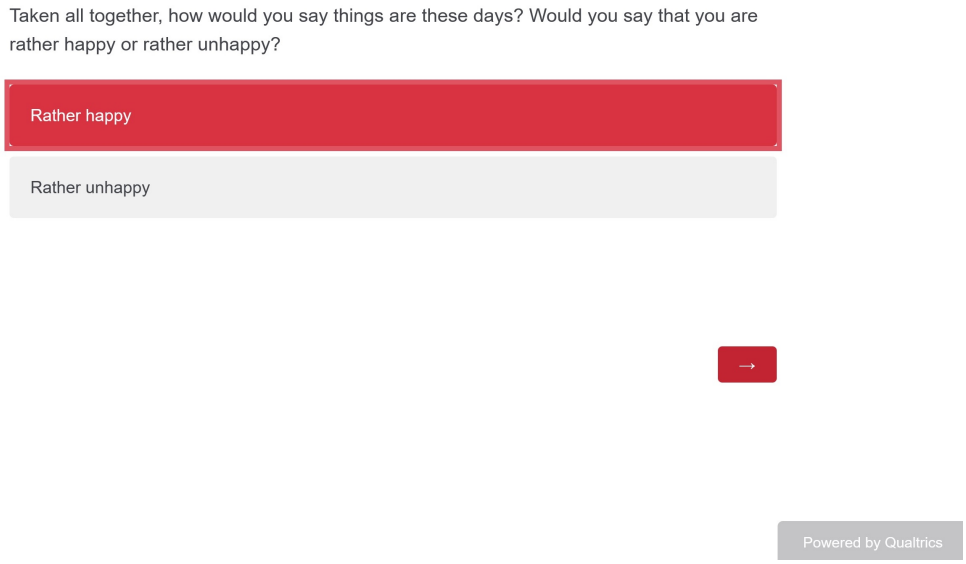


FIGURE 2: Example of survey screen.

part, the subjects were asked 7 substantive questions. These questions elicit information about (i) job satisfaction, (ii) social life satisfaction, (iii) overall happiness, (iv) trust attitude, (v) political attitude, (vi) time preference, and (vii) risk preference. The questions for (i)–(v) were again adapted from the GSS, and for (vi) and (vii) the questions were adapted from the Global Preference Survey introduced by Falk et al. (2018).

We implemented two different versions of the survey, to which we randomly assigned the subjects. In one version, the possible response to each substantive question was binary, e.g., "rather happy" and "rather unhappy" for the overall happiness question. The other version had three response categories, e.g., "rather happy", "neither happy nor unhappy", and "rather unhappy". In addition, both versions of the survey included binary follow-up questions that ask the subjects to refine their initial answer to each substantive question, e.g., after an initial response "rather happy" they are asked to refine between "very happy" and "moderately happy." The complete questionnaires can be found in Appendix C.

Figure 2 provides an example of the survey screen displayed to the subjects. Before choosing the submission button "→" at the right bottom of the screen and moving on to the next page, the subjects first had to select one of the available responses to the question (there was no default answer). They were allowed to change their response as long as the current page had not been submitted, but they could not go back to a previous question after submission of the answer. In addition to the responses to the questions, we collected data on response times, which we define as the total amount of time between the display of the question and a subject's last click before submission. This "time to final response"

captures most closely the duration of the decision process, which may involve changing an initial response by clicking on a different button.

3.2 Summary Statistics

We recruited 8,007 subjects from the US with an MTurk approval rate of at least 95%. Each subject received a fixed compensation of 60 cents for completing the survey. In the initial sample, 286 subjects failed an attention check at the end of the survey ("What is 7 times 2?"). No click and time data were recorded for 253 subjects, presumably because they used keyboard navigation to answer the questions. All these subjects were dropped, so our final sample contains 3,744 subjects in the binary survey and 3,724 subjects in the trinary survey. Table 1 summarizes the demographics of the subjects and shows that they are very similar in the two survey versions, as should be expected given the random assignment.[12]

Roughly 90% of the subjects completed the survey within 5 minutes. The median duration was 123s and the average duration was 167s. Table 2 summarizes the median response times for each question (not including the follow-ups) and each survey version separately. The socio-demographic questions and their possible responses were the same in the two survey versions, and hence the median response times are also approximately the same. The median response times for the substantive questions are smaller in the binary survey than in the trinary survey, reflecting that the latter involves more response categories that have to be read, understood, and considered by the subjects.

The marital status question had the quickest median (and also average) response time in both survey versions, reflecting that the question was short and easy to answer. Furthermore, there are typically no varying uncertainties or intensities about being married that could affect response times. Hence, we will use the response time in the marital status question for individual normalization in our following analysis. That is, we will divide each subject's response time in each of the substantive questions by the subject's response time in the

[12] After the survey was completed, we became aware that a significant number of observations in our dataset had suspiciously similar IP addresses. Specialists of the ETHZ Decision Science Laboratory conjectured that these observations may be from participants using virtual private networks (VPNs), but the ultimate source of the pattern remains unknown. Following the suggestion of the ETHZ Decision Science Laboratory, as a conservative robustness check we excluded all observations where the first three IP blocks appeared more than once, which amounts to about 40% of our data. Appendix D contains all our main results based on the restricted sample. Participants in the restricted sample report to be somewhat less educated, be less often married, have children less often, and have a more spread-out income distribution, but are otherwise similar. The results of our analyses are also largely comparable to those for the full dataset. The chronometric effect is confirmed in the restricted sample, and the results of ordered probit are comparable (for example, we never obtain significant parameter estimates of opposite sign). There are some differences in the p-values of our detection hypotheses, but overall the p-values are clearly positively correlated between the full and the restricted sample.

	binary survey	trinary survey
# participants	3,744	3,724
female	50.08%	51.34%
male	49.92%	48.66%
age		
< 20	0.37%	0.62%
$20 - 29$	24.39%	26.91%
$30 - 39$	34.83%	32.92%
$40 - 49$	21.88%	21.51%
$50 - 59$	11.40%	11.09%
$60 - 69$	6.09%	5.99%
≥ 70	1.04%	0.97%
highest education		
high school	17.17%	17.37%
college or higher	82.29%	82.28%
none	0.53%	0.35%
married	64.64%	65.15%
unmarried	35.36%	34.85%
kids	60.87%	61.52%
no kids	39.13%	38.48%
income		
$< \$40,000$	26.90%	27.12%
$\$40,000 - \$69,999$	43.94%	43.98%
$\geq \$70,000$	29.17%	28.89%

TABLE 1: Summary of subject demographics.

marital status questions (or subtract it in logs). That way, we can account for individual differences in the speed of reading or decision-making more generally (recall the formal argument in Section 2.3.2).

3.3 Testing the Chronometric Effect

In our survey, each substantive question was accompanied by a follow-up question requiring the subjects to refine their initial response. This design makes it possible to directly test for the presence of chronometric effects, and in particular our crucial assumption that response times are monotone in the latent variable within the extreme categories. For example,

	binary survey	trinary survey
complete survey	119	128
demographic questions		
gender	1.66	1.66
age	2.05	2.07
education	2.06	2.08
marital status	1.52	1.51
kids	1.73	1.73
income	2.18	2.16
substantive questions		
work satisfaction	2.58	3.33
social life satisfaction	2.49	2.84
overall happiness	2.94	3.42
trust	3.26	4.03
political attitude	2.12	2.21
time preference	3.98	4.32
risk preference	2.34	2.62

TABLE 2: Median response times in seconds.

consider only those subjects who responded in the extreme "rather happy" category in the initial question about overall happiness. Based on their response in the corresponding follow-up question, we can further distinguish those who are "very happy" from those who are only "moderately happy," with the former having larger values of the latent variable than the latter. If the chronometric effect exists, then the former should have responded faster than the latter in the initial question. In other words, the chronometric assumption that we use for our detection method can be tested by the prediction that more extreme follow-up responses should be associated with faster initial responses.

To test the above prediction, we pool all observations from the binary survey and all observations with non-intermediate responses from the trinary survey, and we estimate the following equation:

$$\log RT_{sq} = \beta_0 + \beta_1 FU_{sq} + \beta_2 \mathbf{X}_s + \gamma_q + \epsilon_{sq}. \tag{5}$$

The dependent variable in (5) is the log of the normalized response time of subject s in initial substantive question q (not including the follow-up). The main explanatory variable of interest is FU_{sq}, a dummy that is one if the subject chose the more extreme response among

	Log Normalized Response Time			
	(1)	(2)	(3)	(4)
Follow-Up Response	-0.449***	-0.371***	-0.150***	-0.101***
	(0.0138)	(0.0133)	(0.0076)	(0.0079)
R-squared	0.0716	0.1066	0.0674	0.0681
Demographics & Treatment	NO	YES	YES	NO
Individual RE	NO	NO	YES	NO
Individual FE	NO	NO	NO	YES

*** p<0.01, ** p<0.05, * p<0.1.

Notes: All regressions include all observations from the binary survey and the observations with non-intermediate responses from the trinary survey. The dependent variable is each subject's log response time in the initial substantive question (not including the follow-up), normalized by subtracting his/her log response time in the marital status question. Follow-Up Response is a dummy that takes the value one if the subject chose the extreme response (e.g. "very happy" or "very unhappy") in the corresponding follow-up question. All regressions include question fixed effects. The demographic controls are gender, age, education, marital status, co-residence with children, and family income. Treatment is a dummy for the survey version (binary versus trinary). Column (3) is a random-effect model with all demographic and treatment controls. Column (4) is a fixed-effect model which controls for heterogeneity at the subject level. Robust standard errors are reported in parentheses, with the ones in columns (1) and (2) being clustered at the subject level. The R-squared values reported in columns (3) and (4) concern the variation within subjects.

TABLE 3: Regression analysis of chronometric effects.

the two given in the corresponding follow-up question (e.g., "very happy" after an initial response of "rather happy," or "very unhappy" after an initial response of "rather unhappy"), and zero otherwise. Other controls include the version of the survey that the subject received (binary versus trinary) and the socio-demographic information that our survey collected, all summarized in $\mathbf{X}_s$. Lastly, the variable γ_q captures question fixed effects.

Table 3 reports the results of estimating (5). As shown in the first row of the table, the coefficient of the dummy variable for an extreme follow-up response is always negative and highly significant. This finding is robust if we include demographic and treatment controls, individual random effects, or if we instead employ a fixed-effect model to control for heterogeneity at the subject level. The regression analysis therefore confirms our central assumption: subjects with more extreme latent values – as revealed by the information that they provide in the follow-up question – respond faster to the initial question.[13]

We can also examine the relation between follow-up responses and response times in

[13]Table B1 in Appendix B shows that non-normalized, raw response times exhibit the same pattern.

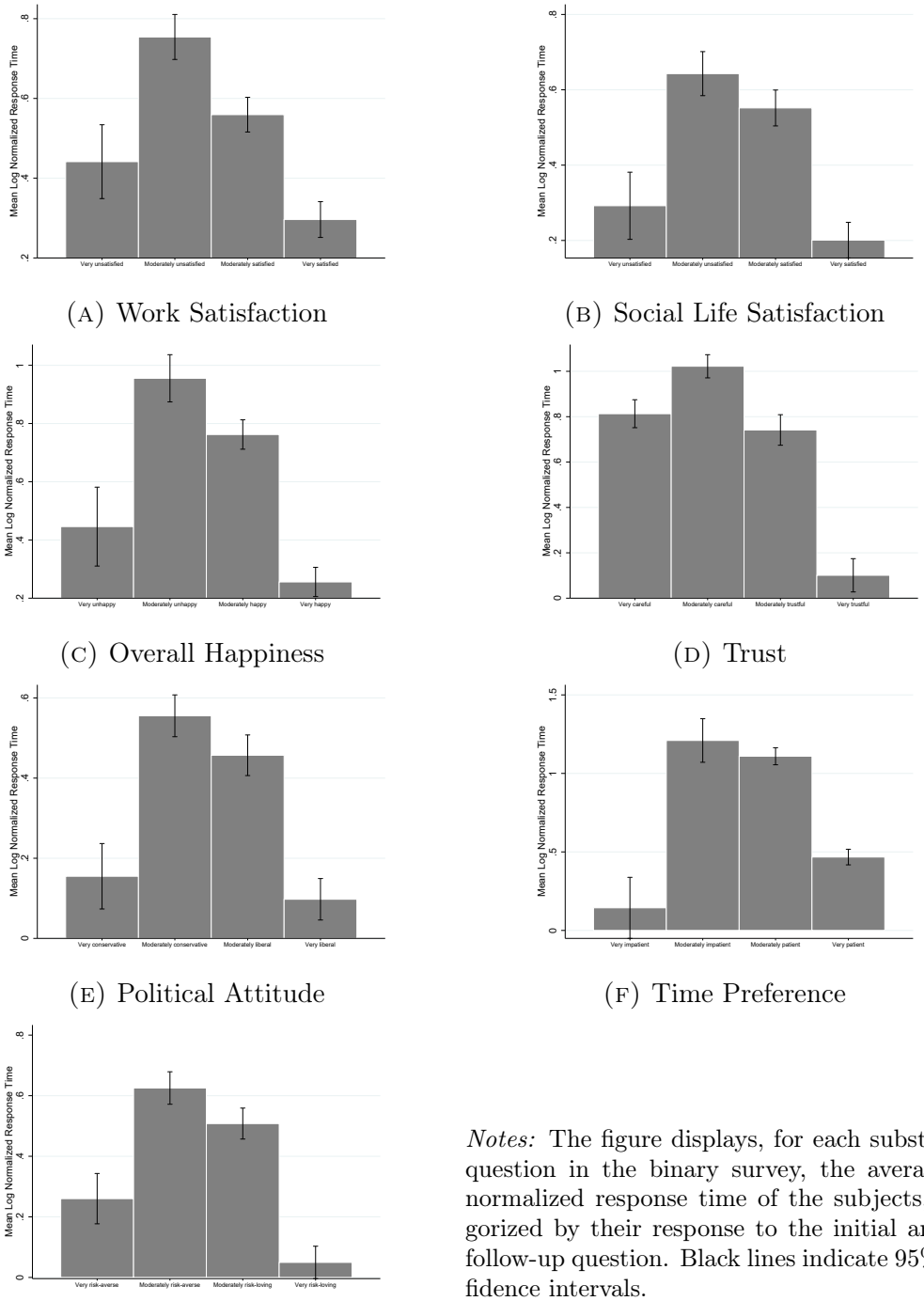
the initial question separately for each substantive question. Figures 3 and 4 summarize our findings for the binary and the trinary survey, respectively. As an illustrative example, consider panel (C) in Figure 3, which concerns the overall happiness question in the binary survey. The subjects are ordered from left to right according to how they responded to the initial question and its follow-up: very unhappy, moderately unhappy, moderately happy, and very happy. Each bar in the graph depicts the average log normalized response time of the respective group in the initial question, along with its 95% confidence interval. The chronometric function becomes visible as a hump-shape. Among the subjects who initially responded to be rather unhappy (bars one and two), those who respond in the follow-up to be very unhappy (first bar) were faster in the initial question than those who respond to be only moderately unhappy (second bar). Analogously, among the subjects who initially responded to be rather happy (bars three and four), those who respond in the follow-up to be very happy (bar four) were faster than those who respond to be only moderately happy (bar three). The hump-shape confirms that subjects with latent values further away from the reporting threshold give their response more quickly on average.

The hump-shape exists for all substantive questions in both versions of the survey. The mean response time is always smaller for the more extreme group than for the less extreme one, and almost all of these differences are statistically significant at the 1% level.[14] Altogether, the evidence strongly supports that survey responses display a chronometric effect.[15]

Since our extended detection results from Section 2.4 use response times from the follow-up questions, it is worthwhile to ask whether the chronometric effect also holds for those questions. If this were true, our theory would predict an intriguing correlation of response times between an initial question and its follow-up. Again, take the overall happiness question as an example, and consider the subjects who first responded "rather happy" and then refined their answer to "very happy." Within this group of subjects, response times should be positively correlated between the initial and the follow-up question, because a larger happiness implies being more distant from the reporting threshold in both stages. By contrast, within the group of subjects who first responded "rather happy" but then refined their answer to "moderately happy", the correlation should be negative, because a larger happiness

[14] Among the 28 pairwise comparisons, 25 are significant at the 1% level according to a t-test (two-sided, unequal variances), with the exceptions being in the trinary survey: the pair "very unsatisfied" and "moderately unsatisfied" in the work satisfaction question ($p = 0.0737$), the pair "very careful" and "moderately careful" in the trust question ($p = 0.3779$), and the pair "very impatient" and "moderately impatient" in the time preference question ($p = 0.0107$). Figures B1 and B2 in Appendix B show that similar patterns, albeit less pronounced, exists for non-normalized, raw response times. Only 7 out of the 28 pairwise comparisons are statistically significant at 1%, but all of them in the direction implied by the chronometric effect.

[15] Another interesting observation is that initial responses in the high category are almost always faster than responses in the low category, which could be explained either by asymmetric distributions of the latent variables or by category-specific chronometric functions (see Section 2.3.1).



(A) Work Satisfaction

(B) Social Life Satisfaction

(C) Overall Happiness

(D) Trust

(E) Political Attitude

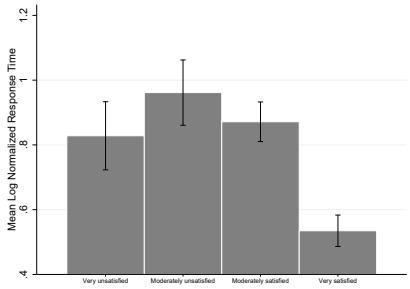
(F) Time Preference

(G) Risk Preference

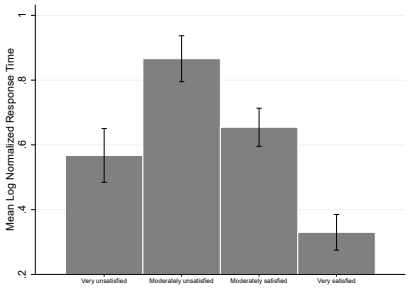
Notes: The figure displays, for each substantive question in the binary survey, the average log normalized response time of the subjects, categorized by their response to the initial and the follow-up question. Black lines indicate 95% confidence intervals.

FIGURE 3: Chronometric effect by question in the binary survey.

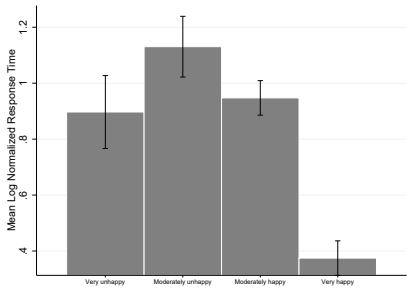
means being closer to the reporting threshold in the follow-up stage. We did not find such differentiated patterns in our data. As Tables B2 – B5 in Appendix B show, response times are always positively correlated across stages regardless of which follow-up response we focus



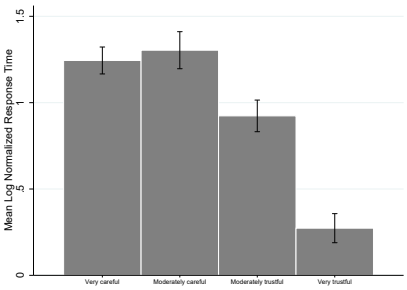
(A) Work Satisfaction



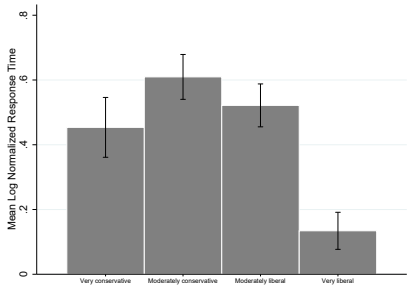
(B) Social Life Satisfaction



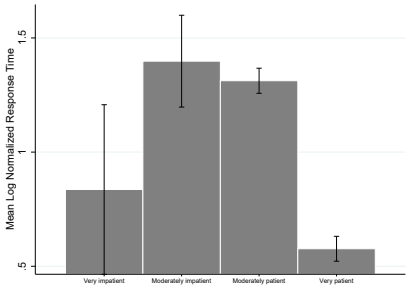
(C) Overall Happiness



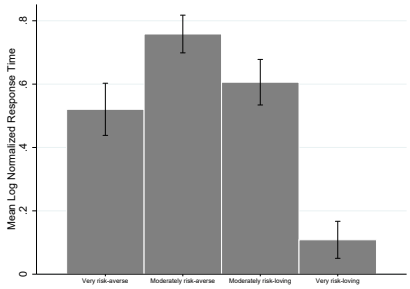
(D) Trust



(E) Political Attitude



(F) Time Preference



(G) Risk Preference

Notes: The figure displays, for each substantive question in the trinary survey, the average log normalized response time of the subjects, categorized by their (non-intermediate) response to the initial and the follow-up question. Black lines indicate 95% confidence intervals.

FIGURE 4: Chronometric effect by question in the trinary survey.

on, even when controlling for individual fixed effects, for both the normalized and the raw response time data. One explanation for the absence of chronometric effects in the follow-up questions is that subjects already made up their mind about the issue, e.g., about how happy

they are, when answering the initial question, and they do not have to think carefully again when answering the follow-up question. Because we cannot validate chronometric effects for the follow-up questions in our dataset, we will in the following not report detection results based on our analysis from Section 2.4.[16]

3.4 Analysis of Binary Survey

Having verified the key premise of our approach – the chronometric effect – we now apply our detection criteria to the binary survey. We divide the sample into socio-demographic groups and, for each substantive question, make pairwise comparisons between the groups to check whether the conditions for detection are satisfied by the data. We do this separately for each socio-demographic characteristic, e.g. we compare the happiness between females and males, and between the young and the middle-aged. Finer divisions of the sample can of course be made, but since our focus here is not on a causal interpretation of the results, we prefer keeping the number of pairwise comparisons low.

Table 4 reports estimates from a traditional ordered probit model, for all our combinations of socio-demographic groups and substantive questions. Each cell corresponds to a regression of the response to the question in the column on a dummy for membership to the group in the row. The ordered probit coefficients are reported along with their robust standard errors. For example, from row six in column one we learn that married subjects are significantly more satisfied with their work than unmarried subjects.

Are these traditional results reliable, if only qualitatively, or do they depend on unjustified distributional assumptions? To answer this question, we observe that in each group the fraction of subjects responding in each category – the empirical counterpart of r_j^i – is always strictly positive, for all substantive questions. Hence, the conditions of Corollary 1 for rank-order and on-average detection without response times are never satisfied. Consistent with Bond and Lang (2019), little can be learned from our data within the traditional ordered response framework without making distributional assumptions.

Next, we examine the conditions in Corollaries 2 and 3 which make use of response and response time data. We normalize each participant's log response time by subtracting his/her log response time in the marital status question. We can then construct the empirical cumulative distribution functions – the empirical counterpart of F_j^i – of these log-normalized response times. Detecting a rank-order relationship requires us to check whether

[16]The analysis in Section 2.4 is still valuable because the chronometric effect may exist in follow-up questions of other surveys. One can even contemplate designing a survey to induce the chronometric effect, for example by posing the follow-up questions not immediately after the original questions, with the goal of restarting the subjects' thinking process.

	work satisfac.	social satisfac.	overall happiness	trust	liberal-ism	patience	risk-taking
0: female	0.009	0.040	0.047	0.133***	−0.078*	−0.055	0.329***
1: male	(0.0471)	(0.0441)	(0.0466)	(0.0410)	(0.0417)	(0.0539)	(0.0422)
0: young	0.181***	0.031	0.169***	0.052	−0.174***	−0.105*	−0.193***
1: middle-age	(0.0519)	(0.0479)	(0.0512)	(0.0443)	(0.0450)	(0.0576)	(0.0453)
0: middle-age	−0.004	−0.080	0.037	0.005	−0.076	0.034	−0.320***
1: old	(0.1009)	(0.0902)	(0.1007)	(0.0846)	(0.0849)	(0.1093)	(0.0848)
0: none	−0.127	−0.460	−0.238	−0.302	−0.310	0.464	−0.194
1: high-school	(0.2925)	(0.3090)	(0.3092)	(0.2885)	(0.2991)	(0.3114)	(0.2856)
0: high-school	0.783***	0.523***	0.502***	0.664***	0.114**	0.086	0.522***
1: college	(0.0572)	(0.0557)	(0.0577)	(0.0570)	(0.0549)	(0.0698)	(0.0548)
0: unmarried	0.930***	0.796***	0.808***	0.639***	−0.130***	0.257***	0.512***
1: married	(0.0495)	(0.0460)	(0.0485)	(0.0439)	(0.0438)	(0.0550)	(0.0437)
0: no kids	0.835***	0.742***	0.651***	0.565***	−0.040	0.199***	0.582***
1: kids	(0.0491)	(0.0455)	(0.0478)	(0.0427)	(0.0428)	(0.0544)	(0.0431)
0: poor	0.740***	0.477***	0.556***	0.472***	0.010	0.199***	0.287***
1: middle-income	(0.0567)	(0.0532)	(0.0554)	(0.0507)	(0.0512)	(0.0628)	(0.0512)
0: middle-income	−0.113*	−0.061	0.033	−0.230***	−0.114**	0.216***	−0.159***
1: rich	(0.0607)	(0.0544)	(0.0595)	(0.0490)	(0.0497)	(0.0691)	(0.0504)

TABLE 4: Ordered probit analysis of the binary survey. Each cell corresponds to a regression of the question in the column on a dummy for membership to the group in the row. Coefficients are reported along with their robust standard errors in parentheses. Asterisks indicate statistical significance (*10%,**5%,***1%).

the empirical $r_A^0 F_A^0(t) - r_B^0 F_B^0(t)$ is below zero and the empirical $r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ is above zero for all observed t. Detecting the ranking of the averages requires that the first of these expressions is always smaller than the second. Naturally, noise affects these conditions when applied to empirical distributions. To make statistical inference, we draw upon the test for stochastic dominance proposed by Barrett and Donald (2003). A null hypothesis of stochastic dominance boils down to an inequality between two distribution functions, so the core of the Barrett-Donald test is to construct a supremum-type statistic from the original sample and compute critical values from bootstrap samples. Our detection conditions also involve inequalities between distribution functions, with the intricacy that these functions are weighted by the response fractions. To account for this special feature of our setting, we treat the empirical fractions r_j^i as fixed in the test and stratify the bootstrap samples so that they all have the same empirical response frequencies as the original sample.[17]

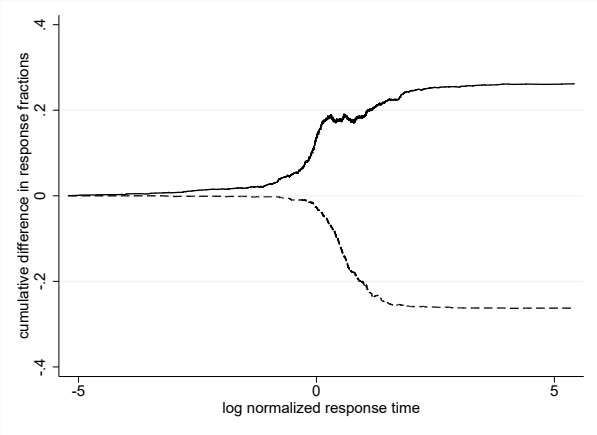
[17]To be more precise, for the null hypothesis that $r_A^0 F_A^0(t) \leq r_B^0 F_B^0(t)$ for all t, we compute the statistic $\hat{S} = \sup_t [r_A^0 F_A^0(t) - r_B^0 F_B^0(t)]$ and use equation (11) in Barrett and Donald (2003, p. 82) for bootstrapping,

To illustrate our findings, consider first the question about work satisfaction ("How satisfied are you with the work you do?") and compare the groups of participants who are married (group A) and who are unmarried (group B). The solid curve in Figure 5(A) plots $r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$, the cumulative difference in response fractions between these two groups for the response category "rather satisfied," with t varying on the x-axis. This curve always lies above zero, meaning that the fraction of married subjects who responded to be satisfied with their job is always larger than that of the unmarried subjects, even when restricting attention to responses which took place before any time t. Similarly, the dashed curve plots the cumulative difference for the answer category "rather unsatisfied," $r_A^0 F_A^0(t) - r_B^0 F_B^0(t)$. It always lies below zero, because the fraction of subjects who responded to be unsatisfied with their job is always smaller for the married than for the unmarried, again for all response times. Since the inequalities hold perfectly in the data, the Barrett-Donald test cannot reject the null hypothesis of a rank-order ($p = 1.000$). Taken together, subjects who are married are detected to be rank-order more satisfied in the work domain than those who are unmarried. The timing of responses rules out that the latent variable follows distributions for which the findings of traditional ordered response model would be reversed.

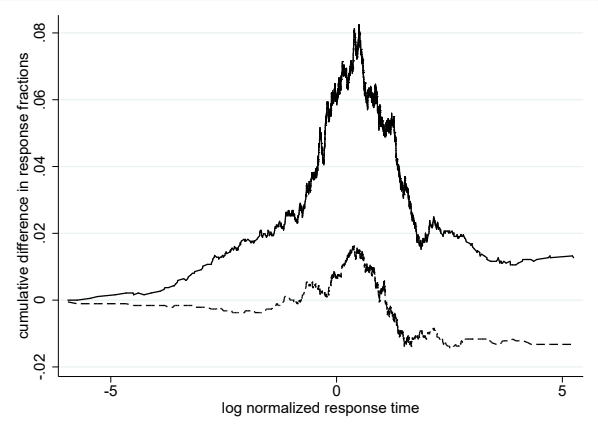
As a second example, consider the relationship between overall happiness and gender. From Table 4 we see that males are happier than females according to an ordered probit model, but the effect is not significant. Figure 5(B) suggests that there is no detectable rank-order relation of happiness between these two groups. The dashed curve goes above zero for some intermediate values of t, meaning that there are some systematically faster "rather unhappy" responses in the male group than in the female group (despite our normalization that would take care of gender-specific speed differences). The Barrett-Donald test indicates that this violation of our detection condition is relatively unlikely to be a coincidence ($p = 0.083$). Hence, we cannot rule out that some male participants are so unhappy as to invalidate the traditional assumption of first-order stochastic dominance in the happiness distributions. Nevertheless, the fast "rather unhappy" responses in the male group are offset by even faster and more frequent "rather happy" responses, reflected in the fact that the dashed curve always lies below the solid curve in Figure 5(B). Male participants are detectably on-average happier than female participants ($p = 1.000$).

Finally, Figure 5(C) illustrates that sometimes even on-average detection fails unequivo-

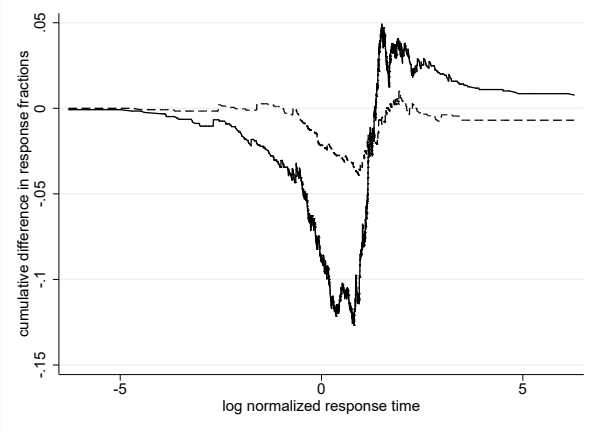
keeping the fractions r_A^0 and r_B^0 fixed. The same approach is used for the null hypothesis that $r_B^1 F_B^1(t) \leq r_A^1 F_A^1(t)$ for all t. Since both conditions have to be satisfied simultaneously for rank-order detection, our p-values count how often both bootstrapped values exceed their respective statistic in 1000 repetitions. For on-average detection, the analogous procedure is used to test the null hypothesis that $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ for all t. Our paper is accompanied by Stata ado-files which implement these tests for surveys with an arbitrary number of response categories, and with or without follow-up questions.



(A) work satisfaction, married vs. unmarried



(B) overall happiness, male vs. female



(C) patience, old vs. middle-age

Notes: Each subfigure refers to a different question in the binary survey. The first sociodemographic group described in the caption is coded as group A, the second as group B. Solid curves depict $r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ and dashed curves depict $r_A^0 F_A^0(t) - r_B^0 F_B^0(t)$.

FIGURE 5: Examples of our detection analysis.

cally. We compare the patience ("How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?") between age groups. We think of the latent variable for this question as being a time preference parameter, such that higher values capture greater patience. From Table 4 we learn that old participants (age over 60) are more patient than middle-aged participants (age $40-59$), but the effect is not significant. The solid curve in Figure 5(C) shows that many middle-aged participants very quickly responded that they are are willing to give up immediate rewards for a future benefit, even though overall a higher fraction of participants in the old group responded in this category. Hence, the response times reveal that some middle-aged subjects are particularly patient. As a consequence, the Barrett-Donald test clearly rejects the null hypothesis of a rank-order of the latent distributions between the groups ($p = 0.000$) and also the null hypothesis of a detection on-average ($p = 0.014$).

Figure 6 summarizes all our detection results for the binary survey. Each circle represents a comparison between two socio-demographic groups. The circles in white display the statistical confidence that we have in identifying a rank-order relationship (i.e. the p–value of the Barrett-Donald test) on the x-axis, for the different substantive survey questions stacked on the y-axis. The circles in grey give the same information for on-average detection. The figure shows a reference line at $p = 0.100$ as an orientation for statistical significance.

Figure 6 documents systematic patterns. For example, for the risk preference question and for the two satisfaction questions about work and social life, our results broadly support the traditional assumptions of ordered-response models. The statistical confidence in a detection of the correct ranking of averages is large in all pairwise comparisons. In fact, our conditions for on-average detection are never rejected at a 10% significance level, and quite often the detection is exact ($p = 1.000$) or almost exact. Even the stronger conditions for rank-order detection are rejected only in few comparisons. On the other extreme, for the time preference question a substantial fraction of the circles is concentrated at low p–values. We conclude that time preference parameters are not likely to follow the distributions assumed in traditional ordered-response models. The contrast between the questions about risk preferences and time preferences may be of interest for the literature that explores the relation between preferences in the risk and the time domain (Andreoni and Sprenger, 2012). The questions about overall life happiness, trust, and political attitude are somewhere in between. Our confidence in correct on-average detection is still large for all but few pairwise comparisons, but the stronger hypothesis of rank-order detection achieves small p–values in a substantial fraction of all pairwise comparisons, indicating that first-order stochastic dominance of the latent distributions cannot be taken for granted in these questions.

Let us return to the relation between income and happiness discussed in the Introduction. For the ordered probit model, Table 4 shows that higher income is associated with significantly higher overall happiness when we move from low income ($< \$40,000$) to middle income ($\$40,000$–$\$69,000$). The effect is still positive but much smaller and not significant when we move from middle income to high income ($\geq \$70,000$). In Figure 6, the comparison between low and middle income is depicted as circle 8. We can see that the associated increase in average happiness is clearly detectable in the data ($p = 0.984$), and so is a rank-order of the happiness distributions ($p = 0.328$). The comparison between middle and high income is depicted as circle 9. An additional increase in happiness is no longer detectable, neither in the rank-order nor in the on-average sense (both $p = 0.000$). Our detection results thus support the results of the ordered probit model, according to which there is a positive association between income and happiness (within country at a fixed point in time) for small but not for large incomes. With the appropriate data, our techniques could be used

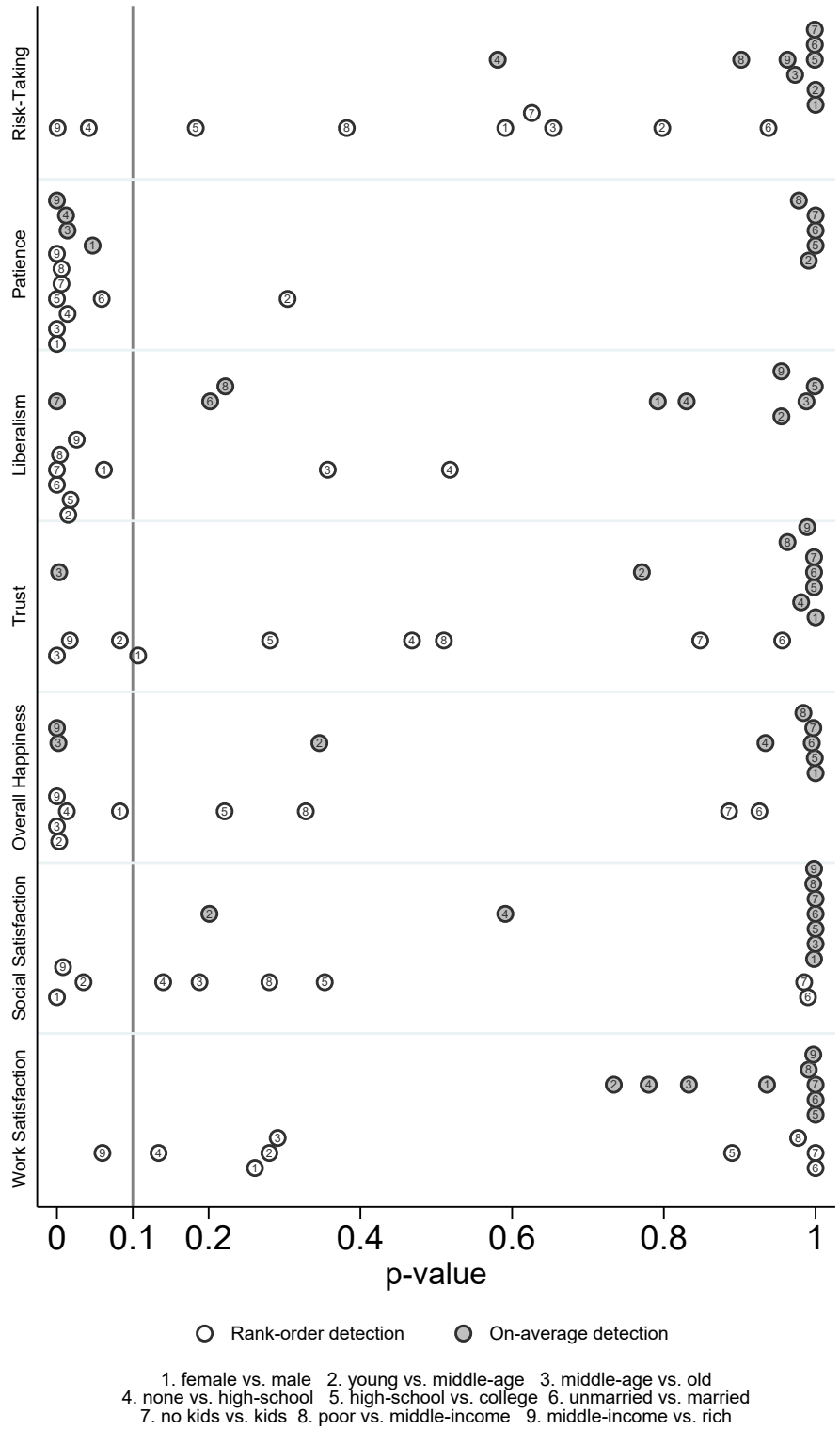


FIGURE 6: Detection analysis in the binary survey.

to examine the income-happiness relation also across countries or over time.

We close this section with two remarks. First, there is no one-to-one relationship between the ability of detection and the significance of the estimated ordered probit coefficient. We sometimes achieve detection but not significance, and sometimes significance but not detection.[18] Altogether, however, the detection analysis seems to support the qualitative validity of the ordered probit estimates. With the exception of one case (see footnote 18), our statistical confidence in correct on-average detection is high whenever the ordered probit coefficient is significant at 10%. These results highlight the value of analysing response times before turning to the standard estimation procedure. Second, we can also check our robust sufficient condition for on-average detection from Proposition 7, which addresses potential issues with between-group heterogeneity in chronometric functions (but remember that our response times are already normalized). The empirical implementation is straightforward. The condition is rejected at 10% significance level for all but four cases.[19]

3.5 Analysis of Trinary Survey

We now turn to the analysis of the survey with three response categories. Our approach is the same as for the binary survey. Table 5 reports the estimated coefficients of the ordered probit model using the trinary survey data. Comparing the estimation results of the two survey versions, we sometimes obtain different parameter signs (5 out of 63 times), but then at least one of the two different estimates is always insignificant. Each of the two survey versions is sometimes "more significant" than the other. Overall, the two versions of the survey seem to generate comparable results based on ordered probit estimation.

Not too surprisingly, the conditions of Proposition 1 are never satisfied, consistent with the findings of Bond and Lang (2019). Just like in the binary survey, even the qualitative results of ordered probit estimation depend on the distributional assumptions of this model. Importantly, condition (*iii*) from Proposition 1 is never satisfied. First, this shows that the problems highlighted by Bond and Lang (2019) are not only due to extreme assumptions on the distribution of the latent variable in the lowest or highest response category. Second, it shows that response times are not helpful for detection in the trinary survey. There is no need to even construct the empirical response time distributions to know that detection fails in

[18] For example, the gender difference in work satisfaction is not significant but rank-order detected ($p = 0.261$) and also on-average detected ($p = 0.936$). The comparison of patience between middle and high incomes is a case where the ordered probit coefficient is significant (with richer subjects being more patient) but detection is not achieved ($p = 0.000$ for both rank-order and on-average detection).

[19] Among these four cases of robust on-average detection, three are for the trust question (the comparisons high-school/college, married/unmarried, and kids/no kids). The other one is for the risk preference question and the comparison high-school/college.

	work satisfac.	social satisfac.	overall happiness	trust	liberal-ism	patience	risk-taking
0: female	-0.007	0.093^{**}	-0.001	0.117^{***}	-0.061^{*}	-0.034	0.232^{***}
1: male	(0.0397)	(0.0386)	(0.0394)	(0.0366)	(0.0368)	(0.0424)	(0.0376)
0: young	0.090^{**}	0.037	0.001	-0.039	-0.101^{**}	-0.021	-0.234^{***}
1: middle-age	(0.0434)	(0.0421)	(0.0431)	(0.0399)	(0.0401)	(0.0463)	(0.0413)
0: middle-age	-0.106	-0.099	0.112	0.164^{**}	-0.009	-0.103	-0.174^{**}
1: old	(0.0777)	(0.0829)	(0.0838)	(0.0791)	(0.0811)	(0.0906)	(0.0795)
0: none	-0.360	-0.462^{*}	0.149	-0.021	-0.108	0.643^{***}	-0.050
1: high-school	(0.2992)	(0.2612)	(0.2882)	(0.2810)	(0.2905)	(0.2221)	(0.2394)
0: high-school	0.625^{***}	0.524^{***}	0.484^{***}	0.547^{***}	0.062	0.194^{***}	0.475^{***}
1: college	(0.0499)	(0.0525)	(0.0521)	(0.0501)	(0.0482)	(0.0568)	(0.0498)
0: unmarried	0.718^{***}	0.669^{***}	0.626^{***}	0.538^{***}	-0.243^{***}	0.150^{***}	0.486^{***}
1: married	(0.0413)	(0.0411)	(0.0412)	(0.0391)	(0.0386)	(0.0443)	(0.0398)
0: no kids	0.662^{***}	0.546^{***}	0.512^{***}	0.436^{***}	-0.217^{***}	0.180^{***}	0.601^{***}
1: kids	(0.0407)	(0.0399)	(0.0404)	(0.0379)	(0.0382)	(0.0435)	(0.0392)
0: poor	0.503^{***}	0.348^{***}	0.375^{***}	0.405^{***}	-0.146^{***}	0.202^{***}	0.354^{***}
1: middle-income	(0.0475)	(0.0471)	(0.0474)	(0.0450)	(0.0449)	(0.0507)	(0.0461)
0: middle-income	0.015	0.013	0.162^{***}	-0.124^{***}	-0.040	0.091^{*}	-0.187^{***}
1: rich	(0.0490)	(0.0471)	(0.0492)	(0.0437)	(0.0446)	(0.0525)	(0.0453)

TABLE 5: Ordered probit analysis of the trinary survey. Each cell corresponds to a regression of the question in the column on a dummy for membership to the group in the row. Coefficients are reported along with their robust standard errors in parentheses. Asterisks indicate statistical significance ($^{*}10\%$,$^{**}5\%$,$^{***}1\%$).

all questions for all pairwise comparisons.[20] This can happen even if the true distributions satisfy FOSD, because Propositions 4 and 6 hold for binary but not for trinary surveys. Hence, it is not a contradiction when we obtain rank-order or on-average detection in the binary but not in the trinary survey.

The stringency of condition (*iii*) obviously does not apply to binary surveys, where intermediate categories do not exist. Given this important advantage, we expect the combination of binary surveys and response time analysis to have great potential in future research.

[20] As pointed out before, we refrain from reporting detection results using follow-up questions based on Proposition 11, because such results would rely on the assumption that the follow-up responses also exhibit the chronometric effect, which, as we discussed in Section 3.3, is not supported by our data.

4 Related Literature

The use of self-reported survey data has long been controversial among economists (see, e.g., Boulier and Goldfarb, 1998; Bertrand and Mullainathan, 2001). A major concern was the fear that self-reported data is not reliable. However, recent studies have shown that surveys can be a reliable source of data. For instance, Falk et al. (2018) have experimentally validated their survey questions, showing that survey responses about preferences predict actual behavior in the lab. In a similar vein, Tannenbaum et al. (2022) have used behavioral data from field experiments to validate survey measures of social capital. The problem forcefully demonstrated by Bond and Lang (2019) is not non-reliability of self-reported data, but that the coarseness of ordered response data gives rise to fundamental identification problems. Several other papers (e.g. Oswald, 2008; Bond and Lang, 2013; Schroeder and Yitzhaki, 2017; Kaiser and Oswald, 2022) make the related point that ordinal data cannot simply be treated as cardinal, and they conclude that results from subjective well-being and test score research, respectively, can be sensitive to the choice of the cardinal scale.

Some recent papers have provided responses to the startling critique of Bond and Lang (2019). For example, Kaiser and Vendrik (2020) argue that, although theoretically possible, reversing standard estimation results using Bond and Lang (2019)'s method may involve conditions that are empirically implausible. Kaplan and Zhuo (2019) show that partial identification of group differences can be possible with semi-parametric assumptions on the latent distributions (e.g. symmetry, unimodality). Chen et al. (2019) propose that analysis of ordinal data should focus on the median instead of the mean, since the ranking of medians between groups is invariant to monotone transformations. In contrast to all these studies, we aim at learning the necessary distributional properties from extended data, rather than judging the plausibility of (semi-)parametric assumptions or reformulating the question.

We are not the first to investigate response times in surveys. For example, Hess and Strathopoulos (2013) assume that survey participants differ in their unobservable engagement with the survey, and that engagement influences both response time (for completing the entire survey) and the individual response scale. Response time data is then useful to control for individual scale heterogeneity. Studer and Winkelmann (2014) show that unhappy participants tend to respond more slowly. Furthermore, they illustrate that including survey response times in happiness regressions modulates the effect of income, but not of other explanatory variables.

More generally, there is a growing interest among economists to explore what can be learned from response times. For instance, Rubinstein (2007, 2016) proposes a typology of choices and players in strategic games based on response times. Achtziger and Alós-

Ferrer (2014) show that response time can measure the extent to which an agent's decision-making process under uncertainty is consistent with the rational paradigm of Bayesian belief-updating. The literature has also suggested that response time data can be used to reveal how decision-makers allocate their limited attention between different problems (Avoyan and Schotter, 2020), to facilitate social learning by serving as an observable signal of agents' private information (Frydman and Krajbich, 2022), to alleviate misspecification bias in the estimation of structural preference parameters (Webb, 2019), and to improve out-of-sample predictions of behavior (Clithero, 2018a; Alós-Ferrer et al., 2021), among several others.

5 Conclusion

In this paper, we have shown that response time data can solve a fundamental identification problem of ordered response models. Since survey data are typically discrete and ordinal, while comparing averages across groups requires continuous and cardinal information, the traditional ordered response models rely on assumptions about the distribution of a latent variable. Their results can change drastically when this distribution is transformed. We have shown, both theoretically and empirically, that response times are a source of information about the distribution of the latent variable. Through the chronometric function, properties of the distribution become observable and distributional assumptions become testable. Our empirical application has shown that the traditional assumptions appear to be a reasonable approximation for some survey questions but less so for others.

We have in mind two ways in which our results can be used in practice. First, surveys are increasingly conducted online, and recording response times is easy and costless in that case. We think that response time data should be collected on par with response data, and their analysis could become a natural part of any investigation. We have repeatedly advocated the use of binary surveys combined with a measurement of response times. Our empirical analysis confirms that identification can become possible that way, while surveys with more than two response categories fail in achieving identification. Of course, causal analysis will be an important concern in many applications, which implies that the groups to be compared have to be much finer than in our simple empirical illustration. One could also try to integrate response time data into a multivariate regression analysis. We leave to future research the question how this could be done, but we conjecture that one could attempt to change the outcome variable in the traditional regression analysis from response to response time, or possibly to response weighted by response time to capture the intensity of the response.

Second, one could use our techniques in auxiliary studies, with the goal of verifying in

a representative sample that the latent variable of interest follows distributions for which traditional ordered response models are appropriate. Once enough evidence of this type has been accumulated, the analyst can proceed as usual and does not have to bother about response time data any more.

References

Achtziger, A. and Alós-Ferrer, C. (2014). Fast or rational? A response-times study of Bayesian updating. *Management Science*, 60(4):923–938.

Alós-Ferrer, C., Fehr, E., and Netzer, N. (2021). Time will tell: Recovering preferences when choices are noisy. *Journal of Political Economy*, 129(6):1828–1877.

Andreoni, J. and Sprenger, C. (2012). Risk preferences are not time preferences. *American Economic Review*, 102(7):3357–3376.

Avoyan, A. and Schotter, A. (2020). Attention in games: An experimental study. *European Economic Review*, 124. Article 103410.

Barrett, G. F. and Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica*, 71(1):71–104.

Bertrand, M. and Mullainathan, S. (2001). Do people mean what they say? Implications for subjective survey data. *American Economic Review Papers and Proceedings*, 91(2):67–72.

Boes, S. and Winkelmann, R. (2006). Ordered response models. *Allgemeines Statistisches Archiv*, 90(1):167–181.

Bond, T. N. and Lang, K. (2013). The evolution of the black-white test score gap in grades k-3: The fragility of results. *Review of Economics and Statistics*, 95(5):1468–1479.

Bond, T. N. and Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy*, 127(4):1629–1640.

Boulier, B. L. and Goldfarb, R. S. (1998). On the use and nonuse of surveys in economics. *Journal of Economic Methodology*, 5(1):1–21.

Chabris, C. F., Morris, C. L., Taubinsky, D., Laibson, D., and Schuldt, J. P. (2009). The allocation of time in decision-making. *Journal of the European Economic Association*, 7(2-3):628–637.

Chen, L.-Y., Oparina, E., Powdthavee, N., and Srisuma, S. (2019). Have econometric analyses of happiness data been futile? A simple truth about happiness scales. IZA Discussion Paper No. 12152.

Clithero, J. A. (2018a). Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior and Organization*, 148:344–375.

Clithero, J. A. (2018b). Response times in economics: Looking through the lens of sequential sampling models. *Journal of Economic Psychology*, 69:61–86.

Davis, J. A. and Smith, T. W. (1991). *The NORC General Social Survey: A User's Guide*. Newbury Park: Sage Publications.

DellaVigna, S. and Pope, D. (2018). What motivates effort? Evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069.

Easterlin, R. A. (1974). Does economic growth improve the human lot? Some empirical evidence. In David, P. A. and Reder, M. W., editors, *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, pages 89–125. Academic Press, New York.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.

Frydman, C. and Krajbich, I. (2022). Using response times to infer others' beliefs: An application to information cascades. *Management Science*, 68(4):2970–2986.

Goodman, J. K. and Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research*, 44(1):196–210.

Hanock, G. and Levy, H. (1969). The efficiency analysis of choices involving risk. *Review of Economic Studies*, 36(3):335–346.

Hess, S. and Strathopoulos, A. (2013). Linking response quality to survey engagement: A combined random scale and latent variable approach. *Journal of Choice Modelling*, 7:1–12.

Kaiser, C. (2022). Using memories to assess the intrapersonal comparability of well-being reports. *Journal of Economic Behavior and Organization*, 193:410–442.

Kaiser, C. and Oswald, A. J. (2022). Inequality, well-being, and the problem of the unknown reporting function. *Proceedings of the National Academy of Sciences*, 119(50).

Kaiser, C. and Vendrik, M. C. (2020). How threatening are transformations of reported happiness to subjective wellbeing research? IZA Discussion Paper No. 13905.

Kaplan, D. M. and Zhuo, L. (2019). Comparing latent inequality with ordinal data. Mimeo.

Kellogg, W. N. (1931). The time of judgment in psychometric measures. *American Journal of Psychology*, 43(1):65–86.

Konovalov, A. and Krajbich, I. (2019). Revealed strength of preference: Inference from response times. *Judgment & Decision Making*, 14(4):381–394.

Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10):1292–1298.

Kuziemko, I., Norton, M. I., Saez, E., and Stantcheva, S. (2015). How elastic are preferences for redistribution? Evidence from randomized survey experiments. *American Economic Review*, 105(4):1478–1508.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22:5–55.

Liu, S. and Netzer, N. (2020). Happy times: Identification from ordered response data. University of Zurich, Department of Economics, Working Paper No. 371.

Moffatt, P. G. (2005). Stochastic choice and the allocation of cognitive effort. *Experimental Economics*, 8(4):369–388.

Moyer, R. S. and Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology*, 8(2):228–246.

Oswald, A. J. (2008). On the curvature of the reporting function from objective reality to subjective feelings. *Economics Letters*, 100:369–372.

Palmer, J., Huk, A. C., and Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, 5:376–404.

Paolacci, G. and Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188.

Rossi, P. H., Wright, J. D., and Anderson, A. B. (1983). Sample surveys: History, current practice, and future prospects. In Rossi, P. H., Wright, J. D., and Anderson, A. B., editors, *Handbook of Survey Research*, chapter 1, pages 1–20. Academic Press, New York.

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *The Economic Journal*, 117:1243–1259.

Rubinstein, A. (2016). A typology of players: Between instinctive and contemplative. *The Quarterly Journal of Economics*, 131(2):859–890.

Schroeder, C. and Yitzhaki, S. (2017). Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, 92:337–358.

Stevenson, B. and Wolfers, J. (2008). Economic growth and subjective well-being: Reassessing the Easterlin paradox. *Brookings Papers on Economic Activity*, Spring 2008:1–87.

Studer, R. and Winkelmann, R. (2014). Reported happiness, fast and slow. *Social Indicators Research*, 117:1055–1067.

Tannenbaum, D., Cohn, A., Zünd, C., and Maréchal, M. A. (2022). What do cross-country surveys tell us about social capital? *Review of Economics and Statistics*, forthcoming.

Webb, R. (2019). The (neural) dynamics of stochastic choice. *Management Science*, 65(1):230–255.

A Omitted Proofs

A.1 On-Average Detection with Ordered Responses

On-average detection requires that

$$\mathbb{E}_{G_A}[\tilde{h}] \geq \mathbb{E}_{G_B}[\tilde{h}],$$

for all (G_A, G_B, τ) that generate the data. A conceptually stronger requirement would be

$$\mathbb{E}_{G_A}[q(\tilde{h})] \geq \mathbb{E}_{G_B}[q(\tilde{h})],$$

for all (G_A, G_B, τ) that generate the data and all strictly increasing $q : \mathbb{R} \to \mathbb{R}$. We claim, however, that these two requirements are equivalent.

The second requirement obviously implies the first by using $q(h) = h$. To see why the first requirement implies the second, assume that $\mathbb{E}_{G_A}[q(\tilde{h})] < \mathbb{E}_{G_B}[q(\tilde{h})]$ for some (G_A, G_B, τ) that generates the data and some strictly increasing q. Let $\hat{G}_j$ describe the distribution of $q(\tilde{h})$ under G_j, which exists and is continuous because q is strictly increasing. It satisfies $\mathbb{E}_{\hat{G}_j}[\tilde{h}] = \mathbb{E}_{G_j}[q(\tilde{h})]$ by construction. Define $\hat{\tau} = (\hat{\tau}^1, \hat{\tau}^2, \ldots, \hat{\tau}^n)$ by $\hat{\tau}^i = q(\tau^i)$. It follows that

$$\begin{aligned}
\hat{G}_j(\hat{\tau}^{i+1}) - \hat{G}_j(\hat{\tau}^i) &= \Pr[q(\tilde{h}) \leq \hat{\tau}^{i+1}] - \Pr[q(\tilde{h}) \leq \hat{\tau}^i] \\
&= \Pr[\tilde{h} \leq \tau^{i+1}] - \Pr[\tilde{h} \leq \tau^i] \\
&= G_j(\tau^{i+1}) - G_j(\tau^i),
\end{aligned}$$

so that $(\hat{G}_A, \hat{G}_B, \hat{\tau})$ generates the data and satisfies $\mathbb{E}_{\hat{G}_A}[\tilde{h}] < \mathbb{E}_{\hat{G}_B}[\tilde{h}]$, which implies that the first requirement is also violated.

A.2 Example for On-Average Detection with $n = 2$

Suppose that the true distribution of happiness in group A is given by

$$G_A(h) = \begin{cases} 0 & \text{if } h < -1.2, \\ \epsilon(h + 1.2) & \text{if } -1.2 \leq h < 0.2, \\ (1.75\epsilon - 1.25)(1 - h) + 1 & \text{if } 0.2 \leq h < 1, \\ 1 & \text{if } h \geq 1, \end{cases}$$

where $\epsilon \in (0, 5/7)$, so G_A is well-defined. The distribution in group B is

$$G_B(h) = \begin{cases} 0 & \text{if } h < -1, \\ \frac{h}{2} + \frac{1}{2} & \text{if } -1 \leq h < 1, \\ 1 & \text{if } h \geq 1. \end{cases}$$

We assume that ϵ is sufficiently small such that $\mathbb{E}_{G_A}[\tilde{h}] > 0 = \mathbb{E}_{G_B}[\tilde{h}]$, so group A is happier than group B on average.

Consider first a survey with $n = 1$. Suppose that $\tau^1 = 0$ is the reporting threshold and $c(\delta) = 1/\delta$ the chronometric function of the true data-generating process. Then, for response time $t = 1$, we have

$$r_A^0 F_A^0(1) - r_B^0 F_B^0(1) = G_A(\tau^1 - c^{-1}(1)) - G_B(\tau^1 - c^{-1}(1)) = G_A(-1) - G_B(-1) = 0.2\epsilon$$

and

$$r_A^1 F_A^1(1) - r_B^1 F_B^1(1) = G_B(\tau^1 + c^{-1}(1)) - G_A(\tau^1 + c^{-1}(1)) = G_B(1) - G_A(1) = 0.$$

Hence, the data generated by the binary survey will violate the condition in Proposition 5 or Corollary 3, leading to a failure in achieving on-average detection.

Now consider a survey with $n = 2$. Suppose that in this case the reporting thresholds are $\tau^1 = -0.5$ and $\tau^2 = 0.2$ (so one threshold is larger and one smaller than the unique threshold considered in the binary case), while the chronometric function is still $c(\delta) = 1/\delta$. Condition (ii) in Proposition 5 then requires

$$r_A^0 + r_A^1 = G_A(\tau^2) = 1.4\epsilon < r_B^0 = G_B(\tau^1) = 1/4,$$

which is satisfied whenever ϵ is sufficiently small. We will now show that, if ϵ is sufficiently small, condition (i) in Proposition 5 is also satisfied. Hence, the data generated from the survey with $n = 2$ will satisfy the conditions in Proposition 5 and we can correctly detect that group A is on-average happier than group B. We distinguish two cases.

Case 1: $t \in [10/7, +\infty)$. In this case we have $-1.2 \leq \tau^1 - c^{-1}(t) < -0.5$, and therefore

$$\begin{aligned} r_A^0 F_A^0(t) - r_B^0 F_B^0(t) &= G_A(\tau^1 - c^{-1}(t)) - G_B(\tau^1 - c^{-1}(t)) \\ &\leq G_A(-0.5 - 1/t) \\ &= \epsilon(0.7 - 1/t). \end{aligned}$$

Furthermore, $0.2 < \tau^2 + c^{-1}(t) \leq 0.9$, and therefore

$$\begin{aligned} r_A^2 F_A^2(t) - r_B^2 F_B^2(t) &= G_B(\tau^2 + c^{-1}(t)) - G_A(\tau^2 + c^{-1}(t)) \\ &= G_B(0.2 + 1/t) - G_A(0.2 + 1/t) \\ &= \frac{1.2 + 1/t}{2} - (1.75\epsilon - 1.25)(0.8 - 1/t) - 1 \\ &= \frac{1.2 + 1/t}{2} + \frac{1.75\epsilon - 1.25}{t} - 1.4\epsilon \\ &= 0.6 - \frac{0.75}{t} + \epsilon\left(\frac{1.75}{t} - 1.4\right). \end{aligned}$$

Since $t \geq 10/7$, it follows that

$$\begin{aligned} &\lim_{\epsilon \to 0}\left[0.6 - \frac{0.75}{t} + \epsilon\left(\frac{1.75}{t} - 1.4\right) - \epsilon(0.7 - 1/t)\right] \\ =& \frac{3}{5} - \frac{3}{4t} \\ \geq& \frac{3}{5} - \frac{21}{40} \\ =& 0.075. \end{aligned}$$

Therefore, if ϵ is sufficiently small, we have, for all $t \in [10/7, +\infty)$,

$$r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq r_A^2 F_A^2(t) - r_B^2 F_B^2(t).$$

Case 2: $t \in (0, 10/7)$. In this case we have $\tau^1 - c^{-1}(t) < -1.2$, and therefore

$$r_A^0 F_A^0(t) - r_B^0 F_B^0(t) = G_A(\tau^1 - c^{-1}(t)) - G_B(\tau^1 - c^{-1}(t)) = 0.$$

Furthermore, $0.9 < \tau^2 + c^{-1}(t)$, and therefore

$$r_A^2 F_A^2(t) - r_B^2 F_B^2(t) = G_B(\tau^2 + c^{-1}(t)) - G_A(\tau^2 + c^{-1}(t)) \geq 0,$$

where the inequality follows for $0.9 < \tau^2 + c^{-1}(t) < 1$ because

$$\left(\frac{h}{2} + \frac{1}{2}\right) - (1.75\epsilon - 1.25)(1 - h) - 1 = (1.75 - 1.75\epsilon)(1 - h) > 0$$

whenever $0.9 < h < 1$, and it follows trivially when $1 \leq \tau^2 + c^{-1}(t)$.

A.3 Robustness Results

Proof of Proposition 7. Let $(G_A, G_B, \tau^1, c_A, c_B)$ be any possible data-generating process. Under the stated condition, it follows exactly like in the proof of Theorem 1 in Alós-Ferrer et al. (2021) that $\mathbb{E}_{G_B}[\tilde{h}] \leq \tau^1 \leq \mathbb{E}_{G_A}[\tilde{h}]$ must hold. □

Proof of Proposition 8. Let $(G_A, G_B, \tau^1, c, \phi)$ be any data-generating process. We have $\hat{t} = t/t^b = c(|\tau^1 - h|)/\phi$. It follows that $r_j^0 \hat{F}_j^0(t) = G_j(\tau^1 - c^{-1}(\phi \cdot t))$ and $r_j^1 \hat{F}_j^1(t) = 1 - G_j(\tau^1 + c^{-1}(\phi \cdot t))$ whenever $\underline{t} < \phi \cdot t < \bar{t}$, where G_j refers to the marginal distribution of happiness.

Consider first the if-statement for rank-order detection. The condition stated in the proposition implies

$$G_A(\tau^1 - c^{-1}(\phi \cdot t)) - G_B(\tau^1 - c^{-1}(\phi \cdot t)) \leq 0$$

whenever $\underline{t} < \phi \cdot t < \bar{t}$. We claim that this implies $G_A(h) \leq G_B(h)$ for all $h \leq \tau^1$. This is immediate for any (large enough) h for which there exists t with $\underline{t} < \phi \cdot t < \bar{t}$ such that $h = \tau^1 - c^{-1}(\phi \cdot t)$. For $h = \tau^1$ it follows from continuity of G_j. For any h with $c(\tau^1 - h) = \underline{t}$ it follows because $G_j(h) = 0$ in that case, as there are no atoms in the distributions of normalized response times. By an analogous argument, we also conclude that $G_A(h) \leq G_B(h)$ for all $h \geq \tau^1$, which establishes first-order stochastic dominance. Consider then the only-if-statement for rank-order detection. Suppose that $r_A^0 \hat{F}_A^0(t^*) - r_B^0 \hat{F}_B^0(t^*) > 0$ for some $t^* > 0$. Note that $\hat{F}_A^0(t^*) > 0$ must be true in that case. Furthermore, $\hat{F}_j^0(t^*) < 1$ can be assumed for at least one $j = A, B$ without loss of generality (because otherwise $r_A^0 > r_B^0$ and, by continuity of $\hat{F}_j^0$, we can decrease t^* until we get the desired property). Hence $\underline{t} < \phi \cdot t^* < \bar{t}$ holds in any possible data-generating process (if not, smaller or larger normalized response times than t^* could not be generated). From the above arguments we can conclude that $G_A(\tau^1 - c^{-1}(\phi \cdot t^*)) > G_B(\tau^1 - c^{-1}(\phi \cdot t^*))$, so that G_A FOSD G_B is not true. An analogous argument applies when $r_A^1 \hat{F}_A^1(t^*) - r_B^1 \hat{F}_B^1(t^*) < 0$ for some $t^* > 0$.

Consider then the condition for on-average detection. It implies

$$G_B(\tau^1 + c^{-1}(\phi \cdot t)) + G_B(\tau^1 - c^{-1}(\phi \cdot t)) - G_A(\tau^1 + c^{-1}(\phi \cdot t)) - G_A(\tau^1 - c^{-1}(\phi \cdot t)) \geq 0$$

whenever $\underline{t} < \phi \cdot t < \bar{t}$. It follows as above that

$$G_B(\tau^1 + h) + G_B(\tau^1 - h) - G_A(\tau^1 + h) - G_A(\tau^1 - h) \geq 0$$

for all $h > 0$. The rest of the proof is analogous to Proposition 5. □

Proof of Proposition 9. Given the true data-generating process, define

$$P_j^0(t,\eta) = \begin{cases} G_j(\tau^1) & \text{if } t/\eta \geq \bar{t}, \\ G_j(\tau^1 - c^{-1}(t/\eta)) & \text{if } \underline{t} < t/\eta < \bar{t}, \\ 0 & \text{if } t/\eta \leq \underline{t}, \end{cases}$$

and

$$P_j^1(t,\eta) = \begin{cases} 1 - G_j(\tau^1) & \text{if } t/\eta \geq \bar{t}, \\ 1 - G_j(\tau^1 + c^{-1}(t/\eta)) & \text{if } \underline{t} < t/\eta < \bar{t}, \\ 0 & \text{if } t/\eta \leq \underline{t}, \end{cases}$$

for all $t > 0$ and $\eta > 0$. Since the random variable $\tilde{\eta}$ is independent of happiness, we obtain

$$r_j^0 F_j^0(t) = \mathbb{E}\left[P_j^0(t,\tilde{\eta})\right] \text{ and } r_j^1 F_j^1(t) = \mathbb{E}\left[P_j^1(t,\tilde{\eta})\right]$$

for all $t > 0$, where the expectation is with respect to $\tilde{\eta}$. Since $\tilde{\eta}$ is i.i.d. in the two groups, we can write

$$r_A^0 F_A^0(t) - r_B^0 F_B^0(t) = \mathbb{E}\left[P_A^0(t,\tilde{\eta}) - P_B^0(t,\tilde{\eta})\right] \leq 0,$$

where the inequality holds because G_A FOSD G_B implies $P_A^0(t,\eta) \leq P_B^0(t,\eta)$ for all $t > 0$ and $\eta > 0$, so that we are taking the expectation of a weakly negative function. The analogous argument implies $r_A^1 F_A^1(t) - r_B^1 F_B^1(t) = \mathbb{E}[P_A^1(t,\tilde{\eta}) - P_B^1(t,\tilde{\eta})] \geq 0$. □

Proof of Proposition 10. Fix the true data-generating process and denote $\mu_j = \mathbb{E}_{G_j}[\tilde{h}]$ for $j = A, B$. Using the same notation as in the proof of Proposition 9, we obtain, for all $t > 0$ and $\eta > 0$,

$$\begin{aligned} P_A^0(t,\eta) &= \begin{cases} G_A(\tau^1) & \text{if } t/\eta \geq \bar{t}, \\ G_A(\tau^1 - c^{-1}(t/\eta)) & \text{if } \underline{t} < t/\eta < \bar{t}, \\ 0 & \text{if } t/\eta \leq \underline{t}, \end{cases} \\ &= \begin{cases} 1 - G_A(2\mu_A - \tau^1) & \text{if } t/\eta \geq \bar{t}, \\ 1 - G_A(2\mu_A - \tau^1 + c^{-1}(t/\eta)) & \text{if } \underline{t} < t/\eta < \bar{t}, \\ 0 & \text{if } t/\eta \leq \underline{t}, \end{cases} \end{aligned}$$

$$
\begin{aligned}
&\leq \begin{cases} 1 - G_A(\tau^1) & \text{if } t/\eta \geq \bar{t}, \\ 1 - G_A(\tau^1 + c^{-1}(t/\eta)) & \text{if } \underline{t} < t/\eta < \bar{t}, \\ 0 & \text{if } t/\eta \leq \underline{t}, \end{cases} \\
&= P_A^1(t, \eta),
\end{aligned}
$$

where the second equality follows from symmetry of G_A and the inequality follows from the assumption that $\tau^1 \leq \mu_A$. Analogous arguments reveal that $P_B^0(t, \eta) \geq P_B^1(t, \eta)$. Hence

$$
\begin{aligned}
r_A^0 F_A^0(t) - r_B^0 F_B^0(t) &= \mathbb{E}_A \left[P_A^0(t, \tilde{\eta}) \right] - \mathbb{E}_B \left[P_B^0(t, \tilde{\eta}) \right] \\
&\leq \mathbb{E}_A \left[P_A^1(t, \tilde{\eta}) \right] - \mathbb{E}_B \left[P_B^1(t, \tilde{\eta}) \right] \\
&= r_A^1 F_A^1(t) - r_B^1 F_B^1(t),
\end{aligned}
$$

where the expectations are with respect to the group-specific distribution of $\tilde{\eta}$. □

A.4 Follow-Up Questions

Let $(G_A, G_B, \tau, c, (\check{\tau}^i, \check{c}^i)_i)$ be any process that could have generated the data. Exactly as in the proof of Proposition 3, conditions (i) and (ii) are equivalent to $G_A(h) \leq G_B(h)$ for all $h \leq \tau^1$ and $h > \tau^n$. We argue that conditions (iii) and (iv) are jointly necessary and sufficient for concluding that we also have $G_A(h) \leq G_B(h)$ for all $h \in (\tau^1, \tau^n]$.

To see this, note that for all $j = A, B$, $k = 1, ..., n-1$, and $t \in (\underline{t}, \bar{t})$, we have

$$
\sum_{i=0}^{k-1} r_j^i + \underline{r}_j^k \underline{F}_j^k(t) = G_j(\tau^k) + \max\{0, G_j(\check{\tau}^k - (\check{c}^k)^{-1}(t)) - G_j(\tau^k)\}.
$$

For large enough t such that $\check{\tau}^k - (\check{c}^k)^{-1}(t) > \tau^k$, we can skip the max-operator and condition (iii) is equivalent to

$$
G_A\left(\check{\tau}^k - (\check{c}^k)^{-1}(t)\right) \leq G_B\left(\check{\tau}^k - (\check{c}^k)^{-1}(t)\right).
$$

If that case applies to all $t \in (\underline{t}, \bar{t})$, it follows similar to the proof of Proposition 3 that (iii) is necessary and sufficient for obtaining that $G_A(h) \leq G_B(h)$ for all $h \in (\tau^k, \check{\tau}^k]$. If there exist small t for which $\check{\tau}^k - (\check{c}^k)^{-1}(t) \leq \tau^k$, condition (iii) immediately implies $G_A(h) \leq G_B(h)$ for all $h \in (\tau^k, \check{\tau}^k]$. The converse implication follows because $\underline{r}_j^k \underline{F}_j^k(t) = 0$ for these small t,

and therefore (iii) for all these t becomes

$$\sum_{i=0}^{k-1} r_A^i = G_A(\tau^k) \leq G_B(\tau^k) = \sum_{i=0}^{k-1} r_B^i,$$

which is true by continuity of G_j. Similarly,

$$\sum_{i=k+1}^{n} r_j^i + \overline{r}_j^k \overline{F}_j^k(t) = 1 - G_j(\tau^{k+1}) + \max\{0, G_j(\tau^{k+1}) - G_j(\check{\tau}^k + (\check{c}^k)^{-1}(t))\}$$

holds for all $j = A, B$, $k = 1, ..., n-1$, and $t \in (\underline{t}, \bar{t})$. Analogous arguments to the above then imply that condition (iv) is equivalent to $G_A(h) \leq G_B(h)$ also for all $h \in (\check{\tau}^k, \tau^{k+1}]$.

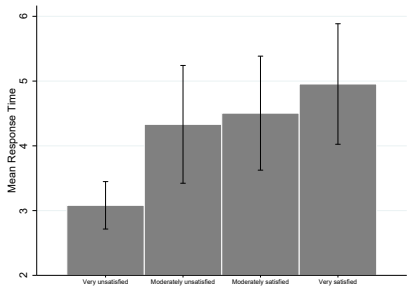
B Additional Tables and Figures

	Response Time			
	(1)	(2)	(3)	(4)
Follow-Up Response	-0.804***	-0.913***	-0.807***	-0.657***
	(0.1736)	(0.1715)	(0.1703)	(0.2036)
R-squared	0.0045	0.0055	0.0052	0.0052
Demographics & Treatment	NO	YES	YES	NO
Individual RE	NO	NO	YES	NO
Individual FE	NO	NO	NO	YES

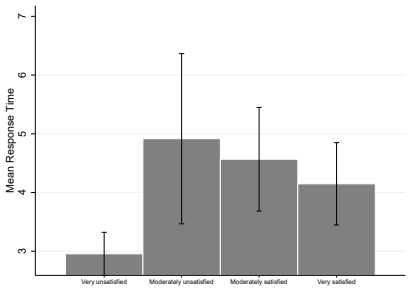
*** p<0.01, ** p<0.05, * p<0.1.

Notes: All regressions include all observations from the binary survey and the observations with non-intermediate responses from the trinary survey. The dependent variable is each subject's response time in the initial substantive question (not including the follow-up). Follow-Up Response is a dummy that takes the value one if the subject chose the extreme response (e.g. "very happy" or "very unhappy") in the corresponding follow-up question. All regressions include question fixed effects. The demographic controls are gender, age, education, marital status, co-residence with children, and family income. Treatment is a dummy for the survey version (binary versus trinary). Column (3) is a random-effect model with all demographic and treatment controls. Column (4) is a fixed-effect model which controls for heterogeneity at the subject level. Robust standard errors are reported in parentheses, with the ones in columns (1) and (2) being clustered at the subject level. The R-squared values reported in columns (3) and (4) concern the variation within subjects.

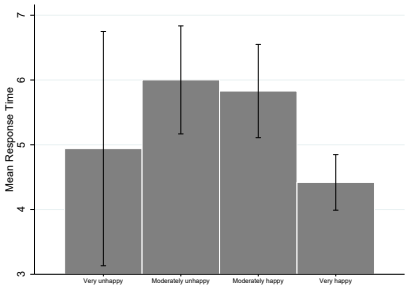
TABLE B1: Regression analysis of chronometric effects (raw data).



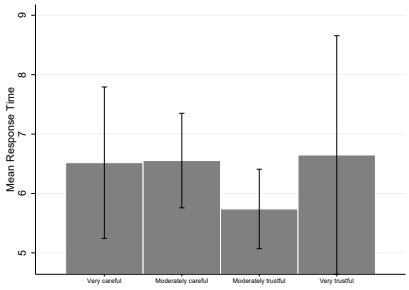
(A) Work Satisfaction



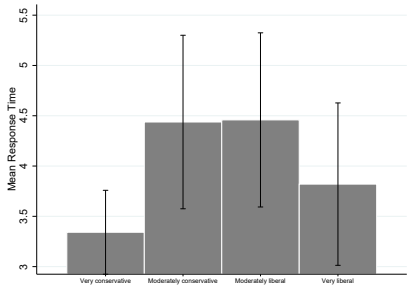
(B) Social Life Satisfaction



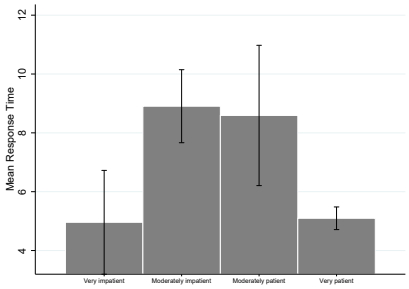
(C) Overall Happiness



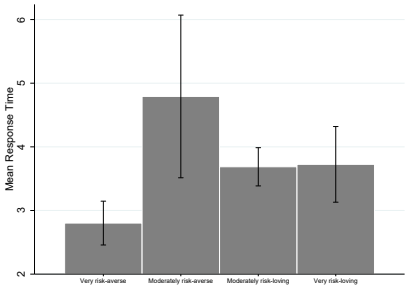
(D) Trust



(E) Political Attitude



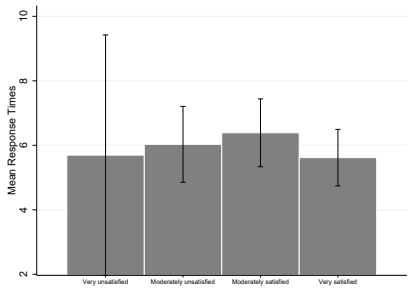
(F) Time Preference



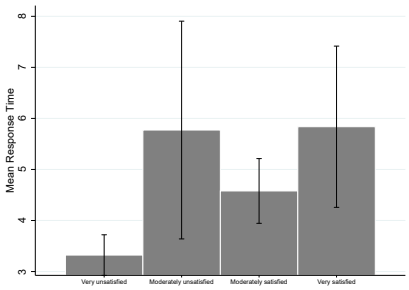
(G) Risk Preference

Notes: The figure displays, for each substantive question in the binary survey, the average response time of the subjects, categorized by their response to the initial and the follow-up question. Black lines indicate 95% confidence intervals.

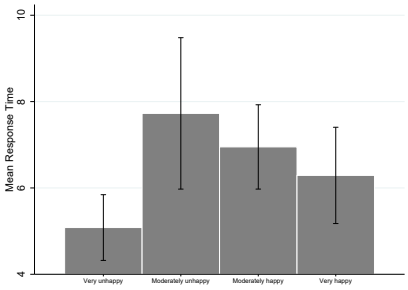
FIGURE B1: Chronometric effect by question in the binary survey (raw data).



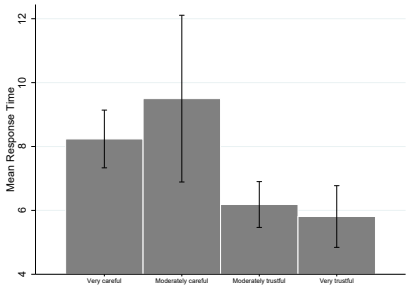
(A) Work Satisfaction



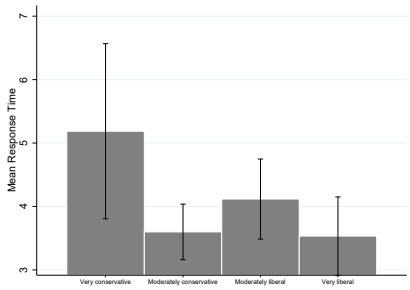
(B) Social Life Satisfaction



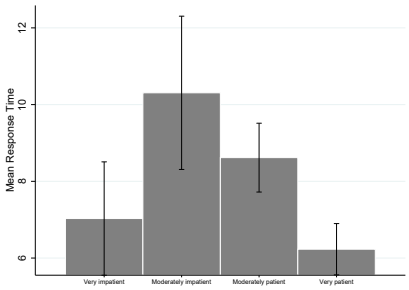
(C) Overall Happiness



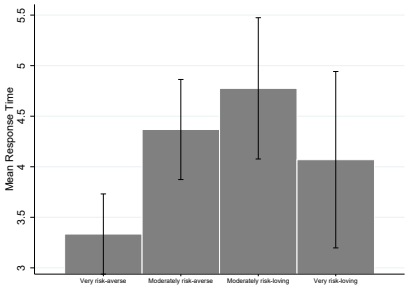
(D) Trust



(E) Political Attitude



(F) Time Preference



(G) Risk Preference

Notes: The figure displays, for each substantive question in the trinary survey, the average response time of the subjects, categorized by their (non-intermediate) response to the initial and the follow-up question. Black lines indicate 95% confidence intervals.

FIGURE B2: Chronometric effect by question in the trinary survey (raw data).

	Log Normalized RT Follow-Up			
	(1)	(2)	(3)	(4)
Log Normalized RT Initial	0.595***	0.589***	0.465***	0.117***
	(0.0126)	(0.0130)	(0.0126)	(0.0102)
R-squared	0.3783	0.3828	0.0917	0.1250
Demographics & Treatment	NO	YES	YES	NO
Individual RE	NO	NO	YES	NO
Individual FE	NO	NO	NO	YES

*** p<0.01, ** p<0.05, * p<0.1.

Notes: All regressions include all observations with extreme responses to both the initial and the related follow-up question (e.g. "rather unhappy" followed by "very unhappy") from the binary and the trinary survey. The dependent variable is each subject's log normalized response time in the follow-up question. Log Normalized RT Initial is the subject's log normalized response time in the initial question. All regressions include question fixed effects. The demographic controls are gender, age, education, marital status, co-residence with children, and family income. Treatment is a dummy for the survey version (binary versus trinary). Column (3) is a random-effect model with all demographic and treatment controls. Column (4) is a fixed-effect model which controls for heterogeneity at the subject level. Robust standard errors are reported in parentheses, with the ones in columns (1) and (2) being clustered at the subject level. The R-squared values reported in columns (3) and (4) concern the variation within subjects.

TABLE B2: Response time correlation, extreme follow-up response.

	Log Normalized RT Follow-Up			
	(1)	(2)	(3)	(4)
Log Normalized RT Initial	0.523***	0.511***	0.367***	0.071***
	(0.0153)	(0.0153)	(0.0133)	(0.0117)
R-squared	0.4132	0.4224	0.3771	0.4158
Demographics & Treatment	NO	YES	YES	NO
Individual RE	NO	NO	YES	NO
Individual FE	NO	NO	NO	YES

*** p<0.01, ** p<0.05, * p<0.1.

Notes: All regressions include all observations with an extreme response to the initial question and a moderate response to the related follow-up question (e.g. "rather unhappy" followed by "moderately unhappy") from the binary and the trinary survey. The dependent variable is each subject's log normalized response time in the follow-up question. Log Normalized RT Initial is the subject's log normalized response time in the initial question. All regressions include question fixed effects. The demographic controls are gender, age, education, marital status, co-residence with children, and family income. Treatment is a dummy for the survey version (binary versus trinary). Column (3) is a random-effect model with all demographic and treatment controls. Column (4) is a fixed-effect model which controls for heterogeneity at the subject level. Robust standard errors are reported in parentheses, with the ones in columns (1) and (2) being clustered at the subject level. The R-squared values reported in columns (3) and (4) concern the variation within subjects.

TABLE B3: Response time correlation, moderate follow-up response.

	Response Time Follow-Up			
	(1)	(2)	(3)	(4)
Response Time Initial	0.083***	0.083***	0.083***	0.036
	(0.0227)	(0.0227)	(0.0227)	(0.0229)
R-squared	0.0136	0.0142	0.0082	0.0094
Demographics & Treatment	NO	YES	YES	NO
Individual RE	NO	NO	YES	NO
Individual FE	NO	NO	NO	YES

*** p<0.01, ** p<0.05, * p<0.1.

Notes: All regressions include all observations with extreme responses to both the initial and the related follow-up question (e.g. "rather unhappy" followed by "very unhappy") from the binary and the trinary survey. The dependent variable is each subject's response time in the follow-up question. Response Time Initial is the subject's response time in the initial question. All regressions include question fixed effects. The demographic controls are gender, age, education, marital status, co-residence with children, and family income. Treatment is a dummy for the survey version (binary versus trinary). Column (3) is a random-effect model with all demographic and treatment controls. Column (4) is a fixed-effect model which controls for heterogeneity at the subject level. Robust standard errors are reported in parentheses, with the ones in columns (1) and (2) being clustered at the subject level. The R-squared values reported in columns (3) and (4) concern the variation within subjects.

TABLE B4: Response time correlation, extreme follow-up response (raw data).

	Response Time Follow-Up			
	(1)	(2)	(3)	(4)
Response Time Initial	0.038***	0.037***	0.037***	0.003
	(0.0134)	(0.0129)	(0.0129)	(0.0082)
R-squared	0.0490	0.0520	0.0516	0.0545
Demographics & Treatment	NO	YES	YES	NO
Individual RE	NO	NO	YES	NO
Individual FE	NO	NO	NO	YES

*** p<0.01, ** p<0.05, * p<0.1.

Notes: All regressions include all observations with an extreme response to the initial question and a moderate response to the related follow-up question (e.g. "rather unhappy" followed by "moderately unhappy") from the binary and the trinary survey. The dependent variable is each subject's response time in the follow-up question. Response Time Initial is the subject's response time in the initial question. All regressions include question fixed effects. The demographic controls are gender, age, education, marital status, co-residence with children, and family income. Treatment is a dummy for the survey version (binary versus trinary). Column (3) is a random-effect model with all demographic and treatment controls. Column (4) is a fixed-effect model which controls for heterogeneity at the subject level. Robust standard errors are reported in parentheses, with the ones in columns (1) and (2) being clustered at the subject level. The R-squared values reported in columns (3) and (4) concern the variation within subjects.

TABLE B5: Response time correlation, moderate follow-up response (raw data).

C Questionnaires

This appendix contains the exact phrasing of all questions and possible answers from our MTurk survey, in the order in which they appeared. A difference between the binary and the trinary version of the survey exists only for the substantive questions.

1. **Welcome Screen**

   Welcome!

   This survey is carried out for a research project at the University of Zurich, Switzerland. The survey is for scientific purposes only.

   There are no known risks for you if you decide to participate in this survey, nor will you experience any costs when participating in the survey. This survey is anonymous. The information you provide will not be stored or used in any way that could reveal your personal identity.

   For more information please contact descil@ethz.ch.

   Answer possibilities:

   - I have read and understood the consent form and agree to participate in this survey.

2. **Socio-Demographic Question 1: Gender**

   What is your gender?

   Answer possibilities:

   - Female
   - Male

3. **Socio-Demographic Question 2: Age**

   What is your age?

   Answer possibilities:

   - younger than 20
   - $20 - 29$
   - $30 - 39$
   - $40 - 49$

- $50 - 59$
- $60 - 69$
- 70 or older

4. **Socio-Demographic Question 3: Education**

   What is the highest level of education that you completed?

   Answer possibilities:

   - High school
   - College degree or higher
   - None of the above

5. **Socio-Demographic Question 4: Marital Status**

   What is your current marital status?

   Answer possibilities:

   - Married
   - Unmarried

6. **Socio-Demographic Question 5: Children**

   Are there any children currently living in your household?

   Answer possibilities:

   - Yes
   - No

7. **Socio-Demographic Question 6: Income**

   In which of these groups did your total family income, from all sources, fall last year before taxes?

   Answer possibilities:

   - Under $ 40,000
   - $ 40,000 to 69,999
   - $ 70,000 or over

8. **Substantive Question 1: Work Satisfaction**

   How satisfied are you with the work you do?

   Answer possibilities binary:

   - Rather satisfied
   - Rather unsatisfied

   Answer possibilities trinary:

   - Rather satisfied
   - Neither satisfied nor unsatisfied
   - Rather unsatisfied

   Follow-up:

   You have answered that you are rather satisfied with the work you do in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how satisfied you are with the work you do?

   Answer possibilities:

   - Very satisfied
   - Moderately satisfied

   You have answered that you are neither satisfied nor unsatisfied with the work you do in the previous question. We now want to follow up on that question and ask you for a more refined answer. Concerning how satisfied you are with the work you do, in which direction do you tend more?

   Answer possibilities:

   - Tend more toward satisfied
   - Tend more toward unsatisfied

   You have answered that you are rather unsatisfied with the work you do in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how unsatisfied you are with the work you do?

   Answer possibilities:

   - Very unsatisfied

- Moderately unsatisfied

9. **Substantive Question 2: Social Life Satisfaction**

   How satisfied are you with your social life?

   Answer possibilities binary:

   - Rather satisfied
   - Rather unsatisfied

   Answer possibilities trinary:

   - Rather satisfied
   - Neither satisfied nor unsatisfied
   - Rather unsatisfied

   Follow-up:

   You have answered that you are rather satisfied with your social life in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how satisfied you are with your social life?

   Answer possibilities:

   - Very satisfied
   - Moderately satisfied

   You have answered that you are neither satisfied nor unsatisfied with your social life in the previous question. We now want to follow up on that question and ask you for a more refined answer. Concerning how satisfied you are with your social life, in which direction do you tend more?

   Answer possibilities:

   - Tend more toward satisfied
   - Tend more toward unsatisfied

   You have answered that you are rather unsatisfied with your social life in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how satisfied you are with your social life?

   Answer possibilities:

- Very unsatisfied
- Moderately unsatisfied

10. **Substantive Question 3: Overall Happiness**

    Binary: Taken all together, how would you say things are these days? Would you say that you are rather happy or rather unhappy?

    Trinary: Taken all together, how would you say things are these days? Would you say that you are rather happy, neither happy nor unhappy, or rather unhappy?

    Answer possibilities binary:

    - Rather happy
    - Rather unhappy

    Answer possibilities trinary:

    - Rather happy
    - Neither happy nor unhappy
    - Rather unhappy

    Follow-up:

    You have answered that you are rather happy in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how you are these days, taken all together?

    Answer possibilities:

    - Very happy
    - Moderately happy

    You have answered that you are neither happy nor unhappy in the previous question. We now want to follow up on that question and ask you for a more refined answer. Concerning how you are these days, taken all together, in which direction do you tend more?

    Answer possibilities:

    - Tend more toward happy
    - Tend more toward unhappy

You have answered that you are rather unhappy in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how you are these days, taken all together?

Answer possibilities:

- Very unhappy
- Moderately unhappy

11. **Substantive Question 4: Trust**

Binary: Generally speaking, would you say that people can be trusted or that you have to be careful in dealing with people?

Trinary: Generally speaking, would you say that people can be trusted or that you can't be too careful in dealing with people?

Answer possibilities binary:

- People can be trusted
- You have to be careful in dealing with people

Answer possibilities trinary:

- People can often be trusted
- People can sometimes be trusted
- You have to be careful in dealing with people

Follow-up:

You have answered that people can (trinary: often) be trusted in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how much you think people can be trusted?

Answer possibilities:

- Very much
- Moderately much

You have answered that people can sometimes be trusted in the previous question. We now want to follow up on that question and ask you for a more refined answer. Concerning how much you think people can be trusted, in which direction do you tend more?

Answer possibilities:

- Tend more toward trusting people
- Tend more toward being careful in dealing with people

You have answered that you have to be careful in dealing with people in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how careful you think you have to be in dealing with people?

Answer possibilities:

- Very careful
- Moderately careful

12. **Substantive Question 5: Political Attitude**

Binary: Would you say you are a rather liberal or a rather conservative person?

Trinary: Would you say you are a liberal, a moderate, or a conservative person?

Answer possibilities binary:

- Rather liberal
- Rather conservative

Answer possibilities trinary:

- Rather liberal
- Moderate
- Rather conservative

Follow-up:

You have answered that you are a (binary: rather) liberal person in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how liberal you are?

Answer possibilities:

- Very liberal
- Moderately liberal

You have answered that you are a moderate person in the previous question. We now want to follow up on that question and ask you for a more refined answer. In which direction do you tend more?

Answer possibilities:

- Tend more toward liberal
- Tend more toward conservative

You have answered that you are a (binary: rather) conservative person in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how conservative you are?

Answer possibilities:

- Very conservative
- Moderately conservative

13. **Substantive Question 6: Time Preference**

How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?

Answer possibilities binary:

- Rather willing
- Rather unwilling

Answer possibilities trinary:

- Rather willing
- Neither willing nor unwilling
- Rather unwilling

Follow-up:

You have answered that you are rather willing to give up something today for a future benefit in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how willing you are to give up something that is beneficial for you today in order to benefit more from that in the future?

Answer possibilities:

- Very willing
- Moderately willing

You have answered that you are neither willing nor unwilling to give up something today for a future benefit in the previous question. We now want to follow up on that question and ask you for a more refined answer. Concerning how willing you are to give up something that is beneficial for you today in order to benefit more from that in the future, in which direction do you tend more?

Answer possibilities:

- Tend more toward willing
- Tend more toward unwilling

You have answered that you are rather unwilling to give up something today for a future benefit in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how unwilling you are to give up something that is beneficial for you today in order to benefit more from that in the future?

Answer possibilities:

- Very unwilling
- Moderately unwilling

14. **Substantive Question 7: Risk Preference**

In general, how willing are you to take risks?

Answer possibilities binary:

- Rather willing
- Rather unwilling

Answer possibilities trinary:

- Rather willing
- Neither willing nor unwilling
- Rather unwilling

Follow-up:

You have answered that you are rather willing to take risks in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how willing you are to take risks?

Answer possibilities:

- Very willing
- Moderately willing

You have answered that you are neither willing nor unwilling to take risks. We now want to follow up on that question and ask you for a more refined answer. Concerning how willing you are to take risks, in which direction do you tend more?

Answer possibilities:

- Tend more toward willing
- Tend more toward unwilling

You have answered that you are rather unwilling to take risks in the previous question. We now want to follow up on that question and ask you for a more refined answer. What describes best how unwilling you are to take risks?

Answer possibilities:

- Very unwilling
- Moderately unwilling

15. **Attention Check**

What is 7 times 2?

Answer possibilities:

- 2
- 7
- 9
- 14
- 16
- 49

D Results for Restricted Sample

This appendix contains all main results of our empirical analysis when those subjects are excluded whose first three IP address blocks appear more than once. Among the remaining subjects, 74 failed the attention check, and for 84 no click and time data were recorded.

	binary survey	trinary survey
# participants	2,350	2,278
female	52.72%	53.91%
male	47.28%	46.09%
age		
< 20	0.51%	0.75%
$20 - 29$	21.32%	24.41%
$30 - 39$	36.55%	33.63%
$40 - 49$	20.77%	20.81%
$50 - 59$	12.13%	11.85%
$60 - 69$	7.15%	7.16%
≥ 70	1.57%	1.40%
highest education		
high school	24.89%	25.64%
college or higher	74.60%	74.01%
none	0.51%	0.35%
married	51.02%	50.18%
unmarried	48.98%	49.82%
kids	46.43%	46.01%
no kids	53.57%	53.99%
income		
$< \$40,000$	31.36%	31.87%
$\$40,000 - \$69,999$	36.34%	36.83%
$\geq \$70,000$	32.30%	31.30%

TABLE D1: Summary of subject demographics (restricted sample).

Roughly 90% of the subjects completed the survey within 5 minutes. The median duration was 117s and the average duration was 153s.

	binary survey	trinary survey
complete survey	112	122
demographic questions		
gender	1.44	1.44
age	1.89	1.90
education	1.96	1.99
marital status	1.45	1.45
kids	1.65	1.67
income	2.31	2.31
substantive questions		
work satisfaction	2.70	3.45
social life satisfaction	2.69	3.05
overall happiness	3.51	4.25
trust	3.62	4.72
political attitude	2.22	2.32
time preference	5.16	5.72
risk preference	2.57	2.96

TABLE D2: Median response times in seconds (restricted sample).

	Log Normalized Response Time			
	(1)	(2)	(3)	(4)
Follow-Up Response	-0.296***	-0.275***	-0.152***	-0.125***
	(0.0151)	(0.0140)	(0.0082)	(0.0084)
R-squared	0.0916	0.1225	0.1556	0.1559
Demographics & Treatment	NO	YES	YES	NO
Individual RE	NO	NO	YES	NO
Individual FE	NO	NO	NO	YES

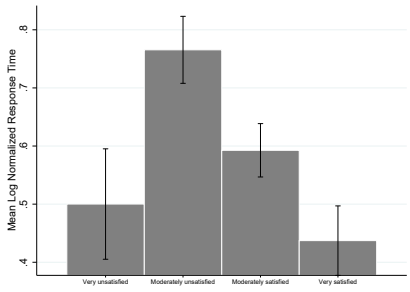
*** p<0.01, ** p<0.05, * p<0.1.

Notes: All regressions include all observations from the binary survey and the observations with non-intermediate responses from the trinary survey. The dependent variable is each subject's log response time in the initial substantive question (not including the follow-up), normalized by subtracting his/her log response time in the marital status question. Follow-Up Response is a dummy that takes the value one if the subject chose the extreme response (e.g. "very happy" or "very unhappy") in the corresponding follow-up question. All regressions include question fixed effects. The demographic controls are gender, age, education, marital status, co-residence with children, and family income. Treatment is a dummy for the survey version (binary versus trinary). Column (3) is a random-effect model with all demographic and treatment controls. Column (4) is a fixed-effect model which controls for heterogeneity at the subject level. Robust standard errors are reported in parentheses, with the ones in columns (1) and (2) being clustered at the subject level. The R-squared values reported in columns (3) and (4) concern the variation within subjects.

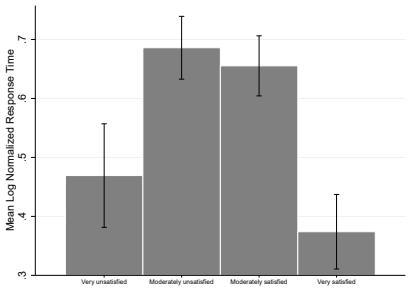
TABLE D3: Regression analysis of chronometric effects (restricted sample).

As the following two figures show, the hump-shape in average response times exists for all substantive questions in both versions of the survey, and most of the relevant pairwise differences are statistically significant at the 1% level.[21]

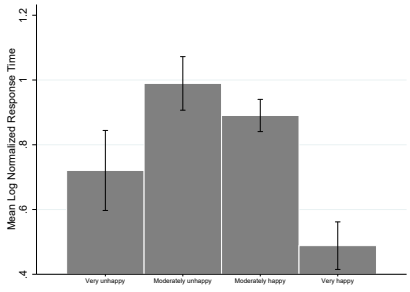
[21] Among the 28 pairwise comparisons, 23 are significant at the 1% level according to a t-test (two-sided, unequal variances), with the exceptions being in the trinary survey: the pair "very unsatisfied" and "moderately unsatisfied" in the work satisfaction question ($p = 0.3000$), the pair "very unhappy" and "moderately unhappy" in the overall happiness question ($p = 0.0225$), the pair "very careful" and "moderately careful" in the trust question ($p = 0.7079$), the pair "very conservative" and "moderately conservative" in the political attitude question ($p = 0.0233$), and the pair "very impatient" and "moderately impatient" in the time preference question ($p = 0.1512$). Using the non-normalized, raw response times, only 1 out of the 28 pairwise comparisons is significant at 1%, but 9 are at 5%, and all of them in the direction implied by the chronometric effect.



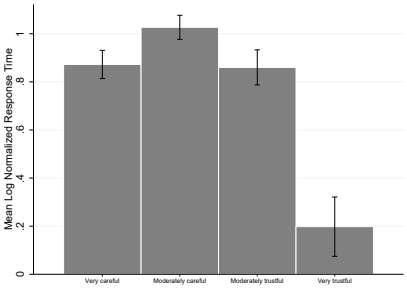
(A) Work Satisfaction



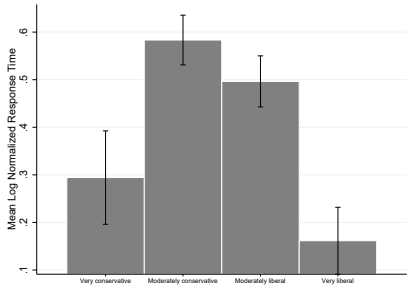
(B) Social Life Satisfaction



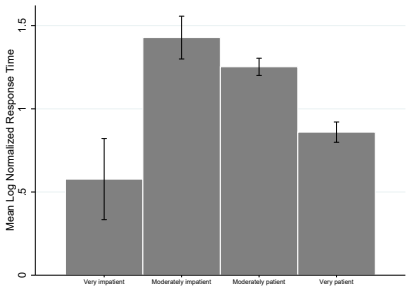
(C) Overall Happiness



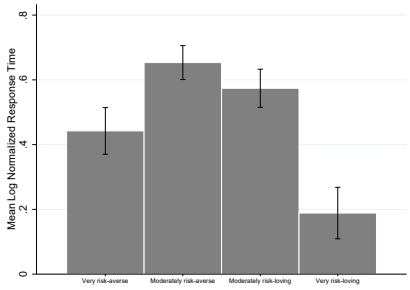
(D) Trust



(E) Political Attitude



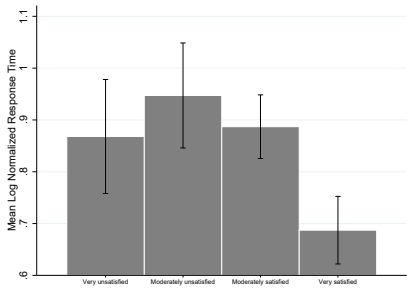
(F) Time Preference



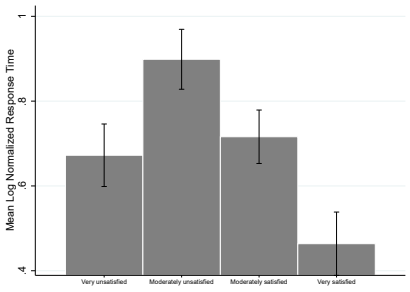
(G) Risk Preference

Notes: The figure displays, for each substantive question in the binary survey, the average log normalized response time of the subjects, categorized by their response to the initial and the follow-up question. Black lines indicate 95% confidence intervals.

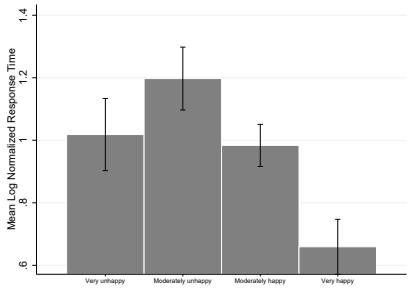
FIGURE D1: Chronometric effect by question in the binary survey (restricted sample).



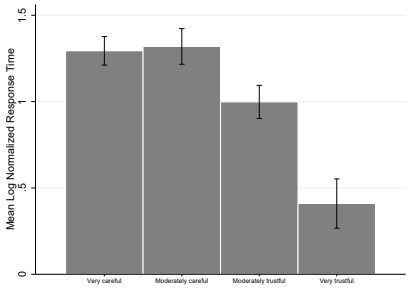
(A) Work Satisfaction



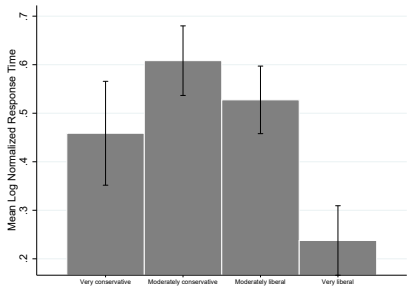
(B) Social Life Satisfaction



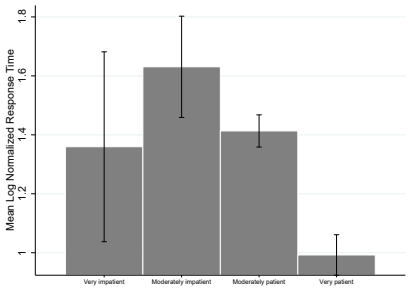
(C) Overall Happiness



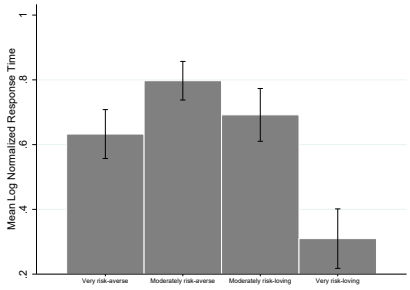
(D) Trust



(E) Political Attitude



(F) Time Preference



(G) Risk Preference

Notes: The figure displays, for each substantive question in the trinary survey, the average log normalized response time of the subjects, categorized by their (non-intermediate) response to the initial and the follow-up question. Black lines indicate 95% confidence intervals.

FIGURE D2: Chronometric effect by question in the trinary survey (restricted sample).

	work satisfac.	social satisfac.	overall happiness	trust	liberal-ism	patience	risk-taking
0: female	−0.044	0.022	−0.001	0.142***	−0.127**	0.007	0.397***
1: male	(0.0553)	(0.0533)	(0.0559)	(0.0526)	(0.0525)	(0.0682)	(0.0523)
0: young	0.197***	0.078	0.168*	0.101*	−0.285***	−0.068	−0.273***
1: middle-age	(0.0608)	(0.0582)	(0.0613)	(0.0572)	(0.0571)	(0.0739)	(0.0566)
0: middle-age	0.191*	−0.018	0.148	0.142	0.004	−0.032	−0.267***
1: old	(0.1130)	(0.1018)	(0.1126)	(0.0988)	(0.0988)	(0.1262)	(0.1000)
0: none	0.218	−0.248	−0.042	−0.187	−0.027	0.477	0.084
1: high-school	(0.3659)	(0.3782)	(0.3784)	(0.3787)	(0.3687)	(0.3993)	(0.3687)
0: high-school	0.534***	0.283***	0.347***	0.443***	0.126**	0.084	0.285***
1: college	(0.0619)	(0.0607)	(0.0627)	(0.0632)	(0.0604)	(0.0775)	(0.0601)
0: unmarried	0.688***	0.602***	0.674***	0.397***	−0.304***	0.303***	0.213***
1: married	(0.0572)	(0.0543)	(0.0578)	(0.0530)	(0.0527)	(0.0691)	(0.0519)
0: no kids	0.587***	0.540***	0.514***	0.298***	−0.153***	0.213***	0.322***
1: kids	(0.0575)	(0.0546)	(0.0577)	(0.0527)	(0.0526)	(0.0694)	(0.0522)
0: poor	0.674***	0.432***	0.552***	0.454***	−0.003	0.333***	0.211***
1: middle-income	(0.0674)	(0.0646)	(0.0673)	(0.0649)	(0.0642)	(0.0799)	(0.0632)
0: middle-income	0.009	0.055	0.077	−0.187***	−0.106*	0.200**	−0.124**
1: rich	(0.0713)	(0.0661)	(0.0716)	(0.0630)	(0.0633)	(0.0904)	(0.0628)

TABLE D4: Ordered probit analysis of the binary survey (restricted sample). Each cell corresponds to a regression of the question in the column on a dummy for membership to the group in the row. Coefficients are reported along with their robust standard errors in parentheses. Asterisks indicate statistical significance (*10%,**5%,***1%).

	work satisfac.	social satisfac.	overall happiness	trust	liberal-ism	patience	risk-taking
0: female	−0.029	0.142***	0.034	0.101**	−0.073	0.028	0.317***
1: male	(0.0490)	(0.0488)	(0.0496)	(0.0469)	(0.0473)	(0.0540)	(0.0478)
0: young	0.095*	0.017	−0.021	−0.023	−0.131**	−0.049	−0.361***
1: middle-age	(0.0531)	(0.0528)	(0.0543)	(0.0513)	(0.0517)	(0.0592)	(0.0524)
0: middle-age	0.052	−0.048	0.198**	0.289***	−0.019	−0.052	−0.044
1: old	(0.0904)	(0.0966)	(0.0962)	(0.0934)	(0.0960)	(0.1045)	(0.0939)
0: none	0.109	−0.741**	0.481	0.482	0.215	0.606*	−0.099
1: high-school	(0.2932)	(0.3280)	(0.3681)	(0.3335)	(0.3546)	(0.3275)	(0.3188)
0: high-school	0.427***	0.351***	0.369***	0.349***	0.095*	0.166***	0.217***
1: college	(0.0547)	(0.0567)	(0.0565)	(0.0552)	(0.0536)	(0.0623)	(0.0544)
0: unmarried	0.538***	0.537***	0.594***	0.348***	−0.358***	0.109**	0.200***
1: married	(0.0497)	(0.0490)	(0.0502)	(0.0470)	(0.0474)	(0.0538)	(0.0476)
0: no kids	0.452***	0.358***	0.438***	0.192***	−0.347***	0.136**	0.339***
1: kids	(0.0498)	(0.0492)	(0.0502)	(0.0468)	(0.0475)	(0.0538)	(0.0478)
0: poor	0.448***	0.297***	0.385***	0.336***	−0.132**	0.251***	0.252***
1: middle-income	(0.0589)	(0.0589)	(0.0590)	(0.0568)	(0.0572)	(0.0643)	(0.0578)
0: middle-income	0.063	0.056	0.179***	−0.061	−0.098*	0.058	−0.180***
1: rich	(0.0607)	(0.0596)	(0.0621)	(0.0565)	(0.0573)	(0.0673)	(0.0577)

TABLE D5: Ordered probit analysis of the trinary survey (restricted sample). Each cell corresponds to a regression of the question in the column on a dummy for membership to the group in the row. Coefficients are reported along with their robust standard errors in parentheses. Asterisks indicate statistical significance (*10%,**5%,***1%).

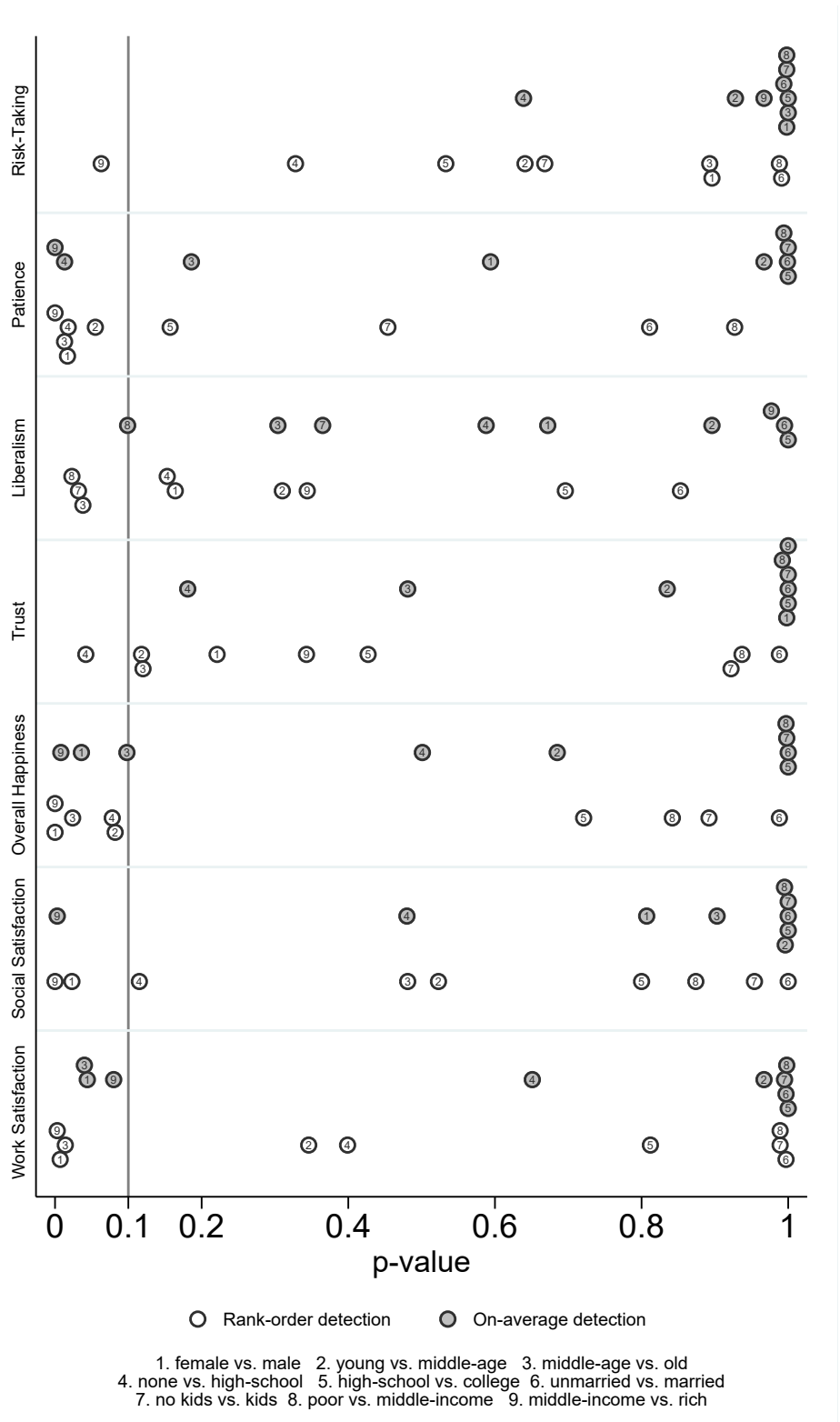


FIGURE D3: Detection analysis in the binary survey (restricted sample).

The robust sufficient condition for on-average detection in the binary survey from Proposition 7 is rejected at 10% significance level for all but five cases.[22] In the trinary survey, condition (*iii*) of Proposition 1 is never satisfied, so detection always fails in the trinary survey even with response time data.

[22] Among these five cases of robust on-average detection, four are for the risk preference question (the comparisons high-school/college, married/unmarried, kids/no kids, and poor/middle-income). The other one is for the work satisfaction question and the comparison none/high-school.