# Happy times: measuring happiness using response times

Liu, Shuo ; Netzer, Nick

DOI: https://doi.org/10.1257/aer.20211051

*American Economic Review 2023, 113(12): 3289–3322*
*https://doi.org/10.1257/aer.20211051*

# Happy Times: Measuring Happiness Using Response Times[†]

*By* Shuo Liu and Nick Netzer*

*Surveys measuring happiness or preferences generate discrete ordinal data. Ordered response models, which are used to analyze such data, suffer from an identification problem. Their conclusions depend on distributional assumptions about a latent variable. We propose using response times to solve that problem. Response times contain information about the distribution of the latent variable through a chronometric effect. Using an online survey experiment, we verify the chronometric effect. We then provide theoretical conditions for testing conventional distributional assumptions. These assumptions are rejected in some cases, but overall our evidence is consistent with the qualitative validity of the conventional models.* (*JEL* C14, D60, D91, I31)

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

Surveys measuring subjective states like happiness or preferences usually generate discrete ordinal data (Likert 1932). For example, the life happiness question in

*Shuo Liu: Guanghua School of Management, Peking University (email: shuo.liu@gsm.pku.edu.cn); Netzer: Department of Economics, University of Zurich (email: nick.netzer@econ.uzh.ch). Isaiah Andrews was the coeditor for this article. This paper was previously circulated as "Happy Times: Identification from Ordered Response Data." We are grateful for very helpful comments to two anonymous referees. We also thank Carlos Alós-Ferrer, Timothy Bond, Yu Gao, Lukas Hensel, Damian Kozbur, Ian Krajbich, Xi Zhi Lim, Andrew Oswald, Rainer Winkelmann, and seminar participants at Central European University, HKBU-Kyoto-Osaka-NTU-Sinica, Jinan University, Nanjing University, Oxford University, Peking University, Renmin University of China, SFU and UBC Vancouver, Shanghai University of Finance and Economics, Shenzhen University, Tsinghua University, Wuhan University, Zhejiang University, the University of Zurich, the CESifo Area Conference on Behavioral Economics 2021, and the SSES Annual Congress 2021. Huaqing Huang and Jindi Huang provided excellent research assistance. Shuo Liu acknowledges financial support from the National Natural Science Foundation of China (grants 72103006 and 72322006). The empirical part of the research described in this paper received IRB approval from the Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich (OEC IRB 2019-042).

[†]Go to https://doi.org/10.1257/aer.20211051 to visit the article page for additional materials and author disclosure statements.

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

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

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

The idea that subjects respond faster when a stimulus is further away from an indecision threshold is not new. This *chronometric effect* has been documented in many studies in psychology, neuroscience and economics. In some of these studies, the stimulus is objective, such as the difference in brightness between two lights. Kellogg (1931) has first shown that subjects identify the brighter light faster if the difference in brightness becomes larger. The same is true in tasks where the larger of two objects has to be identified (Moyer and Bayer 1976), or the direction of

[1] Bond and Lang (2019) point out 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. They also discuss a literature that uses variation in observables to achieve nonparametric identification, such as Cunha, Heckman, and Navarro (2007), but argue that this requires assumptions which are not plausible in the context of happiness surveys.

random dot motion (Palmer, Huk, and Shadlen 2005). Making the decision easier, by magnifying the stimulus away from the indecision threshold, shortens response times. In other studies, the stimulus is subjective, for example the utility difference between two options in an economic choice task. Moffatt (2005) has shown that choice between two lotteries becomes faster when the utility difference between the lotteries becomes larger. The same has been documented for intertemporal choices (Chabris et al. 2009; Konovalov and Krajbich 2019) and choices between food items (Krajbich, Armel, and Rangel 2010). Again, making the decision easier, by increasing the strength of preference away from the indifference point, shortens response times.[2] In the empirical part of our paper, we will later demonstrate that the chronometric effect exists in surveys as well.

In the theoretical part of the paper, we integrate response times into the ordered response model in a way that reflects the chronometric effect. Following Bond and Lang (2019), we first consider a simple version of the model that incorporates neither heterogeneity nor noise. 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 $\tau^1 < \tau^2 < \cdots < \tau^n$, 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. This may appear like a strong requirement, but first-order stochastic dominance is assumed in standard models like ordered probit. It implies that the groups' average happiness can be ranked unambiguously, irrespective of the cardinal scale of happiness. Bond and Lang (2019) show that detecting dominance is possible only under extremely stringent conditions. For instance, in a survey with two response categories, all participants in one group must respond to be happy and all participants in the other group must respond to be unhappy. If there are more than two categories, the condition is stronger than first-order stochastic dominance of the observed response distributions of the groups, and there still cannot be any responses in the lowest (highest) category from the group that is more happy (unhappy).

Now suppose responses display the chronometric effect. Consider first a happiness survey with two response categories. The response time of a participant with happiness $h \leq \tau^1$ is $c^0\left(\tau^1 - h\right)$, where $c^0$ is a strictly decreasing but unknown *chronometric function*, reflecting that the answer becomes easier and thus quicker for the participant when the distance $\tau^1 - h$ between the stimulus $h$ and the indecision threshold $\tau^1$ becomes larger. Similarly, the response time of a participant with happiness $h > \tau^1$ is $c^1\left(h - \tau^1\right)$ for a strictly decreasing chronometric

[2] There are many more studies documenting the chronometric effect in a variety of domains, which we cannot summarize here. See Alós-Ferrer, Fehr, and Netzer (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.

function $c^1$. We assume here that the chronometric function can be category-specific but is the same for all participants. This is analogous to the assumption of identical reporting thresholds for all participants in traditional ordered response models. If the distribution of response times is observed in addition to the survey responses, the conditions for detecting first-order stochastic dominance of the happiness distributions become 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 $h \geq \tau^1 + \left(c^1\right)^{-1}(t)$ 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 necessary and sufficient for dominance detection. These conditions are much weaker than the conditions in Bond and Lang (2019). For $t \to \infty$, they merely imply that the fraction of participants who respond to be happy must be higher in group $A$ than in group $B$, and not that these fractions have to be one and zero. Our conditions are stricter than with traditional ordered response models, because the inequalities have to hold for all response times.

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 $\left[\tau^i, \tau^{i+1}\right]$, it moves away from one indecision threshold but closer towards the other. Hence, any plausible specification of the chronometric effect generates response times that are not monotone in $h$ between two interior reporting thresholds. As a consequence, response times from intermediate response categories are uninformative, and our detection condition coincides with that in Bond and Lang (2019) for these categories. Our results thus make a case for surveys with just two response categories. Due to their continuous and cardinal nature, recording response times may be more important than recording fine-grained responses.

We then generalize the simple baseline model to make it suitable for statistical analysis. We allow for arbitrary differences between individuals or groups in the speed of making their decisions or submitting their responses. We formalize the idea of normalizing response times using the response time from a baseline question to account for such differences. We also allow for stochastic reporting thresholds and general noise or measurement error in the response times under an i.i.d. assumption. Our main result is that whenever the true happiness distributions of two groups exhibit first-order stochastic dominance, as assumed in conventional models, then the above-described detection conditions on the response time distributions are necessarily satisfied. Furthermore, responses in intermediate categories (if they exist) must satisfy conventional first-order stochastic dominance. This result is useful because it allows us to test and possibly falsify assumptions of conventional models. When the conditions are violated, then the true happiness distributions cannot exhibit a first-order stochastic dominance ranking, and consequently the conventional results are sensitive to the choice of the scale and not qualitatively robust.

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 test

standard assumptions of ordered response models. In the words of Bond and Lang (2019), response times can 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, Fehr, and Netzer (2021), which studies the problem of eliciting preferences from choice data when choice is stochastic. While surprising at first glance, the problem with ordered response models is similar to the revealed preference problem in random utility models. In the latter, the utility difference between two choice options of an agent is an unobserved random variable which generates stochastic choices. Without assumptions on its distribution (e.g., logistic in a Luce model) it is not possible to deduce the agent's underlying deterministic utility function from observed choices. Alós-Ferrer, Fehr, and Netzer (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, Fehr, and Netzer (2021). Most importantly, revealed preference questions are questions about the properties of a single distribution (of the utility difference between the choice options). The 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 with about 8,000 participants that we conducted on Amazon Mechanical Turk (MTurk). We asked several sociodemographic questions and several substantive questions about happiness, preferences, trust, and political attitude. These questions were adopted from the GSS and from Falk et al. (2018). We implemented two versions of the survey, one with two answer categories and one with three answer categories. In both versions of the survey, each substantive question was accompanied by a follow-up question in which participants were asked to refine their previous answer. For example, a subject giving the highest possible response "rather happy" in the initial question about overall life happiness subsequently had the choice between "very happy" and "moderately happy" in the follow-up question.

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

We first use responses from the follow-up questions to test for the existence of chronometric effects in our survey. We find that, among subjects who initially gave an identical answer, those who reveal a more extreme position in the follow-up question responded faster on average in the initial question. More specifically, we consider all subjects who responded in the same extreme category in an initial question (e.g., "rather happy") and partition them into two subgroups based on their response in the follow-up question. Those who give a more extreme response in the follow-up (e.g., "very happy") should have larger values of the latent variable than those who give a more moderate response (e.g., "moderately happy"). The chronometric effect then predicts that the former should have responded more quickly in the initial question than the latter. We find this prediction confirmed in our data, for both extreme

response categories in all seven substantive questions and both versions of the survey. Among the 28 pairwise comparisons that we make, 25 are statistically significant at the 1 percent level. We further confirm in pooled regression analyses that giving a more extreme response in the follow-up question is associated with significantly quicker responses in the initial question, even when controlling for demographics or individual fixed effects. In other words, subjects for whom the latent variable is further away from the respective reporting threshold tend to give a quicker response.

Having confirmed the chronometric effect, we test the null hypothesis of first-order stochastic dominance of the latent distributions using the response time data. We compare different sociodemographic groups pairwise and, for each substantive question, visualize the prediction generated by the dominance assumption of traditional models by plotting the evolution of response fractions over response time. To make statistical inference, we exploit the similarity between our novel test conditions and the standard conditions for stochastic dominance, which leads to a bootstrap-based test adapted from Barrett and Donald (2003). Our paper is accompanied by Stata ado-files which implement these procedures.

Using the binary version of the survey, our findings reveal interesting patterns. For our question about time preferences, we often reject the null hypothesis of first-order stochastic dominance, suggesting that distributions of discount rates may be less regular than what is postulated by traditional ordered response models and that the estimated coefficients must therefore be interpreted with caution. This is in contrast to our risk preference question, where the null hypothesis can typically not be rejected. For our two satisfaction questions concerning work and social life and for our trust question, we are also unable to reject the first-order stochastic dominance assumption in almost all comparisons. The questions about overall life happiness and political attitude are somewhere in between, with a rejection of the null hypothesis in some cases but not in others. Overall, we show that significance of the ordered probit coefficient is correlated with the inability to reject the null hypothesis of first-order stochastic dominance. We interpret this as first cautious evidence that significantly estimated parameters of traditional ordered response models tend to be qualitatively robust, but we also demonstrate that our use of response time data is a simple technique to assess the validity of a given estimate.

We conduct the same analysis for the trinary version of the survey. Among the differences, the main one appears to be that the $p$-values of our test are often larger in the trinary survey than in the binary survey, suggesting that the test has higher power in the binary case.

The paper is organized as follows. Section I presents our theoretical results. Section II reports the empirical findings. An extended literature discussion is in Section III, and Section IV concludes. The complete questionnaires of our survey experiment and some additional empirical results can be found in the online Appendix.

## I. Theory

### A. *Baseline Model*

We first introduce a baseline model that has also been studied by Bond and Lang (2019). This model is instructive but incorporates neither heterogeneity nor noise. We will extend the model to make it suitable for statistical analysis in the next subsection.

Consider two groups $j = A,B$ of individuals. The distribution of happiness $h \in \mathbb{R}$ within group $j$ is described by a cumulative distribution function $G_j : \mathbb{R} \to [0,1]$ that is assumed to be continuous. A data analyst does not observe individual happiness but observes only the individuals' survey responses on a finite ordered scale, with categories labeled $i = 0,\ldots,n$ for some $n \geq 1$. The latent variable $h$ generates responses through reporting thresholds $\tau = (\tau^1,\tau^2,\ldots,\tau^n) \in \mathbb{R}^n$ which satisfy $\tau^1 < \tau^2 < \cdots < \tau^n$ and are assumed to be the same for all individuals in both groups. 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 nonparametric detection of first-order stochastic dominance.

DEFINITION 1: *Given* $(r_A, r_B)$, *group A is* ***detectably rank-order happier*** *than group B if*

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

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

Rank-order detection requires $G_A$ to 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 assumed in applications of, e.g., the classical ordered probit model. It is a well-known fact that $G_A$ FOSD $G_B$ is equivalent to

$$\int_{\mathbb{R}} q(h) dG_A(h) \geq \int_{\mathbb{R}} q(h) dG_B(h)$$

for all weakly increasing functions $q : \mathbb{R} \to \mathbb{R}$ (see e.g. Hanock and Levy 1969). Hence, rank-order detection implies a ranking of the groups' average happiness no matter which "cardinalization" (Bond and Lang 2019, p. 1630) we choose to define the happiness scale.

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

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

(*i*) $r_A^0 = 0$,

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

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

PROOF:

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

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

and

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

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 $(G_A, G_B, \tau)$ 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 $(G_A, G_B, \tau)$, 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

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

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 $r_B^n > 0$. ∎

The necessary and sufficient conditions for rank-order detection are particularly striking in the binary response case, where they require that all individuals in group $A$ report to be happy ($r_A^1 = 1$) and all individuals in group $B$ report to be unhappy ($r_B^0 = 1$). In general, conditions (i)–(iii) apply for any number of categories, whether small or large. They are essentially never satisfied in real-world data, as shown by Bond and Lang (2019).

Assume now that 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 $0 \leq \underline{t} < \bar{t} < \infty$. Response times are related to the latent variable $h$ through chronometric functions $c^i : \mathbb{R}_+ \to [\underline{t}, \bar{t}]$, which may be specific to each response category $i = 0, \ldots, n$. Each function $c^i$ is assumed to be continuous, strictly decreasing in $\delta$ whenever $c^i(\delta) > \underline{t}$, and to satisfy $c^i(0) = \bar{t}$ and $\lim_{\delta \to +\infty} c^i(\delta) = \underline{t}$. These chronometric functions are assumed to be the same for all individuals in both groups, analogous to the assumption of identical reporting thresholds. To understand how response times are generated, consider binary surveys ($n = 1$) first. An individual with happiness $h \leq \tau^1$ responds in category $i = 0$ at time $c^0(\tau^1 - h)$. 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 time $c^1(h - \tau^1)$. Allowing the chronometric functions to be category-specific is important when absolute happiness levels directly affect response times, with e.g. more unhappy people being slower (Studer and Winkelmann 2014). We accommodate such effects as long as they do not reverse the monotone chronometric relation within the extreme response categories.[3]

There are various ways how the chronometric effect could be modeled for intermediate response categories $i = 1, \ldots, n-1$ when $n \geq 2$. In the following, we adopt a symmetric specification where response time is driven by the distance between happiness and the closest reporting threshold. Thus, an individual with happiness $\tau^i < h \leq \tau^{i+1}$ responds in category $i$ at time $c^i(\min\{h - \tau^i, \tau^{i+1} - h\})$. Figure 1 depicts an example of response times arising from a data-generating process that satisfies all our requirements and in which the chronometric function is identical in all response categories.

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

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

[3] Our way of adding chronometric functions to an ordered response model is analogous to how Alós-Ferrer, Fehr, and Netzer (2021) add a chronometric function to a random utility model. They consider binary choice problems and assume that response time is monotonically driven by the absolute realized value of the utility difference between the two choice options. In contrast to our setting, they do not allow the chronometric function to be different for the two choice options.

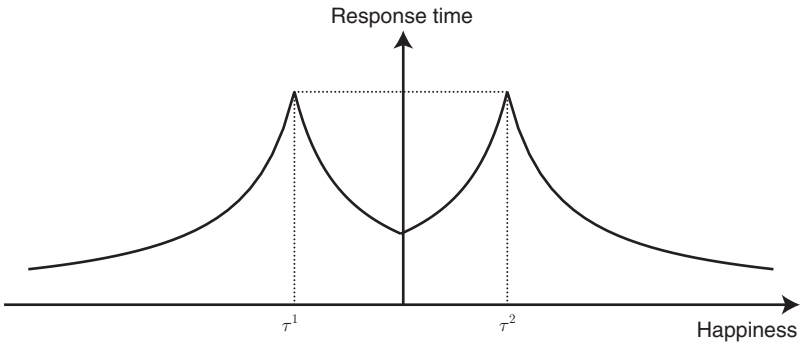


FIGURE 1. EXAMPLE OF RESPONSE TIMES WITH $n = 2, \tau^1 = -2, \tau^2 = 2$, AND $c^i(\delta) = 1/(\delta + 1)$

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

The analyst can now ask the previous question about the happiness distributions, using data on both responses $r_j = \left(r_j^0, r_j^1, \ldots, r_j^n\right)$ and response times $F_j = \left(F_j^0, F_j^1, \ldots, F_j^n\right)$, where each cumulative distribution function $F_j^i$ is assumed to be continuous and to satisfy $F_j^i(\underline{t}) = 0$ and $F_j^i(\bar{t}) = 1$.[4] Following Definition 1, we will say that group $A$ is detectably rank-order happier than group $B$ if $G_A$ FOSD $G_B$ holds in all combinations $\left(G_A, G_B, \tau, c\right)$ of happiness distributions, reporting thresholds, and chronometric functions $c = \left(c^0, \ldots, c^n\right)$ that satisfy (1) and (2) for $i = 0, \ldots, n, j = A, B$, and all $t \in \left[\underline{t}, \bar{t}\right]$. The first result of our paper is a characterization of rank-order detection using response times.

PROPOSITION 2: *Given* $\left(r_A, r_B, F_A, F_B\right)$, *group A is detectably rank-order happier than group B if and only if*

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

(*ii*) $r_A^n F_A^n(t) - r_B^n F_B^n(t) \geq 0$ *for all* $t \in \left[\underline{t}, \bar{t}\right]$, *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 $\left(G_A, G_B, \tau, c\right)$ satisfy (1) and (2) for $i = 0, \ldots, n, j = A, B$, and all $t \in \left(\underline{t}, \bar{t}\right]$. For $i = 0$ this implies

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

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

for all $t \in (\underline{t},\bar{t}]$. Thus, condition (i) implies $G_A\left(\tau^1 - \left(c^0\right)^{-1}(t)\right) \leq G_B\left(\tau^1 - \left(c^0\right)^{-1}(t)\right)$ 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 $h = \tau^1 - \left(c^0\right)^{-1}(t)$. For any $h$ with $c^0\left(\tau^1 - h\right) = \underline{t}$ 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 $t^* \in (\underline{t},\bar{t})$, where assuming $\underline{t} < t^* < \bar{t}$ is without loss of generality by continuity of $F_j^0$. Any $(G_A, G_B, \tau, c)$ that satisfies equations (1) and (2) for $i = 0,\ldots,n, j = A,B$, and all $t \in (\underline{t},\bar{t}]$ must then have $G_A\left(\tau^1 - \left(c^0\right)^{-1}(t^*)\right) > G_B\left(\tau^1 - \left(c^0\right)^{-1}(t^*)\right)$, so that $G_A$ FOSD $G_B$ is not true. An analogous argument applies when $r_A^n F_A^n(t^*) - r_B^n F_B^n(t^*) < 0$ for some $t^* \in (\underline{t},\bar{t})$. Finally, suppose that there exists $k^* = 1,\ldots,n-1$ for which $\sum_{i=0}^{k^*} r_A^i > \sum_{i=0}^{k^*-1} r_B^i$. Starting from any $(G_A, G_B, \tau, c)$ 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 $h \in \left(\tau^{k^*}, \tau^*\right)$, where $\tau^* := \left(\tau^{k^*} + \tau^{k^*+1}\right)/2$, in a specific way:

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

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

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

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

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

this conclusion. Therefore, the power of our weaker conditions becomes particularly apparent in binary surveys, where rank-order detection obtains if and only if $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq 0 \leq r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ for all $t \in [\underline{t}, \bar{t}]$.[5]

## B. *Extended Model*

To make it suitable for statistical analysis, we now generalize the previous baseline model to allow for individual heterogeneity in reporting thresholds and chronometric effects, and to incorporate noise or measurement error more generally. Furthermore, since we can only reject but never confirm null hypotheses using statistical testing, we will derive a general necessary condition that allows us to test and possibly reject the null hypothesis that there is first-order stochastic dominance between the latent distributions of two groups.

As before, consider an individual who answers a happiness question with $n+1$ response categories. The individual responds in category $i = 0, \ldots, n$ when $\tau^i < h \leq \tau^{i+1}$, where $h$ is the individual's happiness and $\tau = (\tau^1, \ldots, \tau^n)$ are the reporting thresholds, and where we fix $\tau^0 = -\infty$ and $\tau^{n+1} = +\infty$ throughout. In the extended model, the response time of the individual is given by

$$t = c(h, \tau) \cdot \eta \cdot \epsilon,$$

where $c : \mathbb{R}^{n+1} \to [\underline{t}, \bar{t}]$ is a general chronometric function, $\eta > 0$ is an individual-specific speed parameter, and $\epsilon > 0$ captures additional uncontrolled factors that may affect response times. We impose assumptions only on the extreme categories of the chronometric function. We assume that, whenever $h \leq \tau^1$, then $c(h, \tau) = c^0(\tau^1 - h)$ for some $c^0 : \mathbb{R}_+ \to [\underline{t}, \bar{t}]$ that satisfies our previous assumptions. Analogously, whenever $\tau^n < h$, then $c(h, \tau) = c^n(h - \tau^n)$ for some $c^n : \mathbb{R}_+ \to [\underline{t}, \bar{t}]$ with our previous assumptions.

The distribution of the individual characteristics $(h, \eta)$ in group $j$ is described by a joint cumulative distribution function $G_j(h, \eta)$, allowing for correlation between $h$ and $\eta$ within a group. These distributions are not necessarily continuous and can be different between the groups, allowing for systematic group-differences in response speed. We denote by $G_j(h)$ the corresponding marginal distribution of happiness and simply refer to it as the happiness distribution of group $j$.

In addition to the happiness question, the individual answers a baseline question. This could be a demographic question about e.g. marital status (possibly, but not necessarily, defining membership to groups $j = A, B$) or a question asking for agreement to participate in the survey. The important assumption is that there are

[5] One may ask whether we can obtain even weaker conditions when requiring directly the detection of the ranking of average happiness between the groups, rather than first-order stochastic dominance. The answer is no for the case without response times, where the conditions in Proposition 1 also characterize detection of the average ranking. The answer is yes for the case with response times, if we are willing to assume that the chronometric function is identical across response categories. In a binary survey, the weaker inequality $r_A^0 F_A^0(t) - r_B^0 F_B^0(t) \leq r_A^1 F_A^1(t) - r_B^1 F_B^1(t)$ for all $t \in [\underline{t}, \bar{t}]$ then already implies that average happiness is higher in group $A$ than in group $B$, for all distributions that could have generated the data. In contrast to the condition for rank-order detection, some fast "unhappy" responses in group $A$ relative to group $B$ can be compensated by even more and faster "happy" responses. See our discussion paper (Liu and Netzer 2020) for details and formal results.

no systematically varying intensities of responses in the baseline question. The response time in the baseline question is

$$t_b = \phi \cdot \eta,$$

where $\eta$ is the individual's speed parameter as before and $\phi > 0$ subsumes all other factors that may affect response time. We ignore the response in the baseline question, other than that it may define group membership. We only use the baseline question to normalize the response time in the happiness question by dividing the latter by the response time in the baseline question, which gives rise to the following:

$$\hat{t} = \frac{t}{t_b} = c(h,\tau) \cdot \frac{\epsilon}{\phi}.$$

In the extended model, we also allow the parameters $(\tau, \epsilon, \phi)$ to be stochastic. In words, individuals may employ noisy reporting thresholds (as in Cunha, Heckman, and Navarro 2007) and there could be additional response-level errors. For example, the "time to response" that the analyst measures in an online survey may be inflated when the individual was distracted for a specific question. Normalization will not take care of such measurement error because the error is not the same across questions for an individual. We assume that the parameters $(\tau, \epsilon, \phi)$ are distributed identically in both groups and independently of all other variables. We describe their distribution by a probability measure $\mu$ on $\mathbb{R}^{n+2}$ with a support which respects that the reporting thresholds are always ordered $\tau^1 < \tau^2 < \cdots < \tau^n$ and that $\epsilon$ and $\phi$ are positive.

We can now state the main result of our paper, which gives properties of responses and distributions of normalized response times that must be satisfied whenever the true happiness distributions exhibit first-order stochastic dominance.

PROPOSITION 3: *Suppose that the happiness distribution of group A first-order stochastically dominates that of group B. The generated normalized data* $(r_A, r_B, \hat{F}_A, \hat{F}_B)$ *then satisfy*

(*i*) $r_A^0 \hat{F}_A^0(t) - r_B^0 \hat{F}_B^0(t) \leq 0$ *for all* $t \geq 0$,

(*ii*) $r_A^n \hat{F}_A^n(t) - r_B^n \hat{F}_B^n(t) \geq 0$ *for all* $t \geq 0$, *and*

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

PROOF:

Consider first the response fractions generated by the described process. We obtain

$$\sum_{i=0}^{k} r_A^i = \int_{\mathrm{supp}(\mu)} G_A\left(\tau^{i+1}\right) d\mu(\tau,\epsilon,\phi) \leq \int_{\mathrm{supp}(\mu)} G_B\left(\tau^{i+1}\right) d\mu(\tau,\epsilon,\phi) = \sum_{i=0}^{k} r_B^i$$

for all $k = 0, \ldots, n$, where the inequality follows from $G_A$ FOSD $G_B$. This implies condition (iii) in the proposition.

Consider next the response times generated in category $i = 0$. Since in this category $\hat{t} = c(h,\tau)\cdot\epsilon/\phi \leq t$ if and only if $c^0(\tau^1 - h) \leq t\cdot\phi/\epsilon$, the following function $P_j^0(t\,|\,\tau,\epsilon,\phi)$ describes the fraction of individuals in group $j$ who respond in category 0 at time $t \geq 0$ or earlier when $(\tau,\epsilon,\phi)$ is held fixed:

$$P_j^0(t\,|\,\tau,\epsilon,\phi) = \begin{cases} G_j(\tau^1), & \text{if } \bar{t} \leq t\cdot\phi/\epsilon; \\ G_j\left(\tau^1 - \left(c^0\right)^{-1}(t\cdot\phi/\epsilon)\right), & \text{if } \underline{t} < t\cdot\phi/\epsilon < \bar{t}; \\ G_j(\tau^1 - \delta^0), & \text{if } t\cdot\phi/\epsilon = \underline{t}; \\ 0, & \text{if } t\cdot\phi/\epsilon < \underline{t}; \end{cases}$$

where $\delta^0 = \lim_{t\downarrow\underline{t}}\left(c^0\right)^{-1}(t)$, which is finite when $c^0(\delta) = \underline{t}$ for sufficiently large $\delta$, and $\delta^0 = +\infty$ and thus $G_j(\tau^1 - \delta^0) = 0$ otherwise. This implies

$$\begin{aligned} r_A^0\hat{F}_A^0(t) &= \int_{\text{supp}(\mu)} P_A^0(t\,|\,\tau,\epsilon,\phi)d\mu(\tau,\epsilon,\phi) \leq \int_{\text{supp}(\mu)} P_B^0(t\,|\,\tau,\epsilon,\phi)d\mu(\tau,\epsilon,\phi) \\ &= r_B^0\hat{F}_B^0(t) \end{aligned}$$

for all $t \geq 0$, where the inequality follows because $G_A$ FOSD $G_B$ implies $P_A^0(t\,|\,\tau,\epsilon,\phi) \leq P_B^0(t\,|\,\tau,\epsilon,\phi)$ for all $(\tau,\epsilon,\phi)$ in the support of $\mu$. This gives condition (i) in the proposition.

Consider then the response times in category $n$. The following function $P_j^n(t\,|\,\tau,\epsilon,\phi)$ describes the fraction of individuals in group $j$ who respond in category $n$ at time $t \geq 0$ or earlier when $(\tau,\epsilon,\phi)$ is held fixed:

$$P_j^n(t\,|\,\tau,\epsilon,\phi) = \begin{cases} 1 - G_j(\tau^n), & \text{if } \bar{t} \leq t\cdot\phi/\epsilon; \\ 1 - G_j\left(\tau^n + \left(c^n\right)^{-1}(t\cdot\phi/\epsilon)\right), & \text{if } \underline{t} < t\cdot\phi/\epsilon < \bar{t}; \\ 1 - G_j(\tau^n + \delta^n), & \text{if } t\cdot\phi/\epsilon = \underline{t}; \\ 0, & \text{if } t\cdot\phi/\epsilon < \underline{t}; \end{cases}$$

where $\delta^n = \lim_{t\downarrow\underline{t}}\left(c^n\right)^{-1}(t)$, which is finite when $c^n(\delta) = \underline{t}$ for sufficiently large $\delta$, and $\delta^n = +\infty$ and thus $1 - G_j(\tau^n + \delta^n) = 0$ otherwise. Thus,

$$\begin{aligned} r_A^n\hat{F}_A^n(t) &= \int_{\text{supp}(\mu)} P_A^n(t\,|\,\tau,\epsilon,\phi)d\mu(\tau,\epsilon,\phi) \geq \int_{\text{supp}(\mu)} P_B^n(t\,|\,\tau,\epsilon,\phi)d\mu(\tau,\epsilon,\phi) \\ &= r_B^n\hat{F}_B^n(t) \end{aligned}$$

for all $t \geq 0$, again by $G_A$ FOSD $G_B$. This gives condition (ii) in the proposition. ∎

The proposition shows that response time conditions (i) and (ii), which were part of the necessary and sufficient conditions for rank-order detection in our simple baseline model, will still be satisfied by the normalized data in our substantially extended model whenever there is first-order stochastic dominance of the true (marginal) happiness distributions. While systematic speed differences between individuals or groups, as captured by $\eta$, may invalidate these conditions in the raw

data, normalization restores them. The noise in the parameters $(\tau, \epsilon, \phi)$ affects both groups equally and thus does not invalidate the conditions either.[6]

Conversely, if these conditions are violated, then the true happiness distributions cannot exhibit a first-order stochastic dominance relation. This statement is stronger than an "only if" statement about rank-order detection as in Proposition 2. From Proposition 2 we learn that at least one possible data-generating process must violate first-order stochastic dominance when the respective conditions are violated, while from Proposition 3 we learn that all possible data-generating processes must violate first-order stochastic dominance when the respective conditions are violated. Therefore, Proposition 3 will form the basis for hypothesis testing in our empirical application.

For the special case of our simple baseline model, condition (iii) for intermediate response categories in Proposition 3 is weaker than the corresponding detection condition in Proposition 2 (since the summation on the right-hand side of the former is up to category $k$ while for the latter it is up to category $k-1$). Hence, even if there is first-order stochastic dominance in the happiness distributions, this dominance relation may not be detectable in a survey with more than two response categories. In that sense, a binary survey may be more useful for detection than a multi-category one when response times are available.

## II. Empirical Application

### A. *Survey Description*

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

Our survey was programmed using the software Qualtrics and was conducted in April and May of 2022 through the ETHZ Decision Science Laboratory.[7] The survey consisted of two parts. The first part included six standard sociodemographic questions concerning gender, age, education, marital status, co-residence with

[6] Noise could invalidate these conditions if it is either systematically different between the groups or correlated with happiness. For example, the two groups may differ in the attention that they bring to a complex question like happiness but not to a simple question like marital status, in which case normalization cannot address the issue and we would get group-specific distributions of $\epsilon$. The presence of noise also makes it difficult to obtain sufficient detection conditions for the extended model, because violations of first-order stochastic dominance of the happiness distributions may be smoothed out by noise and therefore not detectable in the data.

[7] The replication files are publicly available (Liu and Netzer 2023). The first discussion paper version of this paper (Liu and Netzer 2020) contains data from another survey conducted on MTurk already in 2019. This older survey had a smaller number of participants, no question about income, and it did not contain follow-up questions. We are not using those data here.


FIGURE 2. EXAMPLE OF SURVEY SCREEN

children, and family income. These questions are commonly asked in large-scale surveys like the GSS, which is the primary source for US evidence on a broad set of social science issues (Davis and Smith 1991). In the second part, the subjects were asked seven substantive questions. These questions elicit information about (i) job satisfaction, (ii) social life satisfaction, (iii) overall happiness, (iv) trust attitude, (v) political attitude, (vi) time preference, and (vii) risk preference. The questions for (i)–(v) were again adapted from the GSS, and for (vi) and (vii) the questions were adapted from the Global Preference Survey introduced by Falk et al. (2018).

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

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

Table 1—Summary of Subject Demographics

| | Binary survey | Trinary survey |
|---|---|---|
| Number of participants | 3,743 | 3,724 |
| Female (%) | 50.09 | 51.34 |
| Male (%) | 49.91 | 48.66 |
| Age (%) | | |
| $< 20$ | 0.37 | 0.62 |
| 20–29 | 24.39 | 26.91 |
| 30–39 | 34.84 | 32.92 |
| 40–49 | 21.88 | 21.51 |
| 50–59 | 11.38 | 11.09 |
| 60–69 | 6.09 | 5.99 |
| $\geq 70$ | 1.04 | 0.97 |
| Highest education (%) | | |
| High school | 17.18 | 17.37 |
| College or higher | 82.29 | 82.28 |
| None | 0.53 | 0.35 |
| Married (%) | 64.63 | 65.15 |
| Unmarried (%) | 35.37 | 34.85 |
| Kids (%) | 60.86 | 61.52 |
| No kids (%) | 39.14 | 38.48 |
| Income (%) | | |
| $< \$40{,}000$ | 26.90 | 27.12 |
| \$40,000–\$69,999 | 43.92 | 43.98 |
| $\geq \$70{,}000$ | 29.17 | 28.89 |

## B. *Summary Statistics*

We recruited 8,007 subjects from the United States with an MTurk approval rate of at least 95 percent.[8] Each subject received a fixed compensation of 60 cents for completing the survey. In the initial sample, 286 subjects failed an attention check at the end of the survey ("What is 7 times 2?"). No click and time data were recorded for 253 subjects, presumably because they used keyboard navigation to answer the questions, and for one subject several recorded response times were zero. All these subjects were dropped, so our final sample contains 3,743 subjects in the binary survey and 3,724 subjects in the trinary survey. Table 1 summarizes the demographics of the subjects and shows that they are very similar in the two survey versions, as should be expected given the random assignment.[9]

[8] We initially restricted access to subjects with an MTurk approval rate of at least 98 percent. Recruitment became slow over time, so after the first 5,976 subjects, the approval requirement was reduced to 95 percent.

[9] After the survey was completed, we became aware that a significant number of observations in our dataset had suspiciously similar IP addresses. Specialists of the ETHZ Decision Science Laboratory conjectured that these observations may be from participants using virtual private networks (VPNs), but the ultimate source of the pattern remains unknown. Following the suggestion of the ETHZ Decision Science Laboratory, as a conservative robustness check we excluded all observations where the first three IP blocks appeared more than once, which amounts to about 40 percent of our data. Online Appendix C contains all our main results based on the restricted sample. Participants in the restricted sample report to be somewhat less educated, be less often married, have children less often, and have a more spread-out income distribution, but are otherwise similar. The results of our analyses are

Table 2. Median Response Times in Seconds

| | Binary survey | Trinary survey |
|---|---|---|
| Complete survey | 119 | 128 |
| Demographic questions | | |
| Gender | 1.66 | 1.66 |
| Age | 2.05 | 2.07 |
| Education | 2.06 | 2.08 |
| Marital status | 1.52 | 1.51 |
| Kids | 1.73 | 1.73 |
| Income | 2.18 | 2.16 |
| Substantive questions | | |
| Work satisfaction | 2.58 | 3.33 |
| Social life satisfaction | 2.49 | 2.84 |
| Overall happiness | 2.94 | 3.42 |
| Trust | 3.26 | 4.03 |
| Political attitude | 2.12 | 2.21 |
| Time preference | 3.98 | 4.32 |
| Risk preference | 2.34 | 2.62 |

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

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

### C. *Testing the Chronometric Effect*

In our survey, each substantive question was accompanied by a follow-up question requiring the subjects to refine their initial response. This design makes it

also largely comparable to those for the full dataset. The chronometric effect is confirmed in the restricted sample, and the results of ordered probit are comparable (for example, we never obtain significant parameter estimates of opposite sign). There are differences in the $p$-values of our FOSD tests, with a majority of those values being larger in the restricted sample, presumably reflecting the substantially smaller sample size. Overall, the $p$-values are positively correlated between the full and the restricted sample.

possible to directly test for the presence of chronometric effects, and in particular our crucial assumption that response times are monotone in the latent variable within the extreme categories. For example, consider only those subjects who responded in the extreme "rather happy" category in the initial question about overall happiness. Based on their response in the corresponding follow-up question, we can further distinguish those who are "very happy" from those who are only "moderately happy," with the former having larger values of the latent variable than the latter. If the chronometric effect exists, then the former should have responded faster than the latter in the initial question. In other words, the chronometric assumption can be tested by the prediction that more extreme follow-up responses should be associated with faster initial responses.

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

$$\log RT_{sq} = \beta_0 + \beta_1 FU_{sq} + \beta_2 \mathbf{X}_s + \gamma_q + \epsilon_{sq}. \tag{3}$$

The dependent variable in (3) is the log of the normalized response time of subject $s$ in initial substantive question $q$ (not including the follow-up). The main explanatory variable of interest is $FU_{sq}$, a dummy that is one if the subject chose the more extreme response among the two given in the corresponding follow-up question (e.g., "very happy" after an initial response of "rather happy," or "very unhappy" after an initial response of "rather unhappy") and zero otherwise. Other controls include the version of the survey that the subject received (binary versus trinary) and the sociodemographic information that our survey collected, all summarized in $\mathbf{X}_s$. Lastly, the variable $\gamma_q$ captures question fixed effects.

Table 3 reports the results of estimating equation (3). As shown in the first row of the table, the coefficient of the dummy variable for an extreme follow-up response is always negative and highly significant. This finding is robust if we include demographic and treatment controls, or if we instead employ a fixed-effect model to control for heterogeneity at the subject level. The regression analysis therefore confirms our central assumption: subjects with more extreme latent values—as revealed by the information that they provide in the follow-up question—respond faster to the initial question.[10]

We can also examine the relation between follow-up responses and response times in the initial question separately for each substantive question. Figures 3 and 4 summarize our findings for the binary and the trinary survey, respectively. As an illustrative example, consider panel C in Figure 3, which concerns the overall happiness question in the binary survey. The subjects are ordered from left to right according to how they responded to the initial question and its follow-up: very unhappy, moderately unhappy, moderately happy, and very happy. Each bar in the graph depicts the average log normalized response time of the respective group in the initial question, along with its 95 percent confidence interval. The chronometric function becomes visible as a hump shape. Among the subjects who initially

[10] Table A1 in online Appendix A shows that nonnormalized, raw response times exhibit the same pattern.

Table 3—Regression Analysis of Chronometric Effects

| | log normalized response time | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Follow-up response | −0.449 | −0.371 | −0.101 |
| | (0.0138) | (0.0133) | (0.0079) |
| $R^2$ | 0.0716 | 0.1066 | 0.0681 |
| Demographics and treatment | No | Yes | No |
| Individual fixed effects | No | No | Yes |

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

responded to be rather unhappy (bars one and two), those who respond in the follow-up to be very unhappy (bar one) were faster in the initial question than those who respond to be only moderately unhappy (bar two). Analogously, among the subjects who initially responded to be rather happy (bars three and four), those who respond in the follow-up to be very happy (bar four) were faster than those who respond to be only moderately happy (bar three). The hump shape confirms that subjects with latent values further away from the reporting threshold give their response more quickly on average.

The hump shape exists for all substantive questions in both versions of the survey. The mean response time is always smaller for the more extreme group than for the less extreme one, and almost all of these differences are statistically significant at the 1 percent level.[11] Altogether, the evidence supports that survey responses display a chronometric effect.[12]

It is worthwhile to ask whether the chronometric effect also exists in the follow-up questions. If this were true, our theory would predict an intriguing correlation of response times between an initial question and its follow-up. Again, take the overall happiness question as an example, and consider the subjects who first responded "rather happy" and then refined their answer to "very happy." Within this group of subjects, response times should be positively correlated between the

[11] Among the 28 pairwise comparisons, 25 are significant at the 1 percent level according to a *t*-test (two-sided, unequal variances), with the exceptions being in the trinary survey: the pair "very unsatisfied" and "moderately unsatisfied" in the work satisfaction question ($p = 0.0737$), the pair "very careful" and "moderately careful" in the trust question ($p = 0.3799$), and the pair "very impatient" and "moderately impatient" in the time preference question ($p = 0.0107$). Figures A1 and A2 in online Appendix A show that similar patterns, albeit less pronounced, exist for nonnormalized, raw response times. Only 7 out of the 28 pairwise comparisons are statistically significant at 1%, but all of them in the direction implied by the chronometric effect.

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

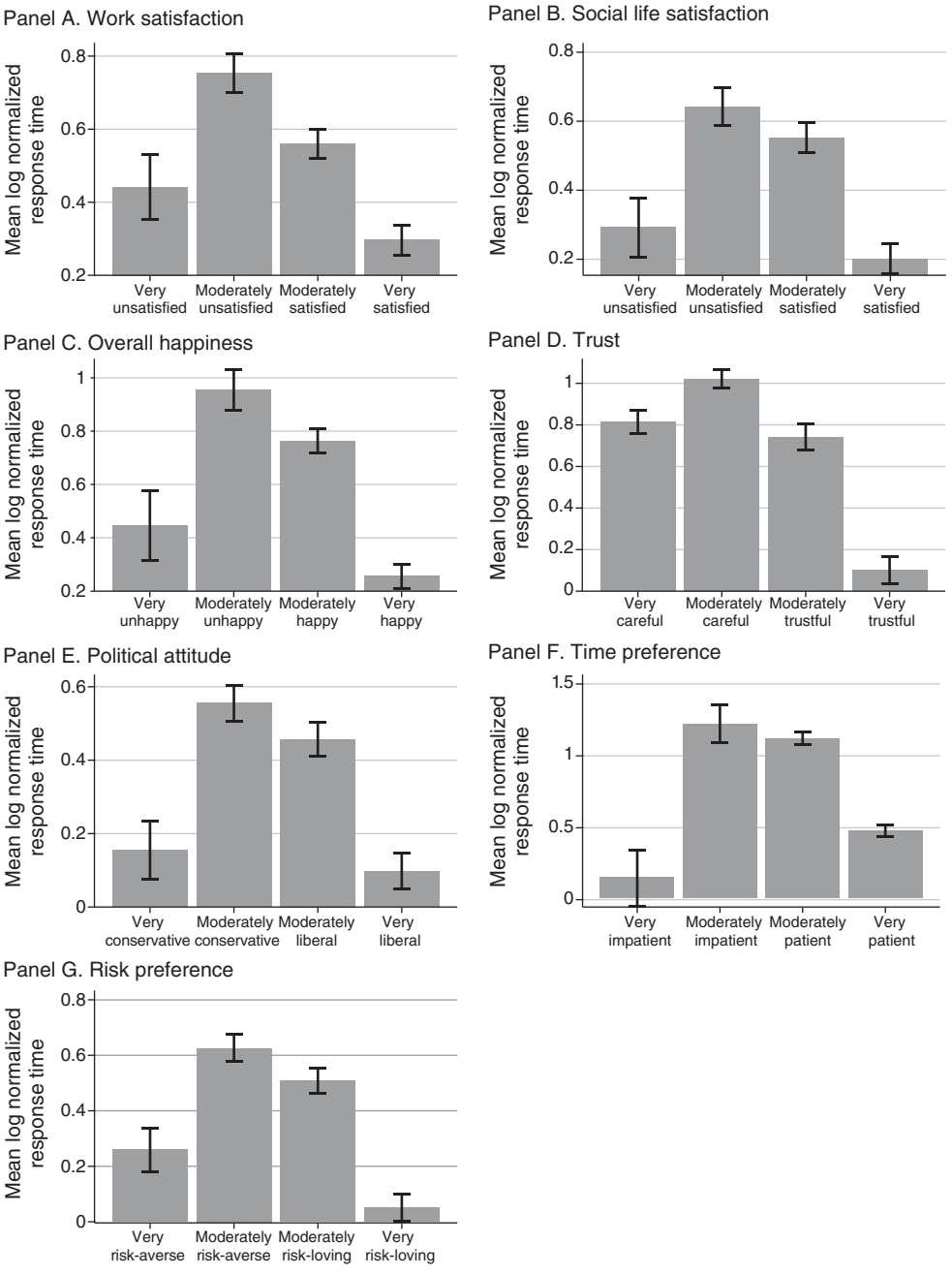


FIGURE 3. CHRONOMETRIC EFFECT BY QUESTION IN THE BINARY SURVEY

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

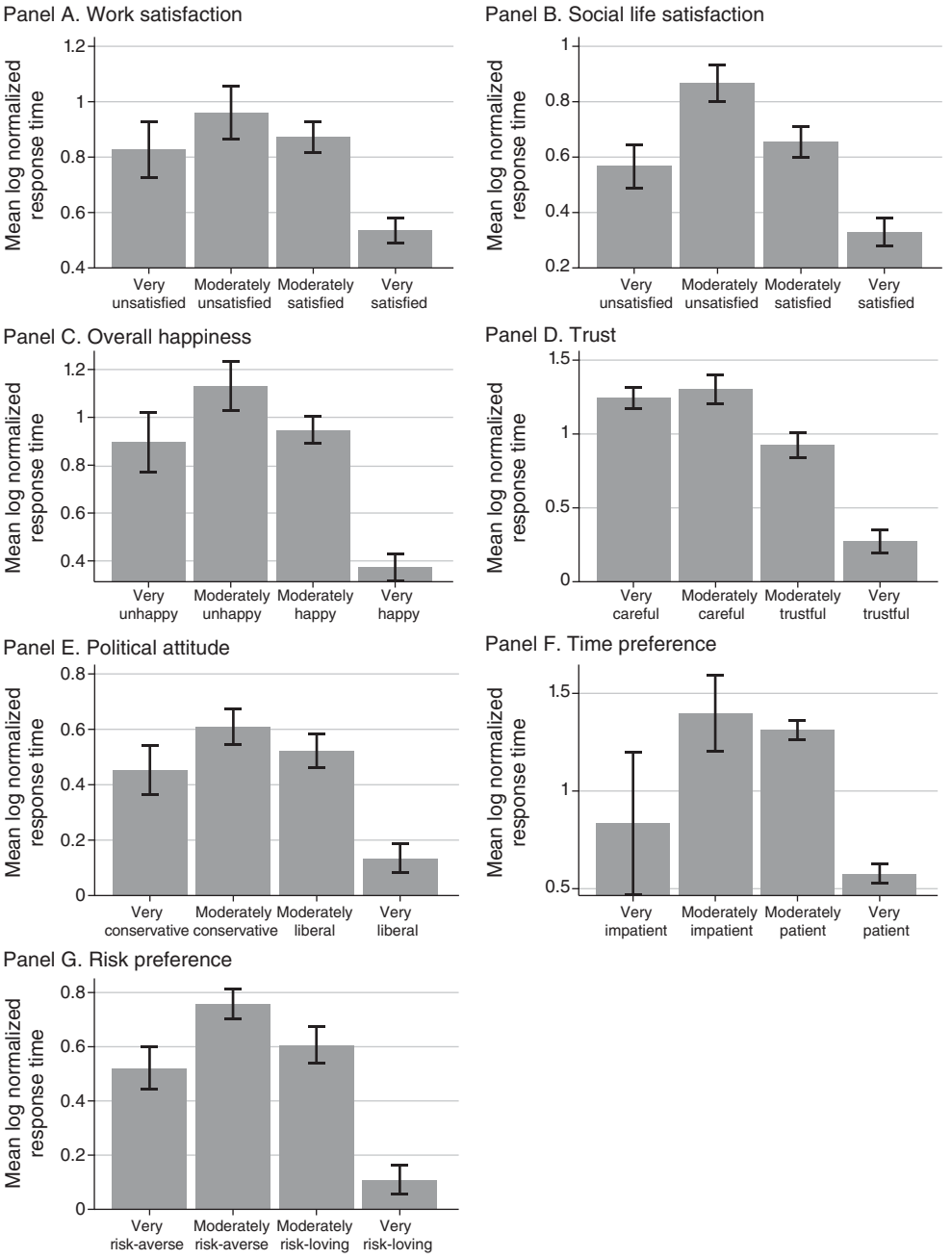


Figure 4. Chronometric Effect by Question in the Trinary Survey

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

initial and the follow-up question, because a larger happiness implies being more distant from the reporting threshold in both stages. By contrast, within the group of subjects who first responded "rather happy" but then refined their answer to "moderately happy," the correlation should be negative, because a larger happiness means being closer to the reporting threshold in the follow-up stage. We did not find such differentiated patterns in our data. As Tables A2–A5 in online Appendix A show, response times are always positively correlated across stages regardless of which follow-up response we focus on, even when controlling for individual fixed effects, for both the normalized and the raw response time data. One explanation for the absence of chronometric effects in the follow-up questions is that subjects already made up their mind about the issue, e.g., about how happy they are, when answering the initial question, and they do not have to think carefully again when answering the follow-up question.

### D. *Analysis of Binary Survey*

Having verified the key premise of our approach – the chronometric effect – we now apply our response time techniques to the binary survey. We divide the sample into sociodemographic groups and, for each substantive question, make pairwise comparisons between the groups to test the null hypothesis of first-order stochastic dominance of the latent distributions. We do this separately for each sociodemographic characteristic, e.g. we compare the happiness between females and males, and between the young and the middle-aged. Finer divisions of the sample can of course be made, but since our focus here is not on a causal interpretation of the results, we prefer keeping the number of pairwise comparisons low. The idea is that, when we reject the null hypothesis of first-order stochastic dominance, then the estimation results using conventional models are qualitatively not robust.

Table 4 reports estimates from a conventional ordered probit model, for all our combinations of sociodemographic groups and substantive questions. Each cell corresponds to a regression of the response to the question in the column on a dummy for membership to the group in the row. The ordered probit coefficients are reported along with their robust standard errors. For example, from row six in column one we learn that married subjects are significantly more satisfied with their work than unmarried subjects ($p = 0.000$).

We now test conditions (i) and (ii) of Proposition 3. We normalize each subject's log response time by subtracting the log response time in the marital status question. We can then construct the empirical cumulative distribution functions of these log-normalized response times and multiply them by the respective response fractions to obtain the empirical counterparts of $r_j^i \hat{F}_j^i(t)$. Under the null hypothesis of first-order stochastic dominance, Proposition 3 predicts that the empirical $r_A^0 \hat{F}_A^0(t) - r_B^0 \hat{F}_B^0(t)$ should be below zero and the empirical $r_A^1 \hat{F}_A^1(t) - r_B^1 \hat{F}_B^1(t)$ should be above zero for all $t$.

Naturally, noise affects these conditions when applied to empirical distributions. To make statistical inference, we draw upon a test for stochastic dominance proposed by Barrett and Donald (2003), which uses a supremum-type statistic from the original sample and computes critical values from bootstrap samples. For condition (i) in Proposition 3, we compute the statistic $S = \max_t \left\{ r_A^0 \hat{F}_A^0(t) - r_B^0 \hat{F}_B^0(t) \right\}$

Table 4—Ordered Probit Analysis of the Binary Survey

| | Work satisfac. | Social satisfac. | Overall happiness | Trust | Liberalism | Patience | Risk-taking |
|---|---|---|---|---|---|---|---|
| 0: female | 0.009 | 0.040 | 0.047 | 0.132 | −0.078 | −0.055 | 0.328 |
| 1: male | (0.0471) | (0.0441) | (0.0466) | (0.0410) | (0.0417) | (0.0539) | (0.0422) |
| 0: young | 0.181 | 0.030 | 0.168 | 0.051 | −0.174 | −0.106 | −0.193 |
| 1: middle-age | (0.0519) | (0.0479) | (0.0512) | (0.0443) | (0.0450) | (0.0576) | (0.0454) |
| 0: middle-age | −0.004 | −0.080 | 0.038 | 0.006 | −0.075 | 0.035 | −0.320 |
| 1: old | (0.1009) | (0.0902) | (0.1007) | (0.0846) | (0.0849) | (0.1093) | (0.0848) |
| 0: none | −0.127 | −0.460 | −0.238 | −0.302 | −0.310 | 0.464 | −0.194 |
| 1: high-school | (0.2925) | (0.3090) | (0.3092) | (0.2885) | (0.2991) | (0.3114) | (0.2856) |
| 0: high-school | 0.782 | 0.523 | 0.502 | 0.664 | 0.114 | 0.086 | 0.522 |
| 1: college | (0.0572) | (0.0557) | (0.0577) | (0.0570) | (0.0549) | (0.0698) | (0.0548) |
| 0: unmarried | 0.930 | 0.796 | 0.808 | 0.639 | −0.131 | 0.257 | 0.512 |
| 1: married | (0.0495) | (0.0460) | (0.0485) | (0.0439) | (0.0438) | (0.0550) | (0.0437) |
| 0: no kids | 0.835 | 0.742 | 0.650 | 0.565 | −0.040 | 0.199 | 0.582 |
| 1: kids | (0.0491) | (0.0455) | (0.0478) | (0.0427) | (0.0428) | (0.0545) | (0.0431) |
| 0: poor | 0.740 | 0.477 | 0.556 | 0.471 | 0.009 | 0.199 | 0.287 |
| 1: middle-income | (0.0567) | (0.0532) | (0.0554) | (0.0507) | (0.0512) | (0.0628) | (0.0512) |
| 0: middle-income | −0.112 | −0.061 | 0.034 | −0.229 | −0.114 | 0.216 | −0.159 |
| 1: rich | (0.0607) | (0.0544) | (0.0595) | (0.0491) | (0.0497) | (0.0692) | (0.0504) |

*Notes:* Each cell corresponds to a regression of the question in the column on a dummy for membership to the group in the row. Coefficients are reported along with their robust standard errors in parentheses.

in the original sample. Using equation (11) in Barrett and Donald (2003), we then compute $S^* = \max_t \left\{ \left[ r_A^{0*} \hat{F}_A^{0*}(t) - r_B^{0*} \hat{F}_B^{0*}(t) \right] - \left[ r_A^0 \hat{F}_A^0(t) - r_B^0 \hat{F}_B^0(t) \right] \right\}$ in each of 1,000 bootstrap samples (marked by $*$), which are generated by sampling from the original data with replacement keeping the number of subjects in each group fixed. The $p$-value associated with condition (i) counts how often $S^*$ exceeds $S$. The analogous procedure is used to obtain a $p$-value for condition (ii) in Proposition 3. To account for the multiple hypothesis nature of the problem, our overall $p$-value for the null hypothesis of first-order stochastic dominance of the latent distributions compares the minimum of the two individual $p$-values to the distribution of this minimum in the bootstrap samples, i.e., it counts how often the smaller of the two $p$-values in the bootstrap samples (using where the bootstrapped statistics fall within the entire bootstrapped distributions) is below the smaller of the two $p$-values in the original sample.[13]

To illustrate our findings, consider again the question about work satisfaction ("How satisfied are you with the work you do?"), and compare the groups of subjects who are married (group $A$) and who are unmarried (group $B$). The solid curve in panel A of Figure 5 plots $r_A^1 \hat{F}_A^1(t) - r_B^1 \hat{F}_B^1(t)$, the cumulative difference in response

[13] See e.g. Romano and Wolf (2005a, b, 2016). We use a single-step rather than a step-down procedure because we are ultimately interested in the joint hypothesis of conditions (i) and (ii) in Proposition 3, not in the two hypotheses separately. Our paper is accompanied by Stata ado-files which implement the test, for surveys with an arbitrary number of response categories.

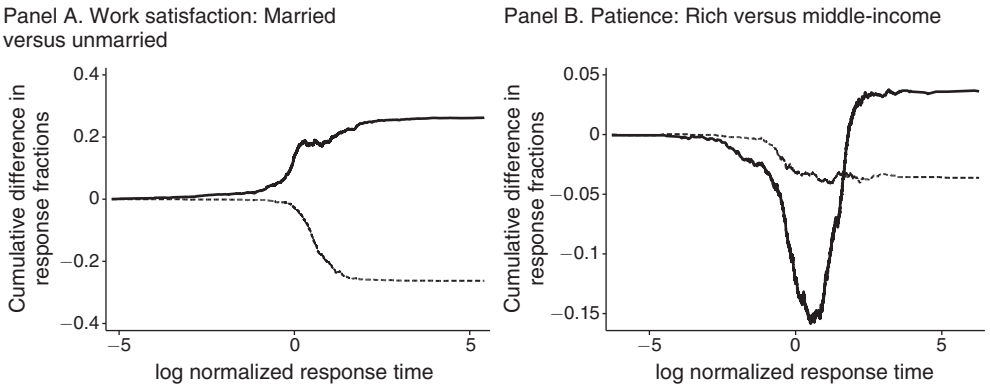


FIGURE 5. EXAMPLES OF EMPIRICAL FIRST-ORDER STOCHASTIC DOMINANCE (FOSD) CONDITIONS

*Notes:* Figure panels refer to different questions in the binary survey. The first sociodemographic group described in the caption is coded as group $A$, the second as group $B$. Solid curves depict $r_A^1 \hat{F}_A^1(t) - r_B^1 \hat{F}_B^1(t)$ and dashed curves depict $r_A^0 \hat{F}_A^0(t) - r_B^0 \hat{F}_B^0(t)$.

fractions between these two groups for the response category "rather satisfied," with $t$ varying on the $x$-axis. This curve always lies above zero, meaning that the fraction of married subjects who responded to be satisfied with their job is always larger than that of the unmarried subjects, even when restricting attention to responses which took place before any time $t$. Similarly, the dashed curve plots the cumulative difference for the answer category "rather unsatisfied," $r_A^0 \hat{F}_A^0(t) - r_B^0 \hat{F}_B^0(t)$. It always lies below zero, because the fraction of subjects who responded to be unsatisfied with their job is always smaller for the married than for the unmarried, again for all response times. These inequalities hold perfectly in the data. Our test cannot reject the null hypothesis of first-order stochastic dominance of the latent distributions ($p = 1.000$).

As a second example, consider the relationship between patience ("How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?") and income. We think of the latent variable for this question as being a time preference parameter, such that higher values capture greater patience. From Table 4 we learn that high income subjects ($\geq$ \$70,000) are significantly more patient than middle-income subjects (\$40,000–\$69,999) ($p = 0.002$). The solid curve in panel B of Figure 5 shows that many middle-income subjects very quickly responded that they are willing to give up immediate rewards for a future benefit, even though overall a higher fraction of participants in the high income group responded in this category. Hence, the response times reveal that some middle-income subjects are particularly patient. Our test clearly rejects the null hypothesis of first-order stochastic dominance of the latent distributions ($p = 0.000$). The estimated relationship between income and patience is qualitatively not robust.

Figure 6 summarizes all our test results for the binary survey. Each circle represents a comparison between two sociodemographic groups. The circles display the $p$-value of our test for first-order stochastic dominance of the latent distributions on the $x$-axis, for the different substantive survey questions stacked on the $y$-axis. The figure shows reference lines at 5 percent and 10 percent as an orientation.

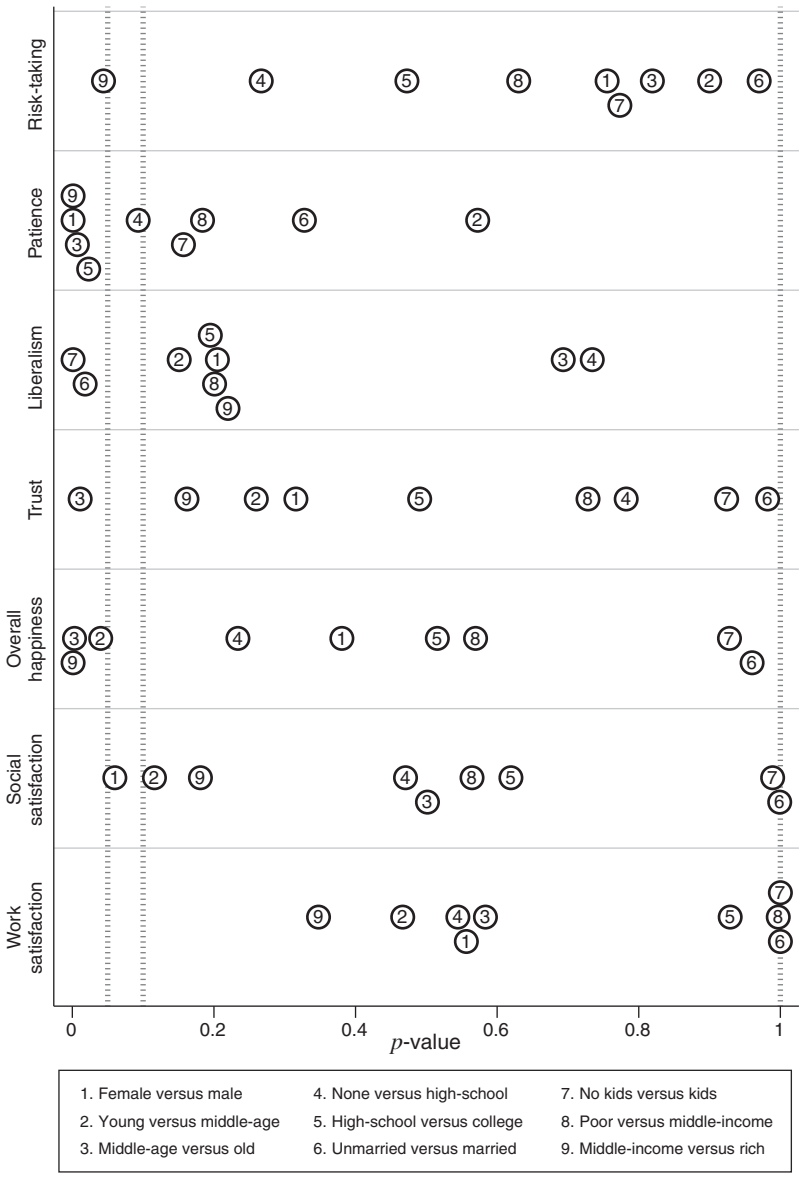


FIGURE 6. FOSD TEST IN BINARY SURVEY

Figure 6 documents interesting patterns. For the time preference question, a substantial fraction of the circles is concentrated at low $p$-values. We reject the null hypothesis of first-order stochastic dominance in four out of nine cases at 5 percent. We conclude that time preference parameters are not likely to follow the distributions assumed in traditional ordered-response models, and the estimated coefficients must therefore be interpreted with caution.[14] This is in contrast to the risk preference question, where $p$-values are generally much larger and only one group

[14] Two of these comparisons have a value $p = 0.000$. The rejections are thus robust even if we explicitly added a conservative multiple hypothesis correction also for our interpretation of the entire Figure 6.

comparison has a $p$-value below 5 percent. The difference between the questions about risk preferences and time preferences may be of interest for the literature that explores the relation between preferences in the risk and time domains (Andreoni and Sprenger 2012). For the two satisfaction questions concerning work and social life and for the trust question, we are also unable to reject the first-order stochastic dominance assumption in almost all comparisons. The questions about overall life happiness and political attitude are somewhere in between, with a clear rejection of the null hypothesis in some cases (two cases $p = 0.000$) and high $p$-values in others.

Let us return to the relation between income and happiness discussed in the Introduction. For the ordered probit model, Table 4 shows that higher income is associated with significantly higher overall happiness when we move from low income ($<$ \$40,000) to middle income ($p = 0.000$). The effect is still positive but much smaller and not significant when we move from middle income to high income ($p = 0.573$). In Figure 6, the comparison between low and middle income is depicted as circle 8. We can see that the null hypothesis of first-order stochastic dominance of the happiness distributions cannot be rejected for this group comparison ($p = 0.569$). The comparison between middle and high income is depicted as circle 9. Here, our test clearly rejects the null hypothesis of first-order stochastic dominance ($p = 0.000$). Taken together, these findings are consistent with the results of the ordered probit model, according to which there is a positive association between income and happiness (within country at a fixed point in time) for small but not for large incomes. With the appropriate data, our techniques could be used to examine the income-happiness relation also across countries or over time.

In general, there is not a one-to-one relationship between the significance of the ordered probit coefficient and our FOSD test. The above discussed relation between patience and income shows that the estimated coefficient can be significant while the FOSD test rejects the null. The relation between work satisfaction and gender is an example where the estimated coefficient is insignificant ($p = 0.847$) while the FOSD test does not reject the null ($p = 0.556$). Altogether, however, our results are broadly consistent with the qualitative validity of the ordered probit estimates. Figure 7 is a scatter plot of the $p$-values of the ordered probit coefficient and our FOSD test. These values are negatively correlated ($\rho = -0.3585$), indicating that more significant estimation results tend to go along with reduced ability to reject the assumption of first-order stochastic dominance of the latent distributions.

### E. *Analysis of Trinary Survey*

We can repeat the analysis for the survey with three response categories. Table 5 reports the estimated coefficients of the ordered probit model using the trinary survey data. Comparing the estimation results of the two survey versions, we sometimes obtain different parameter signs (8 out of 63 times), but then at least one of the two different estimates is always insignificant. Each of the two survey versions is sometimes "more significant" than the other. Overall, the two versions of the survey seem to generate comparable results based on ordered probit estimation.

We now need to test also the additional condition (iii) of Proposition 3. This condition is just another instance of a first-order stochastic dominance

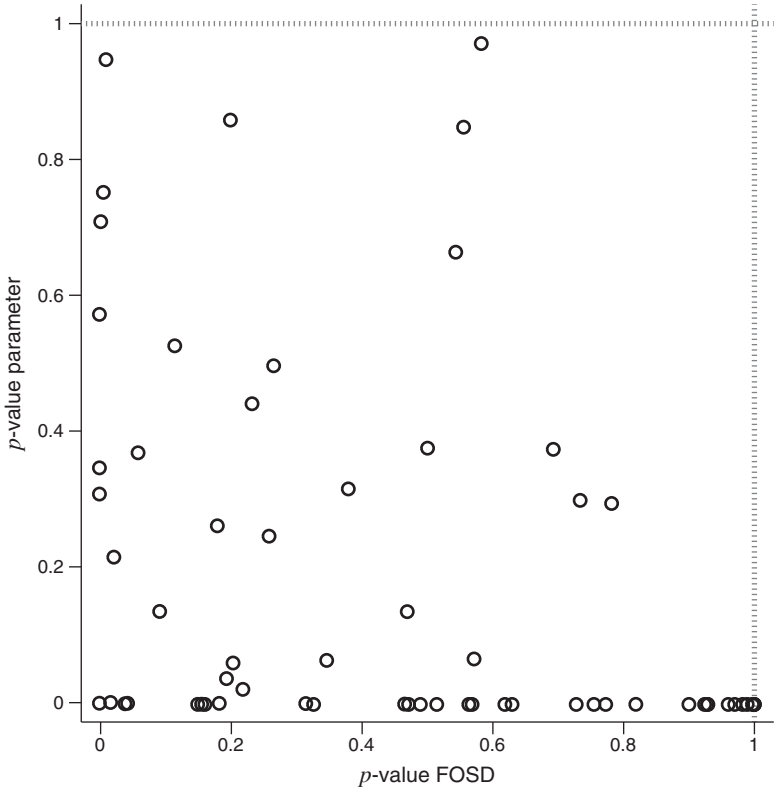


FIGURE 7. *p*-VALUES FOR FOSD TEST AND ORDERED PROBIT ESTIMATION IN BINARY SURVEY

requirement, this time on the distribution of responses in the two groups, so we again apply the procedure of Barrett and Donald (2003) and calculate the statistic $S = \max_{k=0,\ldots,n}\left\{\sum_{i=0}^{k} r_A^i - \sum_{i=0}^{k} r_B^i\right\}$ in the original sample and $S^* = \max_{k=0,\ldots,n}\left\{\left(\sum_{i=0}^{k} r_A^{i*} - \sum_{i=0}^{k} r_B^{i*}\right) - \left(\sum_{i=0}^{k} r_A^i - \sum_{i=0}^{k} r_B^i\right)\right\}$ in each bootstrap sample.[15] We use the same multiple hypothesis correction described earlier.

Figure 8 summarizes the results of our FOSD tests in the trinary survey and Figure 9 plots the *p*-values of the ordered probit coefficient against those of our FOSD test. The correlation between these two *p*-values is negative just like in the binary survey ($\rho = -0.5342$). One difference between the binary and the trinary survey is that, while the *p*-values of our FOSD tests are positively correlated between the binary and the trinary survey ($\rho = 0.6509$), those in the trinary survey tend to be weakly larger (in 47 out of 63 cases). Random assignment of subjects into treatments should imply that the latent distributions are the same in both versions of the survey, suggesting that our tests have a higher power in the binary case. We expect the combination of binary surveys and response time analysis to have great potential in future research.

[15] The proof of Proposition 3 shows that inequality (iii) holds for categories $k = 0, n$ as well, so that this testing procedure is appropriate.

Table 5—Ordered Probit Analysis of the Trinary Survey

| | Work satisfac. | Social satisfac. | Overall happiness | Trust | Liberalism | Patience | Risk-taking |
|---|---|---|---|---|---|---|---|
| 0: female | −0.007 | 0.093 | −0.001 | 0.117 | −0.061 | −0.034 | 0.232 |
| 1: male | (0.0397) | (0.0386) | (0.0394) | (0.0366) | (0.0368) | (0.0424) | (0.0376) |
| 0: young | 0.090 | 0.037 | 0.001 | −0.039 | −0.101 | −0.021 | −0.234 |
| 1: middle-age | (0.0434) | (0.0421) | (0.0431) | (0.0399) | (0.0401) | (0.0463) | (0.0413) |
| 0: middle-age | −0.106 | −0.099 | 0.112 | 0.164 | −0.009 | −0.103 | −0.174 |
| 1: old | (0.0777) | (0.0829) | (0.0838) | (0.0791) | (0.0811) | (0.0906) | (0.0795) |
| 0: none | −0.360 | −0.462 | 0.149 | −0.021 | −0.108 | 0.643 | −0.050 |
| 1: high-school | (0.2992) | (0.2652) | (0.2882) | (0.2810) | (0.2905) | (0.2221) | (0.2394) |
| 0: high-school | 0.625 | 0.524 | 0.484 | 0.547 | 0.062 | 0.194 | 0.475 |
| 1: college | (0.0499) | (0.0525) | (0.0521) | (0.0501) | (0.0482) | (0.0568) | (0.0498) |
| 0: unmarried | 0.718 | 0.669 | 0.626 | 0.538 | −0.243 | 0.150 | 0.486 |
| 1: married | (0.0413) | (0.0411) | (0.0412) | (0.0391) | (0.0386) | (0.0443) | (0.0398) |
| 0: no kids | 0.662 | 0.546 | 0.512 | 0.436 | −0.217 | 0.180 | 0.601 |
| 1: kids | (0.0407) | (0.0399) | (0.0404) | (0.0379) | (0.0382) | (0.0435) | (0.0392) |
| 0: poor | 0.503 | 0.348 | 0.375 | 0.405 | −0.146 | 0.202 | 0.354 |
| 1: middle-income | (0.0475) | (0.0471) | (0.0474) | (0.0450) | (0.0449) | (0.0507) | (0.0461) |
| 0: middle-income | 0.015 | 0.013 | 0.162 | −0.124 | −0.040 | 0.091 | −0.187 |
| 1: rich | (0.0490) | (0.0471) | (0.0492) | (0.0437) | (0.0446) | (0.0525) | (0.0453) |

*Notes:* Each cell corresponds to a regression of the question in the column on a dummy for membership to the group in the row. Coefficients are reported along with their robust standard errors in parentheses.

## III. Related Literature

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

Some recent papers have provided responses to the 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

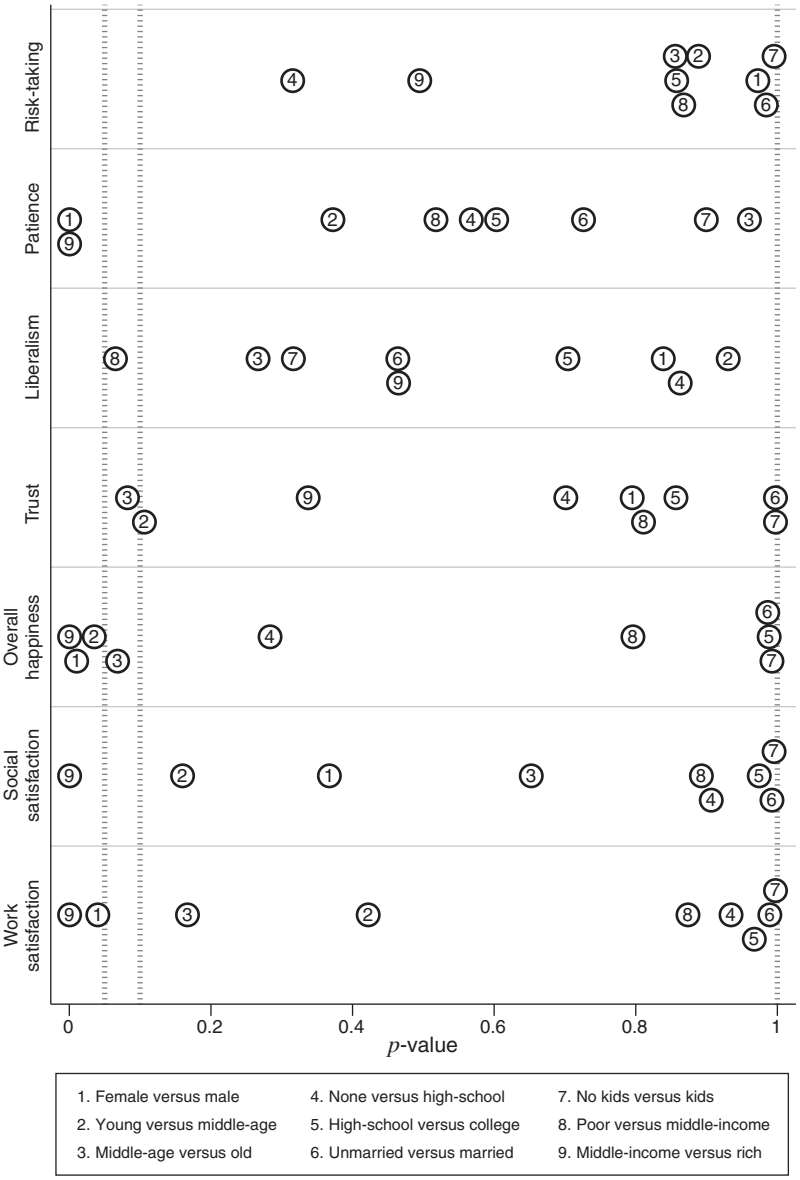


FIGURE 8. FOSD TEST IN TRINARY SURVEY

(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 testing the necessary distributional assumptions with 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

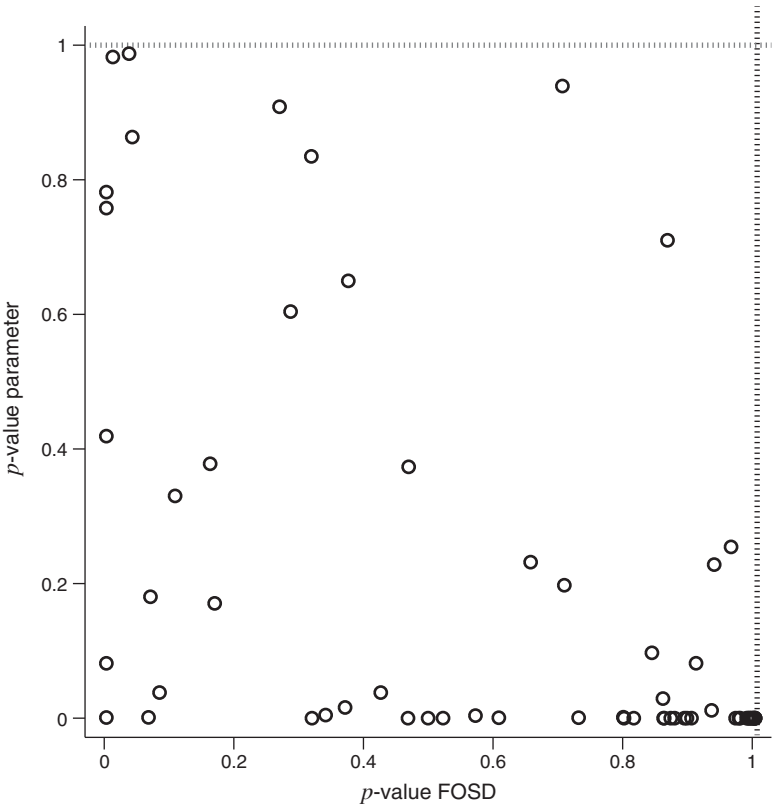


Figure 9. *p*-values for FOSD Test and Ordered Probit Estimation in Trinary Survey

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

More generally, there is a growing interest among economists to explore what can be learned from response times. For instance, Rubinstein (2007, 2016) proposes a typology of choices and players in strategic games based on response times. Achtziger and Alós-Ferrer (2014) show that response time can measure the extent to which an agent's decision-making process under uncertainty is consistent with the rational paradigm of Bayesian belief-updating. The literature has also suggested that response time data can be used to reveal how decision-makers allocate their limited attention between different problems (Avoyan and Schotter 2020), to facilitate social learning by serving as an observable signal of agents' private information (Frydman and Krajbich 2022), to alleviate misspecification bias in the estimation of structural preference parameters (Webb 2019), and to improve out-of-sample predictions of behavior (Clithero 2018a; Alós-Ferrer, Fehr, and Netzer 2021), among several others.

## IV. Conclusion

In this paper, we have shown that response time data can address an identification problem of ordered response models. Since survey data are typically discrete and ordinal, while comparing averages across groups requires continuous and cardinal information, the traditional ordered response models rely on assumptions about the distribution of a latent variable. Their results can change drastically when this distribution is transformed. We have shown, both theoretically and empirically, that response times are a source of information about the distribution of the latent variable. Through the chronometric function, properties of the distribution become observable and distributional assumptions become testable.

We have in mind two ways in which our results can be used in practice. First, surveys are increasingly conducted online, and recording response times is easy and costless in that case. We think that response time data should be collected on par with response data, and their analysis could become a natural part of any investigation. Of course, causal analysis will be an important concern in many applications, which implies that the groups to be compared have to be much finer than in our simple empirical illustration. One could also try to integrate response time data into a multivariate regression analysis. We leave to future research the question how this could be done, but we conjecture that one could attempt to change the outcome variable in the traditional regression analysis from response to response time, or possibly to response weighted by response time to capture the intensity of the response. Second, one could use our techniques in auxiliary studies, with the goal of confirming in a representative sample that the latent variable of interest follows distributions for which traditional ordered response models are appropriate. Once enough evidence of this type has been accumulated, the analyst can proceed as usual and does not have to bother about response time data any more.

## REFERENCES

**Achtziger, Anja, and Carlos Alós-Ferrer.** 2014. "Fast or Rational? A Response-Times Study of Bayesian Updating." *Management Science* 60 (4): 923–38.

**Alós-Ferrer, Carlos, Ernst Fehr, and Nick Netzer.** 2021. "Time Will Tell: Recovering Preferences when Choices Are Noisy." *Journal of Political Economy* 129 (6): 1828–77.

**Andreoni, James, and Charles Sprenger.** 2012. "Risk Preferences Are Not Time Preferences." *American Economic Review* 102 (7): 3357–76.

**Avoyan, Ala, and Andrew Schotter.** 2020. "Attention in Games: An Experimental Study." *European Economic Review* 124: 103403.

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

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

**Boes, Stefan, and Rainer Winkelmann.** 2006. "Ordered Response Models." *Journal of the German Statistical Society* 90 (1): 167–81.

**Bond, Timothy N., and Kevin Lang.** 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): 1468–79.

**Bond, Timothy N., and Kevin Lang.** 2019. "The Sad Truth About Happiness Scales." *Journal of Political Economy* 127 (4): 1629–40.

**Boulier, Bryan L., and Robert S. Goldfarb.** 1998. "On the Use and Nonuse of Surveys in Economics." *Journal of Economic Methodology* 5 (1): 1–21.

**Chabris, Christopher F., Carrie L. Morris, Dmitry Taubinsky, David Laibson, and Jonathon P. Schuldt.** 2009. "The Allocation of Time in Decision-Making." *Journal of the European Economic Association* 7 (2-3): 628–37.

**Chen, Le-Yu, Ekaterina Oparina, Nattavudh Powdthavee, and Sorawoot Srisuma.** 2019. "Have Econometric Analyses of Happiness Data Been Futile? A Simple Truth about Happiness Scales." IZA Discussion Paper 12152.

**Clithero, John A.** 2018a. "Improving Out-of-Sample Predictions Using Response Times and a Model of the Decision Process." *Journal of Economic Behavior and Organization* 148: 344–75.

**Clithero, John A.** 2018b. "Response Times in Economics: Looking Through the Lens of Sequential Sampling Models." *Journal of Economic Psychology* 69: 61–86.

**Cunha, Flavio, James J. Heckman, and Salvador Navarro.** 2007. "The Identification and Economic Content of Ordered Choice Models with Stochastic Thresholds." *International Economic Review* 48 (4): 1273–1309.

**Davis, James A., and Tom W. Smith.** 1991. *The NORC General Social Survey: A User's Guide.* Newbury Park, CA: Sage Publications.

**DellaVigna, Stefano, and Devin Pope.** 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies* 85 (2): 1029–69.

**Easterlin, Richard A.** 1974. "Does Economic Growth Improve the Human Lot? Some Empirical Evidence." In *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, edited by Paul A. David and Melvin W. Reder, 89–125. New York: Academic Press.

**Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde.** 2018. "Global Evidence on Economic Preferences." *Quarterly Journal of Economics* 133 (4): 1645–92.

**Frydman, Cary, and Ian Krajbich.** 2022. "Using Response Times to infer Others' Beliefs: An Application to Information Cascades." *Management Science* 68 (4): 2970–86.

**Goodman, Joseph K., and Gabriele Paolacci.** 2017. "Crowdsourcing Consumer Research." *Journal of Consumer Research* 44 (1): 196–210.

**Hanock, Giora, and Haim Levy.** 1969. "The Efficiency Analysis of Choices Involving Risk." *Review of Economic Studies* 36 (107): 335–46.

**Hess, Stephane, and Amanda Strathopoulos.** 2013. "Linking Response Quality to Survey Engagement: A Combined Random Scale and Latent Variable Approach." *Journal of Choice Modelling* 7: 1–12.

**Kaiser, Caspar, and Andrew J. Oswald.** 2022. "Inequality, Well-Being, and the Problem of the Unknown Reporting Function." *Proceedings of the National Academy of Sciences* 119 (50): e2217750119.

**Kaiser, Caspar, and Maarten C.M. Vendrik.** 2020. "How Threatening are Transformations of Reported Happiness to Subjective Wellbeing Research?" IZA Discussion Paper 13905.

**Kaplan, David M., and Longhao Zhuo.** 2019. "Comparing Latent Inequality with Ordinal Data." Unpublished.

**Kellogg, W. N.** 1931. "The Time of Judgment in Psychometric Measures." *American Journal of Psychology* 43 (1): 65–86.

**Konovalov, Arkady, and Ian Krajbich.** 2019. "Revealed Strength of Preference: Inference from Response Times." *Judgment & Decision Making* 14 (4): 381–94.

**Krajbich, Ian, Carrie Armel, and Antonio Rangel.** 2010. "Visual Fixations and the Computation and Comparison of Value in Simple Choice." *Nature Neuroscience* 13 (10): 1292–98.

**Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva.** 2015. "How Elastic are Preferences for Redistribution? Evidence from Randomized Survey Experiments." *American Economic Review* 105 (4): 1478–1508.

**Likert, Rensis.** 1932. "A Technique for the Measurement of Attitudes." *Archives of Psychology* 22 (140): 5–55.

**Liu, Shuo, and Nick Netzer.** 2020. "Happy Times: Identification from Ordered Response Data." University of Zurich Department of Economics Working Paper 371.

**Liu, Shuo, and Nick Netzer.** 2023."Data and Code for: Happy Times: Measuring Happiness Using Response Times." American Economic Association [publisher], Inter-University Consortium for Political and Social Research [distributor]. http://doi.org/103886/E193215V1.

**Moffatt, Peter G.** 2005. "Stochastic Choice and the Allocation of Cognitive Effort." *Experimental Economics* 8 (4): 369–88.

**Moyer, Robert S., and Richard H. Bayer.** 1976. "Mental Comparison and the Symbolic Distance Effect." *Cognitive Psychology* 8 (2): 228–46.

**Oswald, Andrew J.** 2008. "On the Curvature of the Reporting Function from Objective Reality to Subjective Feelings." *Economics Letters* 100 (3): 369–72.

**Palmer, John, Alexander C. Huk, and Michael N. Shadlen.** 2005. "The Effect of Stimulus Strength on the Speed and Accuracy of a Perceptual Decision." *Journal of Vision* 5 (5): 376–404.

**Paolacci, Gabriele, and Jesse Chandler.** 2014. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." *Current Directions in Psychological Science* 23 (3): 184–88.

**Romano, Joseph P., and Michael Wolf.** 2005a. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association* 100 (469): 94–108.

**Romano, Joseph P., and Michael Wolf.** 2005b. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73 (4): 1237–82.

**Romano, Joseph P., and Michael Wolf.** 2016. "Efficient Computation of Adjusted P-Values for Resampling-Based Stepdown Multiple Testing." *Statistics and Probability Letters* 113: 38–40.

**Rossi, Peter H., James D. Wright, and Andy B. Anderson.** 1983. "Sample Surveys: History, Current Practice, and Future Prospects." In *Handbook of Survey Research*, edited by Peter H. Rossi, James D. Wright and Andy B. Anderson, 1–20. New York: Academic Press.

**Rubinstein, Ariel.** 2007. "Instinctive and Cognitive Reasoning: A Study of Response Times." *Economic Journal* 117 (523): 1243–59.

**Rubinstein, Ariel.** 2016. "A Typology of Players: Between Instinctive and Contemplative." *Quarterly Journal of Economics* 131 (2): 859–90.

**Schroeder, Carsten, and Shlomo Yitzhaki.** 2017. "Revisiting the Evidence for Cardinal Treatment of Ordinal Variables." *European Economic Review* 92: 337–58.

**Stevenson, Betsey, and Justin Wolfers.** 2008. "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox." *Brookings Papers on Economic Activity* 38 (1): 1–87.

**Studer, Raphael, and Rainer Winkelmann.** 2014. "Reported Happiness, Fast and Slow." *Social Indicators Research* 117 (3): 1055–67.

**Tannenbaum, David, Alain Cohn, Christian Zünd, and Michel A. Maréchal.** Forthcoming. "What do Cross-Country Surveys Tell Us About Social Capital?" *Review of Economics and Statistics***.**

**Webb, Ryan.** 2019. "The (Neural) Dynamics of Stochastic Choice." *Management Science* 65 (1): 230–55.