Happy Times: Identification from Ordered Response Data

Shuo Liu and Nick Netzer*

First version: December 2020
This version: June 2021

Abstract

Surveys that are designed to measure subjective states (e.g., happiness) typically generate ordinal data. A fundamental problem is that methods used to analyse ordinal data (e.g., ordered probit) rely on strong and often unjustified distributional assumptions. In this paper, we propose using survey response times to solve that problem. The key assumption of our approach is that individual response time is decreasing in the distance between the value of the latent variable and an indecision threshold. This assumption is supported by a large body of evidence on chronometric effects in psychology, neuroscience and economics. We provide conditions under which the expected value of the latent variable (e.g., average happiness) can be compared across groups, even without making distributional assumptions. We apply our method to an online survey experiment and obtain some evidence that happiness follows distributions for which traditional regression analysis is valid.

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

JEL Classification: C14, D60, D91, I31

*Liu: Guanghua School of Management, Peking University, shuo.liu@gsm.pku.edu.cn. Netzer: Department of Economics, University of Zurich, nick.netzer@econ.uzh.ch. We are grateful for very helpful comments to Carlos Alós-Ferrer, Timothy Bond, Yu Gao, Lukas Hensel, Ian Krajbich, Xi Zhi Lim, Andrew Oswald, Rainer Winkelmann, and seminar participants at Jinan University, Oxford University, Peking University, Renmin University of China, Shanghai University of Finance and Economics, Shenzhen University, Tsinghua University, Zhejiang University, the University of Zurich, and the SSES Annual Congress 2021. 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). With the help of surveys, information can be collected on a diverse range of topics, from objective socio-demographic characteristics of a population to subjective attitudes like political orientation of individuals. Surveys are also used regularly for program evaluation and in market research.

In economics, surveys have been used at least since Easterlin (1974) to measure happiness. The happiness literature has generated many 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. Recently, surveys have also 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.

Many surveys, in particular those measuring subjective states like happiness or preferences (Likert-scale surveys, Likert, 1932), generate ordinal data. For example, the overall 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 supposed to be less happy than those responding "pretty happy," but there is no information by how much less. Similarly, Falk et al. (2018)'s question about the willingness to take risks asks for an answer on a scale from 0 to 10. Experimental validation has shown that individuals responding in a lower category behave in a more risk-averse way than those responding in a higher category, but there is no information in the survey response data that allows us to judge by how much their risk-aversion coefficients differ.

To analyse ordinal survey data, researchers typically rely on ordered response models like ordered probit (see e.g. Boes and Winkelmann, 2006). These models assume that there is a continuous latent variable, e.g. true happiness or the risk-aversion coefficient, which generates discrete survey responses based on reporting thresholds. With assumptions on the distribution of the latent variable, 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, or average risk-aversion between different countries.

The distributional assumptions made in traditional models are strong. Consider the case of happiness surveys, which we will use as our main example throughout the paper. The classic ordered probit approach assumes that happiness follows Gaussian distributions with equal variances within each of the groups of individuals that the analyst wants to compare. This implicitly assumes that the happiness distributions of different groups are always ranked by first-order stochastic dominance.

Recently, Bond and Lang (2019) have shown that results depend on these assumptions in a drastic way. Without distributional assumptions, the comparison of happiness between two groups is possible only under stringent conditions which are never satisfied in real-world survey data. Roughly speaking, since we cannot learn anything from ordinal data about the distribution of the latent variable within a given response category, making suitable assumptions about that distribution allows us to conclude almost anything. More strikingly, Bond and Lang (2019) find that the distributions which are commonly employed in the literature do not have to be twisted too much to reverse the empirical findings. For instance, plausible lognormal transformations that generate happiness distributions which resemble income or wealth distributions are sufficient to overturn standard results.[1]

In this paper, we argue that the use of survey response time data could help to solve the identification 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” (we will discuss the case of more categories later). 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 farther 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 (and more accurately) 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, always 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

[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.

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 theoretical part of the paper, we show how response times can be integrated into an otherwise 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 $\tau^1 < \tau^2 < \ldots < \tau^n$, 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. They show that this 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 (more unhappy). The same conditions apply when merely asking about a ranking of average happiness between the groups, instead of first-order stochastic dominance.

Now assume that 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 it can be relaxed substantially (we will discuss this in detail later).

[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.

Then, if the distribution of response times is observed in addition to the survey responses, the conditions for identifying 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 identification. 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 those with traditional ordered response models, because the inequalities have to hold conditional on all response times.

We derive analogous conditions for the identification of a 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 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 identification condition coincides with that in Bond and Lang (2019) for these categories. Our condition remains substantially weaker for the lowest and the highest category. Also, we will argue later that the identification condition for intermediate categories can be relaxed if simple binary follow-up questions are included in the survey.

The above arguments rest on the assumption that the chronometric function is the same for all survey participants, which may not be satisfied in reality. Fortunately, this assumption can be relaxed. First, our main results continue to hold if there is individual heterogeneity in the chronometric function that is independent of happiness. The argument is similar to the observation that independent heterogeneity in the reporting thresholds is not a problem for traditional ordered response models (see e.g. Di Tella and MacCulloch, 2006, p. 29f). Second, for some results we can allow heterogeneity that correlates with happiness, as long as monotonicity of the chronometric function is preserved within the lowest and the highest response category. This may be relevant because the literature has documented that unhappy people can be slower responding, generating an additional effect of absolute happiness on response times (Studer and Winkelmann, 2014). Third, we will present a condition for identifying the ranking of group averages with arbitrary group-specific chronometric functions. Finally, most surveys ask more than one question. It is therefore possible to normalize individual response times using the response time from a baseline question. In our empirical application, we will use this procedure to account for individual fixed-effects in the speed of reading, deciding, and clicking on 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 the 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 identification problem in 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 logit 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 identification 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, we report results from an online survey that we conducted on Amazon Mechanical Turk (MTurk). We first asked several socio-demographic questions, followed by 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. Based on the responses of about 4,000 participants, we compare different socio-demographic groups and, for each substantive question, check whether it is possible to identify a difference between the groups, for example whether participants with children are happier than those without, 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.

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 fixed-effects, we normalize the raw response times using each individual's response time in the socio-demographic question about marital status, which was answered quickest on average. We then work with non-parametric kernel density estimates of the normalized response time distributions.

Not surprisingly, the conditions for identification without response time data are never satisfied; for none of the substantive questions, no group comparison, and neither the binary nor the trinary version of the survey. In the trinary survey, the condition that applies to the intermediate response categories is always violated. Since our response-time-based conditions coincide with the conventional ones for intermediate categories, we are unable to achieve identification in the trinary survey even with response time data.

In the binary version of the survey, our results vary across the different questions. On the one hand, for the overall life happiness question we obtain identification of either first-order stochastic dominance or a ranking of the averages in a majority of all group comparisons. We interpret this as first cautious evidence that happiness follows distributions for which the results of traditional ordered response models are valid. On the other hand, identification is never achieved for the question about political attitudes. This may suggest that the underlying distributions of political attitudes are much more skewed than what is postulated by traditional ordered response models. Finally, the preference and trust questions are somewhere in between, with identification being possible in about half of the cases.

Overall, our empirical findings support the idea that surveys with just two response categories may be preferable to surveys with multiple categories, and that response times, due to their continuous and cardinal nature, may be no less important than responses.

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. Some omitted proofs, extensions, and the complete questionnaires of our survey experiment are in the 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 are the same for all individuals in both groups and satisfy $\tau^1 < \tau^2 < \ldots < \tau^n$. In particular, 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 identification of first-order stochastic dominance.

Definition 1. Given data (r_A, r_B), group A is *identified to be 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 identification 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 data (r_A, r_B), group A is *identified to be 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 $i = 0, \dots, n$ and $j = A, B$.

In words, on-average identification 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 $q : \mathbb{R} \to \mathbb{R}$. Hence, the definition of rank-order identification could be rephrased accordingly. As we show in Appendix A.1, the definition of on-average identification 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 identification 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 (provided the expectations are well-defined).

We now state a first result about rank-order identification. This result is not new (see e.g. 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 identified to be 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* $k = 1, \dots, n-1$.

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

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

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

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 $r_B^n > 0$. □

Observe that 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 identified to be 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 weakening the notion of identification. Unfortunately, as we show in the next result (which is new to the best of our knowledge), the on-average notion of identification is not more admissible than the rank-order notion.

Proposition 2. *Given (r_A, r_B), group A is identified to be on-average-happier than group B if and only if group A is identified to be 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 identified to be rank-order-happier than group B, so at least one of conditions $(i) - (iii)$ in Proposition 1 is violated. Suppose first that there exists $k^* = 1, \ldots, n-1$ for which $\sum_{i=0}^{k^*} r_A^i > \sum_{i=0}^{k^*-1} r_B^i$. Define $\tau^{k^*} = 0$ and $\tau^{k^*+1} = 1$. Let the other thresholds be ϵ-spaced for some $\epsilon \in (0, 1)$, i.e., $\tau^{k^*+2} = 1 + \epsilon$, $\tau^{k^*+3} = 1 + 2\epsilon$, $\ldots$, and $\tau^{k^*-1} = -\epsilon$, $\tau^{k^*-2} = -2\epsilon$, $\ldots$. Construct G_j by setting $G_j(\tau^k) = \sum_{i=0}^{k-1} r_j^i$ for each

$k = 1, \dots, n$, and $G_j(\tau^n + \epsilon) = 1$ and $G_j(\tau^1 - \epsilon) = 0$. Then define $G_A(\epsilon) = G_A(1)$ and $G_B(1-\epsilon) = G_B(0)$, and complete the construction of each G_j by connecting the defined points in a piece-wise linear way. It follows that (G_A, G_B, τ) satisfies (1) for $i = 0, \dots, n$ and $j = A, B$, for all values of $\epsilon \in (0,1)$. It also follows that

$$\lim_{\epsilon \to 0} \mathbb{E}_{G_A}[\tilde{h}] = 1 - G_A(1) = 1 - \sum_{i=0}^{k^*} r_A^i, \text{ and } \lim_{\epsilon \to 0} \mathbb{E}_{G_B}[\tilde{h}] = 1 - G_B(0) = 1 - \sum_{i=0}^{k^*-1} r_B^i.$$

Hence for sufficiently small $\epsilon > 0$ we obtain $\mathbb{E}_{G_A}[\tilde{h}] < \mathbb{E}_{G_B}[\tilde{h}]$. The cases where $r_A^0 > 0$ or $r_B^n > 0$ are again immediate, because expected values can be decreased or increased arbitrarily. □

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 A than in group B. 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 is 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 $\bar{t}$, respectively (where we allow for $\underline{t} = 0$ and $\bar{t} = +\infty$). Response times are related to the latent variable $\tilde{h}$ through a chronometric function $c : \mathbb{R}_{++} \to [\underline{t}, \bar{t})$. This function is assumed to be continuous, strictly decreasing in δ whenever $c(\delta) > \underline{t}$, and to satisfy $\lim_{\delta \to 0} c(\delta) = \bar{t}$ 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. The assumption can be relaxed, and we will discuss this in dedicated remarks after each of our following results.

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).

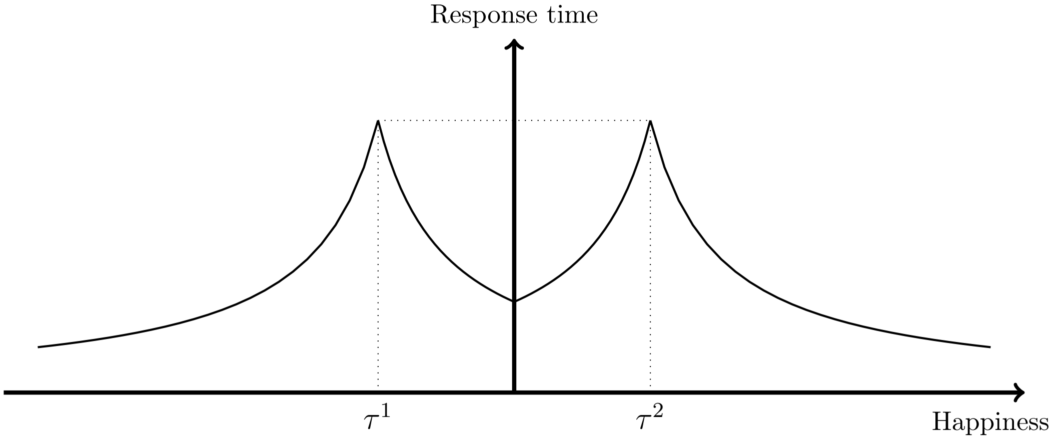


Figure 1: Example of response times with $n = 2$, $\tau^1 = -2, \tau^2 = 2$, and $c(\delta) = 1/(\delta + 1)$.

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

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

Note that 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 $F_j^i(\bar{t}) = 1$,[3] the analyst can again ask the previous identification questions. At the risk of creating redundancy, we state these again as formal definitions.

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

Definition 3. Given data (r_A, r_B, F_A, F_B), group A is *identified to be 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 data (r_A, r_B, F_A, F_B), group A is *identified to be 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 complete characterization of rank-order identification 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 identified to be 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

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

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 $i = 0, \dots, n$, $j = A, B$, 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 $k^* = 1, \dots, n-1$ 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 $i = 0, \dots, n$, $j = A, B$, 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 lower 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 that condition (*iii*) can be weakened if we include simple binary follow-up questions in the survey.

Remark. Our specific formulation of the chronometric function in intermediate response categories is not essential for the conclusion that response times are uninformative in these categories. Proposition 3 holds unaltered 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. Importantly, conditions (*i*) and (*ii*) also do not make comparisons across response categories. Hence Proposition 3 would hold for arbitrary category-specific chronometric functions $c^i(.)$. This is important, because the literature has argued that absolute happiness levels could 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.

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 identified to be rank-order-happier than group B if and only if $r^0_A F^0_A(t) - r^0_B F^0_B(t) \leq 0 \leq r^1_A F^1_A(t) - r^1_B F^1_B(t)$ for all $t \in (\underline{t}, \bar{t})$.*

In fact, if one is interested in rank-order identification, 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 satisfies that $r^0_A F^0_A(t) - r^0_B F^0_B(t) \leq 0 \leq r^1_A F^1_A(t) - r^1_B F^1_B(t)$ 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

$$r^0_j F^0_j(t) = G_j(\tau^1 - c^{-1}(t)) \text{ and } r^1_j F^1_j(t) = 1 - G_j(\tau^1 + c^{-1}(t)),$$

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

$$r^0_A F^0_A(t) - r^0_B F^0_B(t) = G_A(\tau^1 - c^{-1}(t)) - G_B(\tau^1 - c^{-1}(t)) \leq 0$$

and

$$r^1_A F^1_A(t) - r^1_B F^1_B(t) = G_B(\tau^1 + c^{-1}(t)) - G_A(\tau^1 + c^{-1}(t)) \geq 0,$$

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.[4]

[4] When the conditions for rank-order identification are violated, the data from binary surveys allows us to pin down the percentiles of the distributions for which the dominance relation does not hold. For instance,

Remark. We can extend the model and assume that response times are stochastic, modelled by adding i.i.d. zero-mean noise to the log response times that are generated by the deterministic process considered so far. An alternative interpretation is that individuals differ in their chronometric functions, and this heterogeneity is uncorrelated with happiness (but recall our discussion of the possibility of category-specific chronometric functions). Formally, suppose the response time for an individual with happiness h is $c(|h-\tau^1|)\cdot\tilde{\eta}$, where $\tilde{\eta}$ is a non-negative random variable with mean one, assumed to be i.i.d. according to a continuous distribution function Z. The multiplicative structure is equivalent to adding noise to log response times and ensures that response times remain non-negative. Now, the fraction of individuals who respond in each of the two categories at time $t \in (\underline{t}, \overline{t})$ or earlier is

$$F_j^0(t) = \frac{\int_0^{+\infty} G_j(\tau^1 - c^{-1}(t/\eta))\, dZ(\eta)}{G_j(\tau^1)}, \quad F_j^1(t) = \frac{1 - \int_0^{+\infty} G_j(\tau^1 + c^{-1}(t/\eta))\, dZ(\eta)}{1 - G_j(\tau^1)},$$

which replaces equation (2). It is easy to adapt the proof of Proposition 4 to this specification. Hence, if the true happiness distributions of the groups can be ranked by first-order stochastic dominance, the data generated by the model with noisy or heterogeneous chronometric functions still satisfies the condition in Corollary 2, so that our techniques will continue to detect the dominance relation.

The next proposition, which is the second main result of our paper, gives a weaker sufficient condition for on-average identification, which is implied by but does not imply the previous condition in Proposition 3. This shows shows that the on-average notion of identification is indeed weaker than the rank-order notion when response times are being used.[5]

Proposition 5. *Given (r_A, r_B, F_A, F_B), group A is identified to be 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* $t \in (\underline{t}, \overline{t})$*, and*

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

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.

[5] 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 identification are violated.

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

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

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 identification and on-average identification 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 identified to be 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 $t \in (\underline{t}, \overline{t})$.*

For $t \to \overline{t}$, 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 identification, fast "most unhappy" responses in group A relative to group B can be compensated by even faster "most happy" responses. Note that this argument involves a comparison of response times across response categories, hence category-specific chronometric functions are not admissible here.

Unlike for rank-order identification, if one is interested in on-average identification, it can be useful to consider surveys with more than two response categories. The following example illustrates this point.

Example. 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. For the happiness of group B, the distribution 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

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

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 identification. 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 show in Appendix A.2 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 identify that group A is on-average happier than group B.

However, if one is willing to make the assumption that the happiness distribution of each group is symmetric around its mean (which was not the case for group A in the example above), and that in a binary survey the reporting threshold lies between the two groups' average happiness levels, then the generated data will necessarily 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 satisfies 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 $t \in (\underline{t}, \bar{t})$,

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

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 identification 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.[6]

Remark. Like for Proposition 4, the result in Proposition 6 continues to hold with i.i.d. noise in response times or independent heterogeneity in the chronometric function. Our condition in Corollary 3 continues to detect the correct ranking of happiness averages and to serve as a partial test of symmetry.

Finally, we provide an additional sufficient condition for on-average identification in the binary case, which is stronger than in Corollary 3 but is neither implied by nor implies the conditions for rank-order identification in Corollary 2.

Proposition 7. *Given (r_A, r_B, F_A, F_B) for $n = 1$, group A is identified to be 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 $t \in (\underline{t}, \bar{t})$.*

Proof. 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 under the stated condition, for every (G_A, G_B, τ, c) that generates the data. □

For $t \to \bar{t}$, this condition implies $r_A^0 \leq r_A^1$ and $r_B^0 \geq r_B^1$. More generally, it requires that in group A there are more responses in the high than in the low category when considering only those responses that took place before any time t, and conversely for group B.

Remark. The condition in Proposition 7 has the advantage that it does not make direct comparisons of the response time distributions across the two groups. As argued in the proof, if the condition is satisfied, the average happiness of group A must be larger than the reporting threshold and the average happiness of group B must be smaller. A ranking of the averages then obtains from the conventional assumption that both groups use the same reporting threshold. It follows that the condition would continue to achieve identification even if we allowed for arbitrary group-specific chronometric functions $c_j(\cdot)$.

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

3 Empirical Application

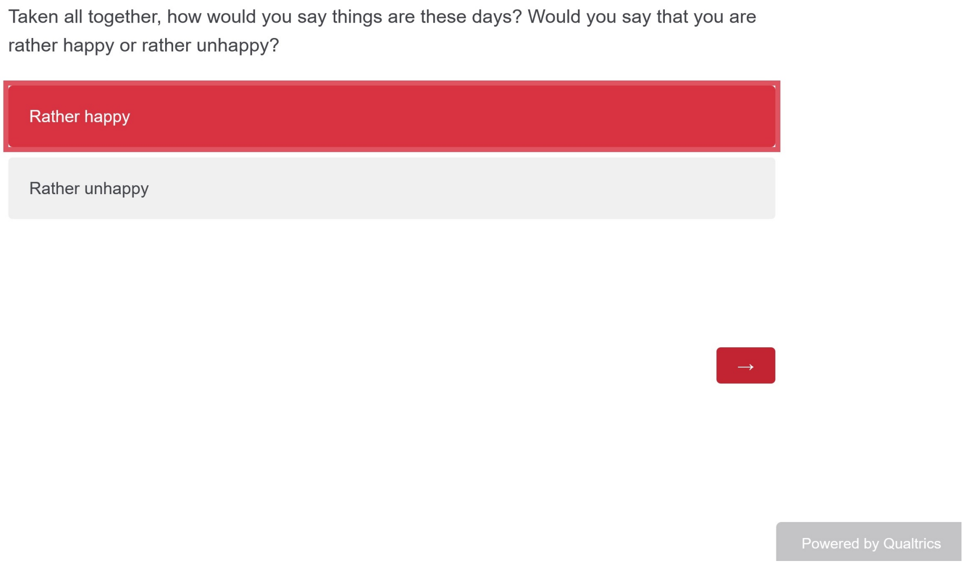
3.1 Survey Description

We now provide an empirical application of our method, to show how it can be implemented in practice and gives rise to new insights. To this end, we designed and conducted a survey experiment on the online platform MTurk. MTurk 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 October 2019 through the ETHZ Decision Science Laboratory. The survey consisted of two parts. After the participants gave their informed consent, the first part included 5 standard socio-demographic questions concerning gender, age, education, marital status, and the number of children currently living in the household. In the second part, the survey participants 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 adapted from the General Social Survey, which is the primary source for U.S. evidence on a broad set of social science issues (Davis and Smith, 1991). 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 participants. In one version, the possible response to each substantive question was binary (e.g., "rather happy" and "rather unhappy" for the overall happiness question). In the other version, there were three response categories (e.g., "rather happy", "neither happy nor unhappy", and "rather unhappy"). The two versions of the survey allow us to investigate whether intermediate response categories are helpful or harmful for the goal of identifying group differences. The full version of both questionnaires can be found in Appendix C.

Figure 2 provides an example of the survey screen displayed to the participants. Before choosing the submission button "→" at the right bottom of the screen and moving on to the next page, the participants first had to select one of the available responses to the question (there was no default answer). The participants 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 initial display of the question and a participant's last click before submission. This "time to



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

Rather happy

Rather unhappy

Figure 2: Example of survey screen.

final response" captures most closely the duration of a participant's decision process, which may involve changing an initial response by clicking on a different button. According to our theoretical analysis, the duration of the decision process is the key ingredient for eliciting cardinal information from the ordinal responses of the survey participants.

3.2 Summary Statistics

We recruited 4,000 participants from the U.S. with an MTurk approval rate of at least 98%. In the initial sample, 34 subjects failed an attention check at the end of the survey ("What is 7 times 2?"). No click and time data were recorded for 94 subjects, presumably because they used keyboard navigation to answer the questions. Finally, one subject asked to have his/her answer removed because he/she did not like the question about political attitude. All these subjects were dropped, so our final sample contains 1,939 subjects in the binary survey and 1,932 subjects in the trinary survey. Roughly 85% of the participants completed the survey within 2 minutes (the median duration was 70s, the average duration was 110s). The participants received a compensation of 50 cents.[7]

Table 1 shows that the demographics of the participants in the two survey versions are very similar, as should be expected given the random assignment. Overall, our sample is in line with the U.S. population regarding gender, marital status, and children in the household. However, the sample is somewhat younger and more educated than the general U.S.

[7] The compensation for the first 100 subjects was 1 dollar. We reduced the payment afterwards because it was very high given the short duration of the survey.

	binary survey	trinary survey
# subjects	1939	1932
female	53.74%	51.55%
male	46.26%	48.45%
age		
< 20	0.88%	0.78%
$20 - 29$	27.08%	25.10%
$30 - 39$	35.59%	36.23%
$40 - 49$	17.48%	18.37%
$50 - 59$	11.60%	12.27%
$60 - 69$	6.45%	6.11%
≥ 70	0.93%	1.14%
highest education		
high school	28.47%	27.33%
college or higher	71.27%	72.41%
none	0.26%	0.26%
married	46.26%	45.13%
unmarried	53.74%	54.87%
kids	41.05%	39.23%
no kids	58.95%	60.77%

Table 1: Summary of subject demographics.

population, which is consistent with previous studies on the representativeness of MTurkers (e.g., Paolacci et al., 2010; Huff and Tingley, 2015).

Table 2 summarizes the median response times for each question. The socio-demographic questions and their possible response categories were the same in the two survey versions, and hence the median response times are also approximately the same. The quickest median responses are for the gender question and the marital status question, reflecting that these questions are short and for many participants easy to answer. The average response time is smaller for the marital status question (2.18s) than for the gender question (4.87s). Therefore, we will use the response time in the marital status question for individual normalization in our later analysis, i.e., we will subtract (in logs) this response time from the response times in each of the substantive questions. This way we can account for individual differences in the speed of reading or decision-making more generally. Note finally that 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 answer categories that have to be

	binary survey	trinary survey
complete survey	67	73
demographic questions		
gender	1.36	1.40
age	1.79	1.81
education	1.90	1.90
marital status	1.40	1.41
kids	1.60	1.61
substantive questions		
work happiness	2.55	3.18
social happiness	1.91	2.39
overall happiness	3.29	4.29
trust	3.54	4.87
conservatism	1.97	2.16
patience	4.84	5.48
risk-aversion	2.38	2.72

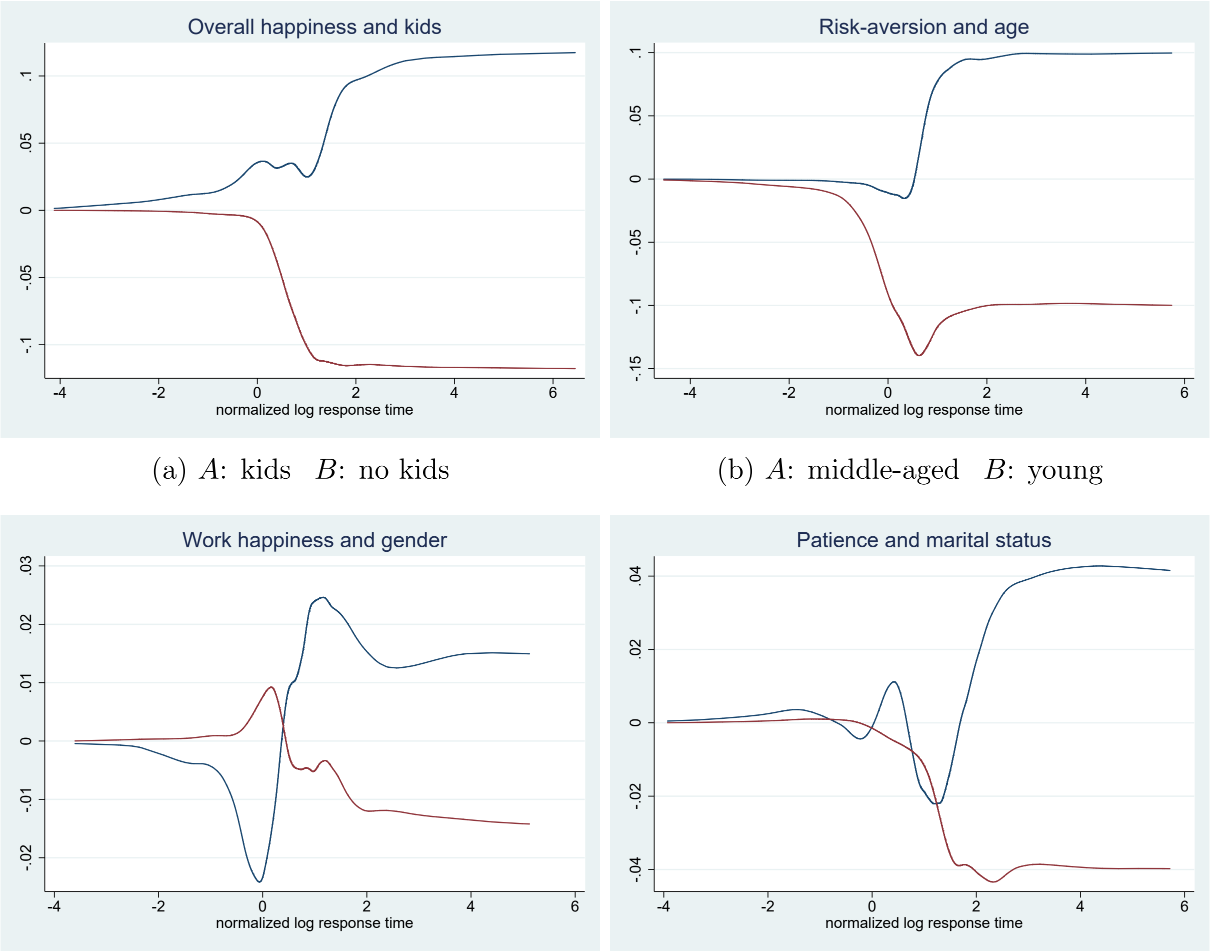
Table 2: Median response times, in seconds.

read, understood, and considered by the participants.

3.3 Analysis of Binary Survey

We start by applying our techniques to the binary survey. We divide the sample into socio-demographic groups and, for each of the substantive questions, make pairwise comparisons between the groups to check whether the conditions in Corollaries 2 and 3 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 on identification rather than a causal interpretation of the results, we prefer to keep the number of pairwise comparisons low.

The key step in the application is to construct the empirical counterparts of $r^i_j F^i_j(\cdot)$ for each demographic group j and each response category i, given the substantive survey question being considered. For r^i_j, we simply use the empirical frequency of responses. For $F^i_j(\cdot)$, we first normalize each participant's log response time by subtracting his/her log response time in the marital status question. As explained before, responses were quickest in the marital status question, and hence we use them as a proxy for general response speed, allowing us to



(a) A: kids B: no kids

(b) A: middle-aged B: young

(c) A: female B: male

(d) A: married B: unmarried

Figure 3: $r_A^0 F_A^0(t) - r_B^0 F_B^0(t)$ (red) and $r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ (blue).

control for individual heterogeneity. We then replace each cumulative distribution function $F_j^i(\cdot)$ by a non-parametric kernel density estimate (Epanechnikov kernel, optimal adaptive bandwidth) of the distribution of normalized log response times. Estimating the distribution of log response times ensures that the non-negativity constraint $t \geq 0$ is never violated.

Consistent with Bond and Lang (2019), the conditions of Proposition 1 are never satisfied by our survey data. In particular, the number of participants responding in each category is non-negligible for all demographic groups and all substantive questions. Our results thus confirm that little can be learned within the traditional ordered response framework without making distributional assumptions.

Fortunately, our conditions based on survey response times do have bite. To illustrate, consider the survey question about overall happiness ("Taken all together, how would you say things are these days? Would you say that you are rather happy or rather unhappy?").

Compare the groups of participants who have children living in the household (group A) and who have not (group B). The blue curve in Figure 3(a) plots $r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$, with t varying on the x-axis. The blue curve always lies above zero, meaning that the fraction of participants with kids who responded to be happy is always larger than the fraction of participants without kids who responded to be happy, even when restricting the sample to responses which took place before any time t. Similarly, the red curve in Figure 3(a) plots $r_A^0 F_A^0(t) - r_B^0 F_B^0(t)$. It always lies below zero, so the fraction of participants with kids who responded to be unhappy is always smaller than the fraction of participants without kids who responded to be unhappy, again for all response times. Taken together, participants who have children are identified to be rank-order happier than those who have not, by Corollary 2. The timing of responses rules out that the latent happiness variable follows distributions for which the findings of traditional ordered response models would be reversed.

As a second example, consider the relationship between risk-attitudes and age. The survey question was “In general, how willing are you to take risks?” and the possible answers were “rather willing” and “rather unwilling.” We think of the latent variable for this question as being a coefficient of risk-aversion, such that higher values capture higher risk-aversion and tend to generate responses in the “rather unwilling” category. Compare the groups of middle-aged participants (group A, age from 40 to 59) and young participants (group B, age below 40). Figure 3(b) shows that there is no rank-order identification of differences in risk-aversion between these two groups. The blue curve falls below zero for some intermediate values of t, meaning that there are some systematically faster “rather unwilling” responses in the young group than in the middle-aged group. Hence, while the overall fraction of participants who responded to be unwilling to take risks is higher in the middle-aged group, we cannot rule out that some young participants are so strongly risk-averse as to reverse the traditional assumption of first-order stochastic dominance in the distributions of risk-aversion coefficients. Nevertheless, the fast “rather unwilling” responses in the young group are offset by even faster and more frequent “rather willing” responses, reflected in the fact that the red curve always lies below the blue curve in Figure 3(b). Thus, middle-aged participants are still identified to be on-average more risk-averse than young participants, by Corollary 3.

Figure 3(c) illustrates that sometimes even on-average identification fails. While the overall fraction of participants who report to be “rather satisfied” with their work is higher for females than for males, the quickest responses in this category are predominantly given by males (blue curve), reflecting a particularly large happiness with their work situation among many males, based on the chronometric relationship. Conversely, the quickest responses in the “rather unsatisfied” category are predominantly given by females (red curve), reflecting that some females are particularly unhappy about their work. Hence, we cannot identify from the binary survey data whether it is true that women are on-average happier with their work than men. The results that traditional models like ordered probit generate thus depend on distributional assumptions which we cannot confirm in the data. Similarly, since the two curves in Figure 3(d) also cross, we cannot compare time preferences (as captured by our survey question on the willingness to postpone rewards) between married and unmarried participants, without relying on assumptions on the distribution of the latent patience variable in the population.

Table 3 summarizes the outcomes of our identification check for each substantive survey question and each socio-demographic characteristic, contrasting them with the results of traditional ordered probit estimation. In each cell of the table, the number is the ordered probit coefficient obtained by regressing the dependent variable of the column on a dummy for the demographic group A of the row (and a constant). Significance levels are indicated by asterisks (*10%,**5%,***1%). The color of each cell indicates whether identification is achieved with our response time method (white: none, light-gray: on-average identification, dark-grey: rank-order identification). For instance, consistent with Figure 3, the table shows that participants with kids are rank-order happier than those without, young participants are on-average less risk-averse than the middle-aged, and no identification is achieved when comparing the work happiness between the genders, or the patience between the married and the unmarried.[8]

Since identification and estimation are distinct problems,[9] it is not surprising that there seems to be no strong connection between the ability of identification and the significance of the estimated ordered probit coefficient. For instance, we sometimes achieve on-average identification but not significance (e.g. for the relation between gender and happiness about social life), and sometimes significance but not identification (e.g. for the relation between education and trust). Notice, however, that in all cases where rank-order identification is achieved, the ordered probit parameter is highly significant.

For the overall life happiness question, our response time method broadly supports traditional assumptions (rank-order or on-average identification in 4 out of 6 cases). Things look slightly worse for the two more specific happiness questions, the two preference questions, and the trust question. The politics question ("Would you say you are a rather liberal or a rather conservative person?") performs worst, with no identification achieved at all.

[8] If systematic inter-group heterogeneity in chronometric effects is a concern, even after our individual normalization, we can invoke Proposition 7. The empirical implementation is straightforward. In our survey, however, the condition is never satisfied.

[9] The former asks whether a certain property of the data-generating process can be unambiguously learned from the data, the latter asks which data-generating process can best fit the data.

	work happiness	social happiness	overall happiness	trust	conserv-atism	patience	risk-aversion
A: female B: male	0.043	-0.047	-0.047	-0.124^{**}	-0.090	-0.307^{***}	0.480^{***}
A: young B: middle-age	-0.260^{***}	-0.004	-0.098	-0.200^{***}	-0.370^{***}	0.038	-0.252^{***}
A: middle-age B: old	0.008	-0.104	0.102	-0.083	0.003	0.086	-0.174
A: highschool B: college	-0.372^{***}	-0.254^{***}	-0.256^{***}	-0.216^{***}	0.045	-0.292^{***}	0.129^{**}
A: married B: unmarried	0.474^{***}	0.481^{***}	0.536^{***}	0.192^{***}	0.413^{***}	0.215^{***}	0.112^{**}
A: kids B: no kids	0.389^{***}	0.204^{***}	0.368^{***}	0.026	0.306^{***}	0.103	0.034

Table 3: Binary survey. Numbers are ordered probit coefficients, and asterisks indicate significance (*10%,**5%,***1%). Colors indicate identification based on response times (white: none, light-gray: on-average identification, dark-grey: rank-order identification).

3.4 Analysis of Trinary Survey

We now turn to the analysis of the survey version with three response categories. Not too surprisingly, the conditions of Proposition 1 are never satisfied, consistent with the findings of Bond and Lang (2019). In all cases, condition (*iii*) from Proposition 1 is not satisfied, showing that the problems highlighted by Bond and Lang (2019) are not only due to extreme assumptions on the distribution of happiness in the lowest or highest response category.

Our general approach is the same as for the binary survey. The results are reported in Table 4. The numbers in each cell of the table are again the estimated ordered probit coefficients. Comparing the estimation results of the binary and the trinary survey versions, we sometimes obtain different parameter signs (8 out of 42 times), but then at least one of the two parameters is always insignificant.[10] Overall, the two versions of the survey seem to generate comparable results based on ordered probit estimation.

The picture is different for identification based on response times. All cells of Table 4 are white, meaning that there is no case where we can learn a ranking of the underlying latent distributions from the trinary survey and response time data. Observe that this can happen

[10] Each of the two survey versions is sometimes "more significant" than the other. The strongest difference in significance is for the middle-age versus old comparison, where the trinary survey is often significant while the binary survey is not.

	work happiness	social happiness	overall happiness	trust	conserv-atism	patience	risk-aversion
A: female B: male	0.029	0.048	0.133**	−0.112**	−0.050	−0.124**	0.308***
A: young B: middle-age	−0.039	−0.020	−0.096	0.067	−0.228***	0.039	−0.225***
A: middle-age B: old	−0.269**	−0.292***	−0.346***	−0.245**	−0.236**	−0.082	−0.136
A: highschool B: college	−0.298***	−0.255***	−0.281***	−0.365***	0.100*	−0.216***	0.186***
A: married B: unmarried	0.387***	0.367***	0.497***	0.293***	0.374***	0.082	0.108**
A: kids B: no kids	0.255***	0.158***	0.359***	0.134**	0.254***	0.110*	−0.079

Table 4: Trinary survey. Numbers are ordered probit coefficients, and asterisks indicate significance (*10%,**5%,***1%). Colors indicate identification based on response times (white: none, light-gray: on-average identification, dark-grey: rank-order identification).

even if the true distributions satisfy FOSD (Proposition 4 holds for binary but not for trinary surveys), and hence it is not a contradiction when we obtain rank-order identification in the binary but not in the trinary survey.

It is worthwhile emphasizing again that identification fails because condition (*iii*) from Propositions 3 or 5 is never satisfied, so there is no need to even construct the empirical response time distributions. The stringency of this 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.

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 usually 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. (2020) 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 standard econometric techniques used to analyse ordinal data suffer from a fundamental identification problem. Bond and Lang (2013) and Schroeder and Yitzhaki (2017) make the related point that ordinal data cannot simply be treated as cardinal, and they conclude that results from test score and subjective well-being research, respectively, can be highly 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 information 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, 2020), 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 ordered response data are ordinal, while comparing averages across groups requires cardinal information, traditional ordered response models must rely on assumptions about the distribution of a latent variable. By contrast, response time distributions reveal cardinal information about the distribution of the latent variable, based on the chronometric effect, and can therefore be used to replace unjustified assumptions.

We have repeatedly advocated the use of binary surveys, combined with a measurement of response times. Our empirical analysis confirms that identification is possible that way, while surveys with multiple categories fail in achieving identification, even with response time data, due to the non-monotonicity of response times in intermediate response categories. Existing surveys usually have more than two answer categories, and applied researchers may therefore prefer working with multiple categories. We formally show in Appendix B that identification can be improved in such surveys with simple binary follow-up questions. For instance, when an individual initially responds in the intermediate category "pretty happy," the survey may in a second step ask which adjacent category the individual feels closer to, "not too happy" or "very happy." The condition for rank-order identification using response time data then becomes weaker than condition (*iii*) in Propositions 3 or 5.

We have in mind two different ways in which our results can be used. First, surveys are increasingly conducted online, and recording response times is easy and costless in that case. Hence 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. Of course, causal analysis will be an important concern in many applications, which implies that the groups to be compared will typically have to be much finer than in our simple empirical illustration. We leave to future research the question how to integrate response time data into a full-fledged multivariate regression analysis. Second, one could use our novel techniques in auxiliary studies, with the goal of verifying in a representative sample that the latent variable of interest typically follows distributions for which traditional ordered response models are appropriate. Our empirical illustration goes in that direction, showing that happiness often seems to be distributed in a way for which the ordered probit results are correctly identified. 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.

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

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.

Di Tella, R. and MacCulloch, R. (2006). Some uses of happiness data in economics. *Journal of Economic Perspectives*, 20(1):25–46.

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. (2020). Using response times to infer others' beliefs: An application to information cascades. Mimeo.

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

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.

Huff, C. and Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of mturk survey respondents. *Research & Politics*, 2(3):1–12.

Kaiser, C. and Vendrik, M. C. (2020). How threatening are transformations of reported happiness to subjective wellbeing research? IZA Disucssion 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.

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.

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.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419.

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.

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. (2020). What do lost wallets tell us about survey measures of social capital? CESifo Working Paper No. 8418.

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

A Omitted Proofs

A.1 On-Average Identification with Ordered Responses

On-average identification 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 with $n = 2$

Consider the survey with $n = 2$, reporting thresholds $\tau^1 = -0.5$ and $\tau^2 = 0.2$, and the chronometric function $c(\delta) = 1/\delta$. 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 $1 \leq \tau^2 + c^{-1}(t)$.

B Follow-Up Questions

In the main text, we pointed out that response times in intermediate response categories are not helpful for identification, as they are not monotone in the latent variable. This appendix shows that responses in intermediate categories become more informative if we include simply binary follow-up questions in the survey.

Specifically, suppose that whenever an individual responds in an intermediate category $i = 1, \ldots, n-1$, she will be asked a binary follow-up question which requires her to indicate whether she feels closer to category $i-1$ or to category $i+1$. Let $\underline{r}_j^i$ be the fraction of individuals within group j who first respond in category i in the initial question and then report to be closer to category $i-1$ in the follow-up question. The response time distribution of these individuals in the follow-up question is denoted by $\underline{F}_j^i$. Similarly, we use $\overline{r}_j^i$ to denote the fraction of individuals within group j who respond to be closer to $i+1$ in the follow-up question, and we let $\overline{F}_j^i$ be the corresponding distribution of response times. Our extended dataset now contains $r_j = (r_j^0, (r_j^i, \underline{r}_j^i, \overline{r}_j^i)_i, r_j^n)$ and $F_j = (F_j^0, (F_j^i, \underline{F}_j^i, \overline{F}_j^i)_i, F_j^n)$.

We assume that the follow-up questions generate data in a way that is similar to the initial questions, so the data-generating process is augmented with reporting thresholds $\check{\tau}^i \in (\tau^i, \tau^{i+1})$ and chronometric functions $\check{c}^i : \mathbb{R}_{++} \to (\underline{t}, \bar{t})$ for the follow-up questions of each intermediate category i. It is straightforward to extend the definition of rank-order identification accordingly. In what follows, we provide a characterization of rank-order identification using response times and binary follow-up questions.

Proposition 8. *Consider a survey with binary follow-up questions. Given (r_A, r_B, F_A, F_B), group A is identified to be 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, for all $k = 1, ..., n-1$,

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

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

Proof. Let $(G_A, G_B, \tau, c, (\check{\tau}^i, \check{c}^i)_i)$ be any data-generating 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 $t \in (\underline{t}, \bar{t})$,

$$\begin{aligned}
&\sum_{i=0}^{k-1} r_j^i + r_j^k \underline{r}_j^k \underline{F}_j^k(t) \\
&= G_j(\tau^k) + \big(G_j(\tau^{k+1}) - G_j(\tau^k)\big) \cdot \frac{G_j\left(\check{\tau}^k - (\check{c}^k)^{-1}(t)\right) - G_j(\tau^k)}{G_j(\tau^{k+1}) - G_j(\tau^k)} \\
&= G_j\left(\check{\tau}^k - (\check{c}^k)^{-1}(t)\right),
\end{aligned}$$

and, similarly,

$$\begin{aligned}
&\sum_{i=k+1}^{n} r_j^i + r_j^k \overline{r}_j^k \overline{F}_j^k(t) \\
&= 1 - G(\tau^{k+1}) + \big(G_j(\tau^{k+1}) - G_j(\tau^k)\big) \cdot \frac{G_j(\tau^{k+1}) - G_j(\check{\tau}^k + (\check{c}^k)^{-1}(t))}{G_j(\tau^{k+1}) - G_j(\tau^k)} \\
&= 1 - G_j\left(\check{\tau}^k + (\check{c}^k)^{-1}(t)\right).
\end{aligned}$$

Hence, conditions (*iii*) and (*iv*) are equivalent to

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

and

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

for all $k = 1, ..., n$ and $t \in (\underline{t}, \bar{t})$. It then follows as in the proof of Proposition 3 that (*iii*) and (*iv*) are necessary and sufficient for obtaining that $G_A(h) \leq G_B(h)$ for all $h \in (\tau^1, \tau^n]$. □

The introduction of follow-up questions does not affect the extreme categories, so conditions (*i*) and (*ii*) remain the same as in Proposition 3. The novel parts in Proposition 8 are conditions (*iii*) and (*iv*). It is an easy exercise to show that they 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.

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. **Substantive Question 1: Work Happiness**

   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

8. **Substantive Question 2: Social Happiness**

   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

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

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

   Answer possibilities binary:

   - Rather happy
   - Rather unhappy

   Answer possibilities trinary:

   - Rather happy
   - Neither happy nor unhappy
   - Rather unhappy

10. **Substantive Question 4: Trust**

    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:

    - Most people can be trusted
    - Need to be very careful

Answer possibilities trinary:

- People can almost always be trusted
- People can sometimes be trusted
- You can't be too careful in dealing with people

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

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

    Answer possibilities binary:

    - Rather liberal
    - Rather conservative

    Answer possibilities trinary:

    - Rather liberal
    - Moderate
    - Rather conservative

12. **Substantive Question 6: Time Preferences**

    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

13. **Substantive Question 7: Risk Preferences**

    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

14. **Attention Check**

    What is 7 times 2?

    Answer possibilities:

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