

The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences

Armin Falk,^{a,b,*} Anke Becker,^c Thomas Dohmen,^{a,d} David Huffman,^e Uwe Sunde^f

^aUniversity of Bonn, D-53113, Bonn, Germany; ^bBehavior and Inequality Research Institute GmbH, 53113 Bonn, Germany; ^cHarvard Business School, Boston, Massachusetts 02163; ^dMaastricht University, School of Business and Economics, 6200 MD, Maastricht, Netherlands;

^eUniversity of Pittsburgh, Pittsburgh, Pennsylvania 15260; ^fUniversity of Munich, Munich D-80539, Germany

*Corresponding author

Contact: armin.falk@briq-institute.org,  <https://orcid.org/0000-0002-7284-3002> (AF); abecker@hbs.edu,

 <https://orcid.org/0000-0003-0838-7644> (AB); t.dohmen@uni-bonn.de,  <https://orcid.org/0000-0002-9321-0319> (TD); huffmand@pitt.edu,

 <https://orcid.org/0000-0002-4714-5560> (DH); uwe.sunde@econ.lmu.de,  <https://orcid.org/0000-0002-2110-7822> (US)

Received: January 11, 2021

Revised: July 2, 2021

Accepted: August 22, 2021


Published Online in Articles in Advance:
October 31, 2022

<https://doi.org/10.1287/mnsc.2022.4455>

Copyright: © 2022 The Author(s)

Abstract. Incentivized choice experiments are a key approach to measuring preferences in economics but are also costly. Survey measures are a low-cost alternative but can suffer from additional forms of measurement error due to their hypothetical nature. This paper seeks to leverage the strengths of both approaches by proposing a new survey module on risk aversion, time discounting, trust, altruism, positive and negative reciprocity, in which survey items are selected based on ability to predict choices in corresponding, incentivized experiments. The methodology and results provided in the paper can also potentially provide a model for researchers who have specific requirements and want to design their own modules.

History: Accepted by Yan Chen, behavioral economics and decision analysis.

 **Open Access Statement:** This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. You are free to download this work and share with others for any purpose, except commercially, if you distribute your contributions under the same license as the original, and you must attribute this work as “*Management Science*. Copyright © 2022 The Author(s). <https://doi.org/10.1287/mnsc.2022.4455>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by-nc-sa/4.0/>.”

Funding: The project received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7-2007-2013) [Grant 209214]. A. Falk and T. Dohmen acknowledge funding from the Deutsche Forschungsgemeinschaft (German Research Foundation) [Grant CRC TR 224 (Project A01)] and Germany’s Excellence Strategy [Grant EXC 2126/1-390838866].

Supplemental Material: Data and the online appendices are available at <https://doi.org/10.1287/mnsc.2022.4455>.

Keywords: survey validation • experiment • preference measurement

1. Introduction

In economic models, preferences are traits that drive decision making. Certain types of preferences—regarding risk, time, and social interactions—are central in economic theory because they affect such a broad range of economic decisions. Having measures of these fundamental preferences is valuable because of the opportunity to better explain economic behavior.

Incentivized choice experiments have emerged as a key approach to measuring preferences. Experiments seek to hold constant the decision environment across individuals, so that differences in choices reveal different preferences. The use of real incentives can help address measurement issues that arise with alternative approaches, specifically survey measures, due to their hypothetical nature. For example, the lack of incentives could lead survey measures to suffer from

measurement error due to inattention.¹ One limitation of incentivized experiments, however, is they are costly in terms of money and also time. Thus, whereas a researcher might prefer to conduct incentivized experiments, it may not always be feasible to do so.

This paper seeks to develop survey modules that leverage the strengths of both experimental and survey approaches. We propose a survey module on risk aversion, time discounting, trust, altruism, positive reciprocity, and negative reciprocity, which is parsimonious and low cost to implement, but where the survey items are selected based on ability to predict choices in corresponding, incentivized choice experiments. The paper describes in detail the methodology used for item selection. The main idea is that there are many different wordings and formats one could choose for survey measures. These may differ in their

accuracy in predicting choices in experiments, for example, because of varying degrees of measurement error, or because what they measure is more or less tightly linked to the determinants of experimental choices.² Because accuracy of survey measures is difficult to judge a priori, we conduct incentivized choice experiments, and from a large set of candidate survey items, identify those that do best in terms of predicting incentivized choices.

Our survey module is suitable for a wide range of applications and settings. One important class of applications is within firms and organizations. Preference measures are potentially valuable to managers due to the role of preferences in determining how employees behave. For example, economic theory predicts a role of risk preference in determining how employees sort into incentive schemes, and how managers make investment decisions; time preference is relevant for how employees respond to threat of being fired in the future and other dynamic incentives; social preferences can shape how employees work in teams. Survey measures of preferences can be easily introduced into the flow of workplace assessments or screenings in the same way as psychometric tools that are already used as part of management practices.³ Alternative methods to measure preferences, such as incentivized choice experiments, are more costly and difficult to implement in such field settings.⁴ Survey measures are also well suited for applications that involve measuring preferences on a large scale, either across a large population of workers in a multinational organization, or across representative population samples in a cross-country survey.⁵ Moreover, it is useful to have access to valid survey measures in applications, ranging from laboratory experiments to collecting observational data, in which researchers or practitioners require preferences measures, but need to allocate the bulk of their time and financial resources to other aspects of the study. The simplicity of administering survey measures also has advantages in the context of certain types of research settings in which logistics are particularly complicated, for example, field experiments.

For our survey item selection exercise, we used a sample of German university students. For each participant, we elicited each preference using both incentivized experimental measures *and* using a comprehensive set of survey items. We conducted multiple experiments for preferences, to reduce measurement error in the dependent variable, and induced a time lag of one week between experiments and corresponding candidate survey measures to minimize spurious correlations arising from consistency bias. When selecting survey items, we considered all possible linear combinations of survey items intended to measure a particular preference, and identified the combination that best predicted

behavior in the respective experimental preference elicitation task. Specifically, we used standard model selection criteria to guide our choice, and, in addition, took into account the risk of overfitting by evaluating out-of-sample predictive power, or alternatively by conducting cross-validation procedures.

We present the module selected through this procedure, which turns out to involve two survey items for the elicitation of each preference. The preference module is symmetric, in that most preferences are measured with one quantitative and one qualitative item. These quantitative questions are typically the single best measure for explaining behavior in the corresponding experiment. The qualitative measures are self-assessments, but are relatively simple and direct, and do contribute additional explanatory power regarding behavior in incentivized choice experiments. Responses to the survey module provide an ordinal measure of preferences. This may be sufficient for many applications, but like with choices in incentivized experiments, it is also possible to transform the predicted choices from the survey measures into cardinal preference parameters using additional assumptions, for example, about functional form of utility.

We provide information on the properties of the survey module in terms of predictive power for choices in experiments. We show that the module does sacrifice some predictive accuracy compared with more costly types of predictors (e.g., incentivized experiments as predictors), but at the benefit of lower cost. We provide information on test-retest correlations for the survey items, which show that they contain measurement error and thus suffer from some attenuation bias when it comes to predicting choices in experiments. One implication is that predictive power of the survey module can be improved if a researcher has the opportunity to implement repeated measurements of the survey module.

Even though our proposed survey modules were selected using German university students, there are conceptual and empirical reasons to expect that they will still be useful proxies for incentivized experiments in a diverse set of nonstudent populations. What is needed is that the types of survey questions that best predict choices in experiments by German students be similar to the types of questions that best predict such choices in a given other population.⁶ In a final section, we discuss findings from other studies, which show that the types of survey measures included in our modules do in fact work well for predicting choices in incentivized experiments, and also predicting relevant life economic outcomes, in nonstudent samples across a wide range of cultures.

Although the proposed survey module was preferred in our validation exercise, researchers might have specific needs that cause them to prefer single

survey items, or different combinations of survey items. For this reason, in an appendix we also show results on the performance of various individual items, as well as different combinations of items, so that users can select their own module out of this set. It could also be that researchers want to develop new survey modules for themselves, which are optimized to a particular population, or application. In this case, our survey-selection methodology provides a potential model for how researchers might develop such survey modules.

This paper ends by providing one example of how our module can be adapted to serve particular purposes. We explain how we modified our preference module for the implementation in applications where time constraints are particularly severe, such as large-scale, international telephone surveys. We call the resulting model the Global Preferences Survey (GPS) module. The GPS version sacrifices a modest amount of explanatory power, in exchange for being even simpler and more time efficient. This module has subsequently been included in the Gallup World Poll 2012, a survey that was conducted with representative samples using telephone and face-to-face interviews in 76 countries around the globe. The resulting data set is described in Falk et al. (2018).⁷

One benefit of the survey modules proposed in this paper stems from the transparency of the methodology for selecting the measures. For most existing survey measures of economic preferences, the criteria and methodology of how the measures were developed is typically not explicit. Even if there was an *ex ante* optimization process for the measures, this is typically not reported. A few previous survey measures have been validated, in the sense that they were found to be correlated with behavior in experiments, but there was not an optimization process that involved a horse race between different types of survey measures.⁸ The transparent methodology helps make the measures less *ad hoc* from the perspective of potential users, and users will be able to cite the underlying design methodology as a reason for confidence, *ex ante*, in the viability of the measures. Another notable feature of the proposed survey preference modules is that they include proxies for a comprehensive set of preference experiments, measured using a consistent methodology. The modules thus provide a low-cost way to capture a whole bundle of preferences.

The remainder of the paper is organized as follows. Section 2 describes the procedures to elicit preferences in experiments and survey questions. Section 3 explains the methodology for selection of items for the preference module. It presents the preference module measuring each of the six preferences, which performed best in out-of-sample prediction. Section 4 discusses important properties of the preference module,

such as explanatory power and viability in nonstudent and non-German samples. Section 5 gives information needed to construct alternative preference modules. It also provides an example of modifying the preference module for the Global Preferences Survey, an international telephone survey. Section 6 concludes.

2. Design of the Survey Module

In this section, we describe the methodology underlying the design of our survey modules. The design involved implementing incentivized choice experiments, asking the same subjects a battery of survey measures, and then selecting the combinations of survey items that did the best job of predicting choices in the experiments in linear, multivariate regression models. To reduce potential measurement error in the dependent variable, we had subjects participate in more than one experiment for a given preference and averaged over the choice-based preference measures. We designed the validation to limit spurious interdependencies in choices and survey responses by never asking survey questions relating to a particular preference experiment in the same session in which the respective preference elicitation experiment was conducted, that is, surveys and experiments were conducted one week apart. We also restricted the subject pool to subjects who had never participated in an experiment before, to help rule out possible biases in behavior due to experiences gained in previous experiments.

2.1. Procedural Details

Four hundred and nine subjects participated in our study. Subjects were students from the University of Bonn who were recruited using the web-based Online Recruitment System for Economic Experiments (ORSEE, see Greiner 2004, 2015). They were required to have never taken part in an experiment before. Subjects signed up for two laboratory sessions. These were scheduled one week apart and run at the Laboratory for Experimental Economics at the University of Bonn in winter 2010/2011. Both sessions consisted of incentivized experiments and nonincentivized surveys, programmed in zTree (Fischbacher 2007). Each session lasted about two hours. Payoffs earned in the incentivized experiments were paid out to subjects at the end of each session.⁹ Average earnings over both sessions amounted to 64 euros (corresponding to approximately 83 U.S. dollars at the time of the experiment), including a fixed fee of 10 euros for participating in both sessions.

To minimize spillovers between the experimental and the survey measures, for example, because individuals might try to avoid cognitive dissonance (Festinger 1957) and strive for giving consistent responses (Falk and Zimmermann 2016; 2018), we never ran survey and experiment for the same preference during the same session. More specifically, we conducted all experiments

Table 1. Overview of Study Design

Group	Week 1	Week 2
Group 1 (n = 198)	Experiments on risk taking and time discounting; surveys on social preferences	Experiments on social preferences; surveys on risk taking and time discounting
Group 2 (n = 211)	Experiments on social preferences; surveys on risk taking and time discounting	Experiments on risk taking and time discounting; surveys on social preferences

relating to social preferences and all surveys relating to time discounting and risk taking in one session. The other session then contained the experiments relating to time discounting and risk taking as well as the surveys on social preferences. In addition, we reversed the order of experimental and survey elicitation of preferences for about half of our subjects to take care of potential order effects, that is, differences in behavior or responses due to differences in the way preferences were measured first. Table 1 gives an overview of the general study design.

We also conducted a pretest with 80 students. This pretest was intended to provide information on the duration and feasibility of the experiment. Experimental measures for negative reciprocity and altruism were not elicited in this pretest and the constraints on the participants regarding previous participation were not applied. Otherwise, the protocol was identical. In Section 3, we use data from this pretest for assessing the out-of-sample predictive performance of different candidate modules.

2.2. Choice Experiments

We elicited choices in standard economic choice experiments on risk taking, time discounting, altruism, trust,

and positive and negative reciprocity, respectively.¹⁰ The experiments that were used in each of the preference dimensions are summarized in Table 2. A detailed description of the experiments is relegated to Online Appendix A. Monetary stakes were presented to subjects in points, where 100 points equaled 80 cents. Subjects received feedback about the outcome of the experiments only at the end of the sessions in order to limit the impact of possible income effects on subsequent choices within a session. All experiments involving social or strategic interaction were one-shot to isolate social preferences from repeated game motives. Specifically, we implemented a perfect stranger random matching protocol implying that subjects never interacted more than once with the same person. Subjects were informed about this at the beginning of each session as well as before each experiment involving social interaction.

For risk taking, time discounting, trust, and positive reciprocity, we conducted two experiments each. These experiments had the same structure, but payoffs in the second experiment differed slightly, such that subjects were never asked to make trade-offs between alternatives that involved the exact same amounts. For instance,

Table 2. Overview: Experimental Measures

Preference	Experiment	Measure
Risk taking	Two multiple price lists in which subjects choose between a lottery and varying safe options.	Average of rows in both price lists in which subjects switch from preferring the lottery to the safe option.
Time discounting	Two multiple price lists in which subjects choose between a payment “today” and a larger payment “in 12 months”.	Average of rows in two price lists in which subjects switch from preferring the early to the delayed payment.
Trust	First mover behavior in two investment games.	Average amount sent as a first mover in both investment games.
Altruism	First mover behavior in a dictator game with a charitable organization as recipient.	Amount of donation.
Positive reciprocity	Second mover behavior in two investment games (contingent response method).	Average amount sent back in both investment games.
Negative reciprocity	Investment into punishment after unilateral defection of the opponent in a prisoner’s dilemma (contingent response method) and minimum acceptable offer in an ultimatum game.	Average score: amount invested into punishment and minimum acceptable offer in an ultimatum game.

the first lottery choice experiment involved 21 choices between a safe payment option, which increased in steps of 50 points from 0 points in the first choice to 1,000 points in the last choice, and a lottery that yields 1,000 points with probability 0.5 and 0 points otherwise. The row in which a subject switches from preferring the safe payment to the lottery gives bounds on the subject's certainty equivalent for the lottery.¹¹ We perturbed the safe payments in the second experiment by adding or subtracting a very small (up to five points) amount from each safe payment alternative. The number of points added or subtracted was determined by a randomly drawn integer value between -5 and $+5$. In the discounting experiments, in which subjects made choices between an immediate payment and a larger payment with a 12-month delay, the switching row gives bounds on the annual internal rate of return that makes the individual willing to wait.¹² We perturbed the delayed payment in the second experiment in the same manner as was done for the risk experiments.

The experimental measure of risk aversion was constructed by averaging over the switching rows in the two lottery choice experiments, which is equivalent in ordinal terms to averaging the implied monetary certainty equivalents.¹³ This averaging reduces measurement error compared with using a single experimental measure. Analogously, we constructed our experimental measure of time preference by averaging the switching rows, or equivalently annual internal rates of return, in the discounting experiments.¹⁴

Trust and positive reciprocity were elicited as first and second mover behavior, respectively, in two versions of the investment game (Berg et al. 1995). Each subject was in the role of the first and the second mover twice, such that overall each subject participated in four investment games. In one version, the amount sent by the first mover was tripled; in the other, it was doubled. For the second mover behavior, we implemented the contingent response method (Selten 1967). As our measure of trust, we again took the averages of the two decisions made as a first mover. For positive reciprocity, we first averaged all second mover decisions from the contingent response method in the two versions of the investment game. The average of these two amounts constitutes our preference measure of positive reciprocity.

For altruism, we conducted a dictator game with a charitable organization as recipient. The size of the donation constitutes our preference measure of altruism. For negative reciprocity, we conducted two different experiments. A subject's minimum acceptable offer in an ultimatum game (Güth et al. 1982) serves as one assessment of negative reciprocity. We obtain a second assessment from a subject's investment into punishment after unilateral defection of their opponent in a prisoner's dilemma (Falk et al. 2005). To obtain our preference measure of

negative reciprocity, we standardized both variables to account for the different response scales and then took the average.

2.3. Candidate Survey Items

For each type of incentivized choice experiment, we identified a set of candidate survey items for predicting choices in the experiment. The set for each experimental measure was on average roughly 30 survey items. In total, we included 188 survey items as candidates for selection into our survey module.¹⁵ Candidate items included both quantitative and qualitative questions. Many survey items were taken or adapted from existing surveys, like the German Socio-Economic Panel Study (SOEP) or the National Longitudinal Study of Youth (NLSY), or from previous research (e.g., Weber et al. 2002, Perugini et al. 2003). Additionally, we designed and included a number of new items. In defining this set of candidate items, we only included items that seemed widely applicable, that is, that were not limited to certain subject pools, for example, university students or employed individuals. In particular, we excluded some items found in the literature that refer to betting on horses, gambling, drug consumption, risky sports, taking a hitchhiker, or require respondents to be employed.¹⁶ Each battery of survey questions for a given preference domain began with a qualitative measure, asking respondents to self-assess their preference "in general" on an 11-point scale.¹⁷ Next, respondents were asked to state how they believe others judge them with respect to that preference and to compare their preference to the preferences of others. Then, respondents had to assess their preference in qualitative terms with respect to different domains, for example, financial decision making. Subsequently, subjects were confronted with a battery of additional qualitative and quantitative survey items.

Quantitative items typically included a hypothetical version of the incentivized choice experiment. Because the multiple price lists used in the lottery choice experiment and in the intertemporal choice experiment involve 30 choices and are rather time consuming, we also included an alternative elicitation procedure in which subjects only had to make five sequential choices. In the five-question measure of risk preference, all subjects first decided between the lottery versus a safe payment that slightly exceeds the expected value of the lottery. In the second decision (and all subsequent decisions), the lottery remained the same. If the participant had chosen the safe option in the first question, the safe option in the subsequent decision was smaller. If the participant had opted for the lottery, the safe payment increased. In the same manner, the safe option was increased or decreased in the third decision when the lottery or the safe payment were preferred in the second decision, respectively.

This procedure was repeated five times. Figure E1 in Online Appendix E.1 illustrates the method underlying this condensed quantitative measure, which is commonly referred to in psychology as the “staircase” method (Cornsweet 1962). For the case of time discounting, an analogous staircase elicitation was used in which the early option was identical in every choice whereas the delayed option varied. The procedures are described in detail in Online Appendix E.1 (for risk taking) and Online Appendix E.2 (for time discounting). Finally, we asked all subjects to rate the reliability of their survey answers.

3. Development of the Preference Module

3.1. Item Selection Procedure

Our aim was to develop a survey preference module that contains the set of items that best predict choices (revealed preferences) in incentivized laboratory experiments.¹⁸ Whereas some previous studies have investigated whether particular survey items are significantly correlated with experimental preference measures, our approach was to identify the combination of survey items from a large menu of alternative items that best predicts choices in incentivized experimental preference elicitation tasks. The basic idea is that different survey wordings and formats may be more or less accurate in predicting choices in experiments, for example, because of varying degrees of measurement error leading to more or less attenuation bias, or due to weaker or stronger links between what the survey items measure and the trait(s) that drive choices in the respective experiment. This is difficult to judge based on intuition alone, so we conduct incentivized choice experiments and use the observed choices as the benchmark for item selection.

We use a model selection approach, in the spirit of best subset selection (see, e.g., Hocking and Leslie 1967, Bertsimas et al. 2016), which consists of testing all possible combinations of our items using information criteria and then selecting the best model in terms of minimizing mean squared prediction error.¹⁹ To identify the best linear combination of items for measuring a particular preference, we proceeded in two stages, the first of which was running ordinary least squares (OLS) regressions of each experimental preference measure on all possible combinations of the respective set of candidate survey items as regressors. We used the results of this stage to identify, for each possible number of regressors, the best model in terms of explanatory power, using statistical criteria.²⁰ For selecting the best model with a given number of regressors it is equivalent to use R^2 , adjusted \bar{R}^2 , the Akaike information criterion (AIC), or the Bayesian information criterion (BIC) as these are identical up to a constant and only differ otherwise in terms of how

they penalize adding independent variables.²¹ We checked robustness to the linearity assumption in our selection procedure. Online Appendix C.4 provides reassurance that linearity is not misleading because the relationships between survey item responses and choices in the experiments are approximately linear.

In the second step, we compared the models identified in the first step using tests of predictive power. Whenever possible, we considered out-of-sample predictive power, making use of a truly independent sample of 80 subjects for whom we had collected data on the same experimental and survey measures on risk taking, time discounting, positive reciprocity, and trust. For each of these, we used the candidate survey models to derive predicted outcomes for each individual in the corresponding experiments.²² For each preference, we then compared the predictions of the alternative models to actual behavior, using the mean squared prediction error (MSPE). Comparing out-of-sample predictive performance helps avoid selecting models that do well in-sample because of overfitting. For all four preference experiments, the two-item model was preferred over modules of other lengths in that it had a lower MSPE.

Because data on altruism and negative reciprocity experiments were lacking in our independent sample, we evaluated the predictive power of the models for these experiments based on a proxy for out-of-sample prediction, provided by cross-validation using the original sample. Cross-validation involves using different subsets of the data for the fitting and prediction exercises, respectively. We ran five-fold and ten-fold cross-validations with 100 repetitions.²³ In line with our out-of-sample prediction results for the other four preference experiments, the two-item models are preferred according to the cross-validation.²⁴ Based on these findings, we selected two-item models as the best predictors for each of the preference experiments.

As a robustness check, we explored the results of using an alternative, popular model selection procedure based on the so-called lasso technique as introduced by Tibshirani (1996).²⁵ For each preference, lasso selects the same items that were identified using our two-step procedure. It also, however, selects a substantial number of additional items to include, leading to less parsimonious models.²⁶ Because parsimony is a key goal of our exercise for practical reasons, we prefer the two-item modules selected using our initial procedure, but Section D.3 in the online appendix displays the items selected by lasso.

3.2. Survey Items Contained in the Preference Module

Table 3 displays the items that were selected for the preference module with two survey questions for each preference dimension. Online Appendix B presents the

Table 3. The Preference Module

Preference		Item description	Weights
Risk	R2	Multiple price list (31 hypothetical choices between a lottery and a safe option).	0.2758
Taking	R3	Are you a person who is generally willing to take risks, or do you try to avoid taking risks?	0.2034
Time	D2	List of 25 hypothetical choices between an early payment “today” and a delayed payment “in 12 months”.	0.4849
Discounting	D4	In comparison with others, are you a person who is generally willing to give up something today in order to benefit from that in the future?	-0.1712
Trust	T24	Hypothetical investment game: first mover behavior.	0.6289
	T16	Self-assessment: As long as I am not convinced otherwise, I assume that people have only the best intentions.	0.1331
Altruism	A11	You won 1,000 euros in a lottery. Considering your current situation, how much would you donate to charity?	0.1845
	A10	How do you assess your willingness to share with others without expecting anything in return when it comes to charity?	0.3210
Positive	PR11	Hypothetical investment game: second mover behavior.	0.4857
Reciprocity	PR9	Hypothetical scenario: Which bottle of wine do you give as a thank-you gift?	0.1640
Negative	NR10	Minimum acceptable offer in hypothetical ultimatum game.	0.3284
Reciprocity	NR1	Are you a person who is generally willing to punish unfair behavior even if this is costly?	0.1479

Notes. The second column displays the item number as listed in Section G in the online appendix. See Section B in the online appendix for the exact wordings of the survey questions. The weights shown in the final column are OLS coefficients in a regression of the standardized experimental measure on the standardized module items. The survey measure for each preference is constructed by multiplying the items by the weights and adding. For details see the regression tables in Section C in the online appendix. Section D lists the survey items with the highest correlations with the experimental measure for each preference.

wording of the survey items in the preference module, translated from German to English; the original wording of the items in German is provided in Section D in the online appendix.

A notable feature of the preference module is its symmetry: For most preference dimensions, it contains a measure based on a hypothetical choice experiment and a qualitative item.²⁷ These two types of measures are complementary in the sense that the quantitative measure is akin to the standard revealed preference approach whereas the qualitative item is a subjective self-assessment. Previous research has shown that subjective assessments with abstract framings can lead to strong all-around predictors of life choices across many different life contexts. For example, a general assessment of willingness to take risks can predict a variety of behaviors ranging from holding risky assets, to being self-employed, to smoking (Dohmen et al. 2011). Quantitative survey measures that involve explicit monetary stakes are no exception, as they are somewhat tied to the context of financial decision making by construction; they may be better predictors of financial decisions in life than qualitative measures of a general disposition, but less predictive of choice in other domains. The preference module has a balance between both approaches.²⁸

The last column of Table 3 shows how the individual survey items for each preference can be combined into a single measure for predicting choices in the experiments, and also what their relative contributions are for predicting choices. The weights are the coefficients from OLS regressions of a given standardized experimental

measure on the standardized responses to the corresponding survey items (more details on the regressions are reported in Online Appendix C.1). The preference measure is obtained by applying the weights to the survey items and adding up. Due to the standardization, the weights directly show the relative contributions of the two items—specifically, by how much choices in the experiments are shifted in the distribution by a one standard deviation change in responses to an item. One can see, for example, that the quantitative item for risk preference has a roughly equal contribution to the qualitative item. In robustness checks, we investigated whether the optimal weights might differ for different demographic groups in our sample. Specifically, we ran regressions of the experimental choices on the survey items, including interaction terms with two observable demographics that have meaningful variation in our student sample: Gender, and an indicator for above median math grades. We do not find significant differences in the weights across these demographics, with the exception of positive reciprocity, for which women are slightly more reciprocal than men in the experiment even after controlling for survey responses.

The combined measure for each preference is an ordinal measure of preferences that ranks individuals in terms of predicted choices in incentivized experiments. For researchers who are interested in mapping survey responses into particular, cardinal representations of preferences (preference parameters), Section C.2 in the online appendix provides the necessary information.²⁹

4. Properties of the Preference Module

4.1. Within-Sample Explanatory Power of the Preference Module

As a first indication of the properties of the survey module, we present the within-sample correlations between the observed experimental choices and the choices predicted by the respective survey measure (each measure is constructed from responses to two survey items). The correlations are 0.41 for risk taking, 0.59 for time discounting, 0.67 for trust, 0.42 for altruism, 0.58 for positive reciprocity, and 0.37 for negative reciprocity. Thus, the survey module has substantial, but also imperfect, explanatory power within sample. One reason for finding correlations less than 1 can be measurement error in the survey measures, which leads to attenuation bias for the purposes of predicting choices.

Although 1 is a possible benchmark, this is not the only relevant benchmark, if the goal is deciding whether to use the survey module. In this case, a relevant benchmark could be the performance of alternative approaches that might be more accurate but entail higher cost. For example, a potentially superior approach for predicting choices in an incentivized experiment, in terms of accuracy, could be choices measured in exactly the same incentivized experiment.

To assess the (within-sample) predictive power provided by incentivized experiments, we use additional experiments with 44 subjects, who participated in preference elicitation experiments twice.³⁰ The experimental sessions were scheduled one week apart (there was no perturbation of experimental parameters across sessions) so the time difference is similar for our survey predictors. The correlations are 0.59 for risk taking, 0.82 for time discounting, 0.77 for trust, and 0.65, 0.66, 0.67 for altruism, positive reciprocity, and negative reciprocity, respectively.³¹ Thus, it is the case that the survey module sacrifices some predictive power, for each of the preference experiments, relative to using corresponding incentivized experiments as predictors, but the difference is less stark than when comparing to a benchmark of 1. At the same time, the survey module has the benefit of being less costly.

Measurement error in the survey module can attenuate explanatory power for incentivized choices in experiments, or other outcome variables, as well as make the module items imperfect statistical controls (for discussions see, e.g., Spearman 1904, Gillen et al. 2019). To provide a measure for the extent of measurement error in the survey module, and the potential benefits of multiple measurements, we also conducted additional sessions, in which 85 subjects answered the survey module questions in one session, and then answered the survey module again when they returned for a second session, one week later. The correlations between the repeated measures of the survey module (test-retest correlations) are

0.76, 0.86, 0.79, 0.84, 0.71, and 0.85 for risk, time, trust, altruism, positive reciprocity, and negative reciprocity, respectively. The fact that these correlations are less than 1 indicates that the survey items do contain measurement error, which contributes to attenuation bias in predicting choices in experiments.³² One implication is that having two or more measurements of the survey module for the same individual can be beneficial because of the potential to reduce measurement error. For example, with two measures of the survey module for each individual, one week apart, one can purge the survey module of measurement error using a standard instrumental variables approach involving instrumenting for survey response at time t with survey response at $t - 1$ (under the assumption that measurement error in the survey is uncorrelated over time; for a discussion see, e.g., Vansteelandt et al. 2009). Our test-retest correlations suggest that this can lead to a nontrivial increase in ability to explain incentivized choices in experiments.³³ This approach comes at a cost, however, of needing to implement the survey twice for each person. As the module does have explanatory power even with a single measure, researchers face a trade-off, and can decide for their particular application whether reduced error justifies the logistical cost of multiple measures.

4.2. Out-of-Sample Prediction of the Preference Module

Another relevant property of the module is its (absolute) performance in out-of-sample prediction. For the subjects in our pretest panel, we used their survey responses to predict their choices in the four experimental preference elicitation tasks (measuring risk and time preferences, trust, and positive reciprocity), and regressed the actual choices on the predicted choices. If our preference module perfectly captured the preferences of individuals in this sample, one would expect the intercept of the regression of actual on predicted choices to be zero and the coefficient of the predicted value to be exactly one. In fact, we cannot reject the hypothesis that the constant is zero and the slope coefficient equals one for all preferences, except for trust, at the 10% significance level. For trust, we find that the slope coefficient is not statistically different from one if we suppress the constant in the regression. It is also reassuring that the out-of-sample predicted and actual choices are strongly and statistically significantly correlated. The correlations are 0.29 for risk preferences, 0.59 for time discounting, 0.26 for trust, and 0.44 for positive reciprocity.

4.3. Evidence on the Viability of Individual Survey Items in Nonstudent and International Samples

Although the selection procedure was based on data from a German student population, there are several

reasons to expect that the resulting module is useful for other populations.

First, although the distribution of preferences might very well differ across populations, the module will be meaningful as long as the correlation structure is not too different. Note that the top two survey predictors for our student sample were typically superior to other measures by a substantial margin, so it is likely that the two measures would perform well if one were to do a similar validation exercise for other populations. Second, the quantitative survey items in our modules closely resemble experimental measures of preferences, which are largely context free and have been widely used to elicit preferences in nonstudent and culturally diverse samples. Third, and most importantly, there are also various pieces of empirical evidence, which show that survey measures similar to, or identical to, the ones used in our modules are significantly correlated with experimental preference measures in nonstudent and non-German samples.

Regarding nonstudent samples, Fehr et al. (2002) used a representative sample of German adults, and documented a significant correlation between subjects' behavior in an incentivized investment game, and survey measures on trust of the type contained in our preference module. Likewise, it has been shown that answers to the qualitative survey question to elicit risk attitudes, contained in our preference module, are significantly correlated with incentivized lottery choices in a large representative subject pool of German adults (Dohmen et al. 2011). In fact, they report a correlation coefficient between the survey measure and behavior in the lottery choice experiment in their representative sample that is almost identical to the one in our validation sample consisting of students.³⁴ It is also notable that the correlation is not significantly different for students versus nonstudents in their representative sample. Similarly, Ziegelmeyer and Ziegelmeyer (2012) predict risk-taking behavior in an alternative lottery choice experiment (Holt and Laury 2002) using the same survey item that is part of our module. In addition, the qualitative survey risk measure contained in our preference module has previously been administered in the German Socio-Economic Panel Study, and other large representative surveys in the United States, Asia, and Australia as well as in other European countries. Various studies have documented that for representative and therefore heterogeneous population samples answers to this question are related to risky behaviors in many contexts of life, for example, occupational choice and self-employment, geographical mobility, ownership of risky assets, as well as smoking (see, e.g., Bonin et al. 2007, Caliendo et al. 2009, Jaeger et al. 2010, Dohmen et al. 2011, Barasinska et al. 2012, Bauernschuster et al. 2014, Fouarge et al. 2014). These findings illustrate that the types of survey items selected in our preference

module provide behaviorally valid preference measures in nonstudent samples.

Moreover, there is previous supporting evidence that items from our preference survey module are valid across a wide range of cultures. For example, recent empirical work by Vieider et al. (2015) uses the same qualitative measure of risk attitudes that is included in our module and documents that it correlates with incentivized lottery choice experiments conducted in 30 different countries. In addition, Hardeweg et al. (2013) replicate the validation exercise of Dohmen et al. (2011) and confirm the significant relationship between this risk question and incentivized lottery choices for a representative sample of 900 inhabitants of rural Northern Thailand. Ding et al. (2010) corroborate these results for a sample of 121 Beijing University students.

Finally, Section 5.2 discusses further evidence on the validity of the items in nonstudent and non-German samples.

4.4. Potential Limitations

Naturally, some aspects of our design choices in this validation exercise imply potential limitations. For example, despite ample evidence discussed in the preceding section that many of our module items have predictive power in non-German and nonstudent samples, we fully acknowledge that we cannot rule out that items or item combinations other than the ones selected for our modules might perform even better in non-German or nonstudent samples. Although this goes beyond the scope of the current paper, we think that running our validation exercise using different samples would provide valuable results on the usefulness of different preference measures, for example, in other countries or in specific subgroups of populations, such as managers or entrepreneurs.

Similarly, we picked a very specific benchmark by which we measured the usefulness of preference measures: incentivized choice experiments that were largely context free. Perhaps somewhat unsurprisingly, survey items that best predict choices in these experiments are largely context free themselves, such as hypothetical versions of these choice experiments or questions about one's general willingness to take risks. In the light of evidence on the context-dependence of preferences (Tversky and Simonson 1993, Barseghyan et al. 2011, Einav et al. 2012, Ellingsen et al. 2012), our approach might come with the caveat that more context-specific items might work even better than the more context-free items selected for our preference modules. However, this does not imply that our modules are not valid in more specific contexts. For example, Dohmen et al. (2011) show that the general risk question often outperforms more context-specific risk questions in predicting domain-specific risk taking.

Moreover, even though preferences affect a range of important life outcomes, such as consumption, labor market, or health-related choices, it might very well be the case that measures other than our selected survey measures perform better at predicting such choices. After all, these choices are consequences not only of preferences, but also of beliefs, constraints, or institutions. Future work might want to shed light on which survey measures perform best in predicting such life outcomes.

5. Recipes for Constructing Alternative Preference Modules

Although our proposed survey module is the best module according to the specified criteria, researchers might have other needs that call for developing alternative preference modules. For example, it might be desirable for certain applications to only use qualitative survey items, or to have a survey module that is even briefer than the one we develop.

5.1. Performance of Individual Survey Items and Alternative Two-Item Modules

For researchers who might want to use individual survey items, or alternative survey modules based on our survey items, we provide additional information in the online appendix. Tables D1 to D6 give the correlations between individual survey measures and the corresponding preference experiment, focusing on the 10 items with the highest correlations for each preference. Notably, the items selected in our preferred preference module are always included in these sets of best individual performers. Table D7 gives the adjusted \bar{R}^2 for alternative two-item survey measures for each preference, focusing on all possible combinations of the set of the 10 best individual measures. Researchers can use these alternative measures if for some reason they prefer the included survey formats, knowing how this performs relative to the benchmark of the best overall measure and a range of alternative measures.

5.2. The Global Preference Survey (GPS) Module

The survey module developed so far offers an easily implementable and lower cost alternative to conducting incentivized experiments, and it is optimal relative to a wide variety of alternative possible survey measures. Nevertheless, there are applications for which this module will not be ideal, as some of the quantitative items either require instructions that are as complex as corresponding experiments (e.g., the hypothetical investment game) or entail a considerable number of decisions (e.g., multiple price lists for eliciting risk and time preferences). Particularly if time constraints are severe or if respondents have limited cognitive capacity, an even

simpler and shorter module seems useful, although this might come at some costs in terms of lower explanatory power.

A prime example of an application for which our main module might not be implementable is a large-scale international survey. In 2012, we wanted to collect preference measures for nationally representative samples in 76 countries around the globe through the professional infrastructure of the Gallup World Poll framework.³⁵ This required us to tailor our initial module version to this specific application in which we faced tight survey time constraints, heterogeneous population samples, and the fact that data collection would be conducted using telephone interviews in the majority of cases. In what follows, we will give an overview over the process of fine-tuning our module to this large-scale cross-cultural study, describe the adjustments we made, and present the resulting GPS module. This can potentially provide a road map for researchers with similar goals. A more detailed description is relegated to Section E in the online appendix.

Developing the GPS module involved two main steps. First, in light of the tight survey time constraints we faced, the heterogeneous population samples, and the implementation method, we discarded the hypothetical versions of our experimental preference elicitation tasks, which are relatively time consuming as they involve a large number of choices or require rather complex instructions that do not seem advisable in telephone surveys. We then implemented the selection procedure described in Section 3 on the set of remaining survey items. As this restricted set still included (simpler) analogues of the discarded items, this restriction ultimately only led to a minimal reduction in explanatory power (R^2) (see Online Appendix E). For example, in the case of risk taking and time discounting the staircase measures were selected. These measures are very comparable to the more complicated quantitative measures based on the multiple price lists for lottery choices and intertemporal choice, respectively, yet their implementation is much more time efficient, as the staircase procedures only require five interdependent choices (lottery vs. safe payments and early vs. delayed payments, respectively).³⁶ Because these preference measures are highly correlated with the respective multiple price list measure and with the respective experimental preference measure (see Section C in the online appendix), the reduction in explanatory power of the streamlined version compared with the original version in terms of R^2 is only 0.02 in the case of risk taking and 0.04 in the case of time discounting.

Second, we tested the resulting preference module, which is based on the modified set of candidate measures, in an in-depth pilot study in 22 countries. In collaboration with Gallup Europe, we surveyed respondents from 10 countries in central Asia (Armenia, Azerbaijan,

Belarus, Georgia, Kazakhstan, Kyrgyzstan, Russia, Tajikistan, Turkmenistan, and Uzbekistan), two countries in South-East Asia (Bangladesh and Cambodia), five countries in Southern and Eastern Europe (Croatia, Hungary, Poland, Romania, and Turkey), four countries in the Middle East and North Africa (Algeria, Jordan, Lebanon, and Saudi Arabia), and one country in Eastern Africa (Kenya).³⁷ In this test phase, in each country 10 to 15 people were interviewed, resulting in more than 220 interviews being conducted overall. In almost all countries, the sample composition was heterogeneous in terms of gender, age, educational background, and area of residence (urban vs. rural). To detect potential difficulties in the understanding of module items and differences in the respondents' interpretation, respondents were explicitly asked to give extensive feedback with respect to the appropriateness and understandability of the module. In particular, we asked respondents to rephrase the items in their own words and to state any concerns or difficulties in understanding of the items that they had or that they thought future respondents of their country or culture might have.³⁸ Likewise, if the meaning of an item was unclear to a respondent, the interviewer would explain it to the individual and then ask the respondent to rephrase it in the person's own words.

Overall, the understanding and implementability of our module was very good. Nevertheless, respondents' feedback induced some additional changes to some items. In terms of wording changes, the use of the term "lottery" in hypothetical risky choices was troubling to some Muslim participants, and some refused to answer the item completely since gambling is taboo (*haram*) in

Islam. As a consequence, we dropped the term "lottery" and replaced it with the more neutral but equally accurate term "random draw." Second, the term "charity" caused confusion in Eastern Europe and Central Asia, so it was replaced with "good cause." Third, some respondents had difficulties answering the question asking about one's willingness to punish unfair behavior without knowing who was treated unfairly. We therefore decided to split the question into two separate items, one item asking for one's willingness to punish unfair behavior toward others, and another asking for one's willingness to punish unfair behavior toward oneself. Fourth, some participants, especially in countries with current or relatively recent phases of volatile and high inflation rates, stated that their answer to questions involving intertemporal trade-offs would depend on the rate of inflation, or said that they would always take the immediate payment due to uncertainty with respect to future inflation. Therefore, we added the following phrase to each question involving hypothetical choices between immediate and future monetary amounts: "Please assume there is no inflation, i.e., future prices are the same as today's prices." The final version of the GPS module is presented in Table 4. Finally, the survey questions were brought into a format that is consistent with the Gallup World Poll questionnaire style, a well-validated format for eliciting responses in an international sample. For example, the first question of the module, which happened to be the qualitative survey question on risk taking, was commenced by the request "Please tell me." The complete module version including exact wordings is relegated to Section F in the online appendix.

Table 4. The GPS Module

Preference	Item description	Weights
Risk taking	1. Staircase measure (five interdependent choices between a lottery and a safe option).	0.2159
	2. Please tell me, in general, how willing or unwilling you are to take risks.	0.2406
Time discounting	1. Staircase measure (five interdependent choices between an early and a delayed amount of money).	0.4417
	2. How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?	0.1791
Trust	1. I assume that people have only the best intentions.	0.2656
	1. Hypothetical donation.	0.1845
Altruism	2. How willing are you to give to good causes without expecting anything in return?	0.3210
	1. Hypothetical choice: size of a "thank-you" gift.	0.2876
Positive reciprocity	2. When someone does me a favor I am willing to return it.	0.2705
	1. If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so.	0.0884
Negative reciprocity	2. How willing are you to punish someone who treats you unfairly, even if there may be costs for you?	0.0741
	3. How willing are you to punish someone who treats others unfairly, even if there may be costs for you?	0.0741

Notes. The second column displays the items as they were adapted to serve the purpose of the GPS study. Online Appendix Section E describes how the wordings etc. were adjusted. The weights shown in the last column are coefficients resulting from OLS regressions using the items with the original wording from the validation sample. The survey measure for each preference can be constructed by multiplying the items by the weights and adding.

For the purpose of implementing the module in the Gallup World Poll, for all items involving hypothetical monetary amounts we adjusted the stake sizes for each country in terms of their real value such that they represent the same share of a country's median income in local currency as the share of the amount in euros of the German median income, where our initial validation study had been conducted. Monetary amounts used in the validation study with the German sample were rounded numbers to facilitate easy calculations (e.g., the expected return of a lottery with equal chances of winning and losing) and to allow for easy comparisons (e.g., 100 euros today vs. 107.50 in 12 months). To proceed in a similar way in all countries, monetary amounts were always adjusted to the next "round and easy" number after adjusting the amounts in terms of their real values.³⁹

A comprehensive analysis of the resulting GPS data on economic preferences from nationally representative samples in 76 countries is presented in Falk et al. (2018). Whereas they document pronounced heterogeneity in preferences both across and within countries, they also show that within countries preferences are systematically related to outcomes in ways which economic theory would predict, and these relationships are similar for almost all countries. For example, patience as measured by the two-item modules is positively correlated with savings and education in more than 90% of the countries. Likewise, risk aversion is negatively associated with being self-employed and with smoking intensity, and there is a positive relationship between altruism and different giving behaviors in the vast majority of countries. This provides a further important and independent check of the validity of our measures and their applicability across cultures.

6. Conclusion

This paper presents survey modules designed to proxy for incentivized measures of economic preferences from experiments—risk aversion, patience, trust, altruism, and positive and negative reciprocity. The guiding methodology for developing the modules is identifying survey items that can predict well the choices in incentivized experiments. Responses to the resulting survey measures provide predictions about choices in such settings and thus reveal preferences, in an ordinal sense, and in a cardinal sense under additional assumptions about, for example, the functional form of utility. The paper offers two versions of the module. One provides the maximum explanatory power, subject to having a parsimonious number of survey items (two items) per preference. This module is particularly well suited for eliciting preferences in studies for which time constraints are not too severe, such as laboratory experiments and many field experiments. This version of the module is also likely to work

well for surveys that use detailed questionnaires, or that are based on written or computer-assisted personalized interviews (CAPI) that can implement more complex types of survey items. The second version of the module, the GPS module, was tailored to the requirements and particular characteristics of a multinational survey with nationally representative population samples: tight time constraints and respondents that are diverse in terms of education, socioeconomic status, and culture. It is streamlined in that it prioritizes time efficiency and simplicity at the expense of a modest reduction in explanatory power.

Both versions of the preference module share several desirable features. First, the module items are experimentally validated. The ability of the items to explain behavior in incentivized choice experiments helps ensure that they are meaningful for predicting choices under real incentives, mitigating one of the major concerns about hypothetical questions. The selected items are not just significant predictors of behavior, but are jointly the best predictors out of a large set of alternative measures. The validation is based on a consistent research design across preferences, and applies state-of-the-art experimental techniques and transparent, quantitative criteria for module selection. Second, the modules consist of a balanced mix of qualitative self-assessments and questions involving quantitative hypothetical trade-offs. This gives the module an attractive balance between different approaches to assessing preferences. Third, the module has a wide range of possible applications. The two versions can be implemented in various survey modes, including modes with tight time constraints. Fourth, by providing an attractive and low-cost approach to measuring preferences, the modules have the potential for widespread adoption, with potentially significant positive externalities in terms of easier comparison of results across studies.

Beyond the specific survey modules provided in the paper, the paper includes information that researchers can use to design their own preference modules. This includes findings about the explanatory power of a wide range of survey items, as well as alternative combinations of the items. Though lacking some of the predictive power of the modules designed in our procedure, these individual questions or alternative modules may suit the purposes of researchers depending on the circumstances they face. The paper also provides a recipe for validating survey modules as proxies for incentivized experiments. This can be used by researchers to develop new types of preference modules.

Directions for future research include developing survey modules that are optimized for particular populations or cultures, or developing survey modules for other important aspects of preferences, for example, present-bias, or loss aversion, or ambiguity aversion. By varying the context embedded in experiments, it

may be possible to develop survey modules optimized to particular contexts, in line with research on the domain specificity of preferences (see, eg., Chapman 1996, Weber et al. 2002). Survey modules on economic preferences might also be used to study the related notion of constructed preferences (Slovic 1995; for a survey, see Warren et al. 2011).

Acknowledgments

The authors thank Thomas Deckers, Fabian Kosse, Mirko Seithe, Benedikt Vogt, Ulf Zölitz, and seminar participants at many institutions for helpful comments. Marco Merle, Patrizia Odyńiec, and Sven Walter provided excellent research assistance.

Endnotes

¹ An alternative methodology for measuring preferences is to use life outcomes as a proxy for preferences. Although this has the advantage of involving real (typically self-reported) behavior, for potentially large stakes, a disadvantage is that a given life outcome may depend on many personal and environmental factors besides the preference of interest. By contrast, both experiments and survey measures can pose individuals with carefully designed scenarios and choice options, which can isolate a particular preference with a reasonably high degree of precision, and which are held exactly the same across respondents. This can help eliminate a major source of unobserved heterogeneity that affects the inference of preferences from life outcomes.

² In psychology, the strength of the relationship between the survey measure and the construct in the absence of measurement error is known as criterion validity. Criterion validity for a survey measure could be low if, for example, it asks about a willingness to engage in a behavior that is mainly determined by other traits besides the trait(s) that drive choices in the respective experiment.

³ Unlike psychometric measures, economic preference measures have an interpretation in the context of economic theory and can be used to generate qualitative or quantitative predictions from economic models.

⁴ Whereas experimental and empirical work—in line with economic theory—has highlighted the role of economic preferences in workplace decisions, most work has used incentivized experiments to measure preferences and therefore relied on student or other convenience samples (see, e.g., Dohmen and Falk (2011) on sorting of employees into incentive schemes; Bandiera et al. (2005) and Falk and Kosfeld (2006) for employees' responses to changes in the incentive structure; Falk et al. (2005) for contract enforcement; and Cohn et al. (2015) for investment behavior of financial professionals).

⁵ Incentivized experiments have been implemented for nonstudent and also representative samples, see, for example, Harrison et al. (2002), Andersen et al. (2008), and Fehr et al. (2002).

⁶ One reason why survey measures that work well for one population might be suboptimal for another is if the survey measures suffer from hypothetical bias, and this bias is different for different populations. For evidence on hypothetical bias see, for example, Blackburn et al. (1994), List and Gallet (2001), Murphy et al. (2005), and Harrison and Rutström (2008).

⁷ Falk et al. (2018) analyze the GPS data and find that the survey preference measures are related to economic outcomes in a similar way across 76 countries. This provides an additional indication that the survey module is useful across a wide range of cultures.

⁸ Fehr et al. (2002), for example, examine six different attitudinal trust questions in terms of their ability to predict behavior in an investment game as introduced by Berg et al. (1995), and find that

self-rated trusting behavior and willingness to trust strangers are most strongly associated with behavior in the incentivized experiment. Dohmen et al. (2011) show that self-rated willingness to take risk in general is significantly correlated with decisions in an incentivized lottery choice experiment. Vischer et al. (2013) relate answers to a survey question asking respondents to rate their general level of impatience to behavior in an experiment involving inter-temporal trade-offs.

⁹ The payments resulting from the choice experiments on time discounting were delivered to the subjects in cash via regular mail, either at the same day of the session or 12 months later, depending on the payoff relevant choice.

¹⁰ There are other types of experiments that can be used to measure the respective preferences. See, for example, Andreoni and Sprenger (2012), Toubia et al. (2013), or Chapman et al. (2018) for alternative measures of time and risk preferences. Future research could explore the relative predictive powers of survey items for these alternative measures.

¹¹ The implied certainty equivalent lies between the safe payment in the switching row and the safe payment in the preceding row.

¹² The implied internal rate of return lies between the rate of return offered in the switching row and the one offered in the preceding row.

¹³ We abstract away from the negligible impact of the perturbed safe payments on the intervals for the certainty equivalent implied by switching row in a given experiment. As is common for this type of elicitation method, some subjects exhibit multiple switching points. We observe that 22% of individuals switch more than once from preferring the lottery to the safe payment in either of the two lottery choices experiments, nine of them have multiple switch points in both experiments. For subjects who make that kind of inconsistent choices, we calculate the average switching row in each choice table and construct the experimental measure of risk aversion as the mean of the two averages.

¹⁴ We abstract away from the negligible impact of the perturbed early payments on the intervals for the internal rate of return implied by switching row in a given experiment. In the discounting experiments, we observe that around 16% of subjects switch more than once in one or the other experiment, and about 3% switch multiple times in both experiments. For these subjects, we construct the experimental measure by taking the mean of the average switching row in the two experiments involving intertemporal choices.

¹⁵ Section A in the online appendix gives a list of all survey items in the candidate set.

¹⁶ Some of these items might work well for particular subsamples of the population, but will most likely be uninformative and inappropriate for large fractions of more general population samples. Although not included in the set of candidate items for the module selection exercise, some of these items were nevertheless included in the questionnaire for the study, because they formed part of standard scales found in the literature.

¹⁷ An example of this type of question is the general risk question that was validated in Dohmen et al. (2011).

¹⁸ Another important ex ante criterion for developing the module was cost efficiency, that is, considering the trade-off between predictive power and conciseness of the module, but as it turns out, the statistical criteria favored combinations that are quite parsimonious in terms of the number of items.

¹⁹ Alternative model selection procedures such as forward selection and backward selection also use information criteria, but do not consider all possible combinations of items, and to some extent suffer from problems of order dependence. Such approaches are also different from ours because they do not include the additional step

of minimizing (out-of-sample) mean squared prediction error, a step that helps address the problem of overfitting.

²⁰ In the following, we will only report results from OLS regressions. However, all results reported here are robust to estimating ordered probit models and selecting items using the criteria of maximum log-likelihood or pseudo- \bar{R}^2 .

²¹ To see this, note that $AIC(\hat{\theta}) = (-2)\log(L) + 2k$ and $BIC(\hat{\theta}) = (-2)\log(L) + k \cdot \log(n)$, with k held constant by the number of model parameters and the sample size n held constant for the purpose of comparing models. We also checked robustness to pooling the survey items for all six preferences, and then using \bar{R}^2 to identify the best model of a given length out of the entire set of roughly 180 survey items. We find for each preference that the same two-item survey modules are selected, for example, we do not find a better two-item module for predicting risk that includes one of the candidate time preference survey items.

²² Predicted values were calculated as the product of the vector of observed answers to the specific preference module and the vector of estimated coefficients from the regression of the experimental preference measure on the respective preference module in the main sample on which the selection procedure was based.

²³ Each cross-validation involved randomly splitting the sample into k partitions (with $k = 5$ or $k = 10$). We used $k - 1$ of the partitions to fit the model (the training sample) and used the resulting coefficient estimates to predict choices for the remaining k th partition (the prediction or hold out sample). This yielded k measures of prediction error, which we averaged. We repeated this procedure 100 times for a given model and took the overall average.

²⁴ This is true for both five- and 10-fold cross-validation. Furthermore, the two-item modules would also be selected based on a range of standard information criteria (BIC, AIC, adjusted R-squared, and likelihood-ratio test (LRT)). Note that due to pessimistic bias in the cross-validation procedure, it is standard to not select the model with the minimum prediction error, but rather the most parsimonious model that falls within a narrowly defined confidence interval of the minimum. In our case, the minimum was obtained with the three item module for each of the two preferences, but the two-item module had only a slightly larger error while being more parsimonious. See Online Appendix C.5 for error plots from the cross-validation. Often, one standard error above the minimum is allowed, but we chose a tighter bound of one-fifth of a standard deviation to sacrifice only minimal prediction accuracy.

²⁵ Lasso is particularly useful when there are more potential explanatory variables than observations, since in such cases there is not a unique solution for OLS. Lasso is also particularly useful when it is not feasible to consider all possible item combinations. Neither is the case in our setting.

²⁶ This is true regardless of whether we run a simple linear lasso with cross-validation or whether we allow the lasso penalty parameter to be adaptive.

²⁷ The only exception is positive reciprocity.

²⁸ One explanation for why the procedure selects a balanced module is that quantitative survey formats may have some form of measurement error in common, and likewise qualitative survey formats may have a common error component, but measurement error may be less correlated across these different types of survey formats. If this is the case, it tends to favor having a balanced module, because this contains more independent information than having two items with the same format.

²⁹ The table in Online Appendix C.2 shows how responses to survey items map into (nonstandardized) monetary values associated with predicted choices in the experiments. For example, in the case of risk, the information allows mapping responses to the risk survey

items into predicted certainty equivalents for the lottery that we use in our risk experiments. By making additional assumptions such as behavior according to expected utility theory (EUT) and a particular functional form of utility, for example, constant relative risk aversion (CRRA) utility, it is possible to infer bounds for a preference parameter.

³⁰ Similar to participants in the main sample, these 44 participants came to the laboratory twice. Both times, they participated in the set of incentivized experiments for each preference. We did not elicit survey measures for these participants.

³¹ A more detailed regression table is relegated to Section C.3 in the online appendix.

³² The test-retest correlations for the incentivized experiments, and the survey module, respectively, allow a measurement error correction of the correlations between experiment choices and choices predicted by the survey module (see, e.g., Fan 2003). Eliminating measurement error in both experiments and the survey module, the correlations would be 0.61 for risk taking, 0.70 for time discounting, 0.86 for trust, 0.57 for altruism, 0.85 for positive reciprocity, and 0.50 for negative reciprocity. This is an average increase in the correlation between observed and predicted choice of about 35%.

³³ The one-week test-retest correlations for the survey module allow calculating the resulting correlation of (instrumented) predicted choices with observed choices in the experiment (assuming experiment choices are measured without error): 0.47 for risk taking, 0.64 for time discounting, 0.75 for trust, 0.46 for altruism, 0.69 for positive reciprocity, and 0.40 for negative reciprocity. This is an average increase of 12% in the correlation between predicted and observed choices (equivalently, a 25% increase in R^2 for a regression of observed choices in a given experiment on responses to the corresponding survey measure). Thus, there is a modest but nontrivial improvement in ability to explain experiment choices, due to reduced measurement error, from implementing the survey module twice for an individual. Researchers may also consider an alternative correction based on having two measures of the survey module for each individual, proposed by Gillen et al. (2019), which is similar but uses each of the two measures to instrument for the other, and takes the average.

³⁴ The correlations are 0.25 in the representative sample of Dohmen et al. (2011), and 0.24 in our validation sample if we focus on the same survey measure for predicting behavior in a single risk experiment (as shown earlier, the correlation is even higher for the validation sample if we use choices from both risk aversion experiments).

³⁵ The World Poll are annual nationally representative surveys conducted in more than 160 countries; see <http://www.gallup.com/analytics/213704/world-poll.aspx> for more information.

³⁶ The staircase procedures are presented in detail in Online Appendix E.1 and E.2.

³⁷ Gallup Europe ensured that the items of the preference module were translated into the major languages of each target country, using state-of-the-art techniques. The translation process involved three steps. As a first step, a translator suggested an English, Spanish, or French version of a German item, depending on the region. A second translator, being proficient in both the target language and in English, French, or Spanish, then translated the item into the target language. Finally, a third translator would review the item in the target language and translate it back into the original language. If differences between the original item and the back-translated item occurred, the process was adjusted and repeated until all translators agreed on a final version.

³⁸ For example, respondents were explicitly asked to explain a “50-percent chance” in their own words and give their own interpretation of “safe payment.”

³⁹ Although this necessarily resulted in some (minor) variations in the real stake size between countries, it minimized cross-country

differences in the understanding and complexity of the quantitative items due to difficulties in assessing the involved monetary amounts.

References

- Andersen S, Harrison G, Lau M, Rutström E (2008) Eliciting risk and time preferences. *Econometrica* 76(3):583–618.
- Andreoni J, Sprenger C (2012) Estimating time preferences from convex budgets. *Amer. Econom. Rev.* 102(7):3333–3356.
- Bandiera O, Barankay I, Rasul I (2005) Social preferences and the response to incentives: Evidence from personnel data. *Quart. J. Econom.* 120(3):917–962.
- Barasinska N, Schäfer D, Stephan A (2012) Individual risk attitudes and the composition of financial portfolios: Evidence from German household portfolios. *Quart. Rev. Econom. Finance* 52(1): 1–14.
- Barseghyan L, Prince J, Teitelbaum JC (2011) Are risk preferences stable across contexts? Evidence from insurance data. *Amer. Econom. Rev.* 101(2):591–631.
- Bauernschuster S, Falck O, Heblich S, Suedekum J, Lameli A (2014) Why are educated and risk-loving persons more mobile across regions? *J. Econom. Behav. Organ.* 98:56–69.
- Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. *Games Econom. Behav.* 10(1):122–142.
- Bertsimas D, King A, Mazumder R (2016) Best subset selection via a modern optimization lens. *Ann. Statist.* 44(2):813–852.
- Blackburn M, Harrison GW, Rutström EE (1994) Statistical bias functions and informative hypothetical surveys. *Amer. J. Agricultural Econom.* 76(5):1084–1088.
- Bonin H, Dohmen T, Falk A, Huffman D, Sunde U (2007) Cross-sectional earnings risk and occupational sorting: The role of risk attitudes. *Labour Econom.* 14(6):926–937.
- Caliendo M, Fossen FM, Kritikos AS (2009) Risk attitudes of nascent entrepreneurs: New evidence from an experimentally validated survey. *Small Bus. Econom.* 32(2):153–167.
- Chapman GB (1996) Temporal discounting and utility for health and money. *J. Experiment. Psych. Learn. Memory Cognition* 22(3): 771–791.
- Chapman J, Snowberg E, Wang S, Camerer C (2018) Loss attitudes in the US population: Evidence from dynamically optimized sequential experimentation (DOSE). NBER Working Paper Series, Cambridge, MA.
- Cohn A, Engelmann J, Fehr E, Marechal MA (2015) Evidence for countercyclical risk aversion: An experiment with financial professionals. *Amer. Econom. Rev.* 105(2):860–885.
- Cornsweet TN (1962) The staircase-method in psychophysics. *Amer. J. Psych.* 75(3):485–491.
- Ding X, Hartog J, Sun Y (2010) Can we measure risk attitudes in a survey? IZA Discussion Paper No. 4807.
- Dohmen T, Falk A (2011) Performance pay and multidimensional sorting: Productivity, preferences, and gender. *Amer. Econom. Rev.* 101(2):556–590.
- Dohmen T, Falk A, Huffman D, Sunde U, Schupp J, Wagner G (2011) Individual risk attitudes: Measurement, determinants, and behavioral consequences. *J. Eur. Econom. Assoc.* 9(3):522–550.
- Einav L, Finkelstein A, Pascu I, Cullen MR (2012) How general are risk preferences? Choices under uncertainty in different domains. *Amer. Econom. Rev.* 102(6):2606–2638.
- Ellingsen T, Johannesson M, Mollerstrom J, Munckhammar S (2012) Social framing effects: Preferences or beliefs? *Games Econom. Behav.* 76(1):117–130.
- Falk A, Kosfeld M (2006) The hidden costs of control. *Amer. Econom. Rev.* 96(5):1611–1630.
- Falk A, Zimmermann F (2016) Consistency as a signal of skills. *Management Sci.* 63(7):2197–2210.
- Falk A, Zimmermann F (2018) Information processing and commitment. *Econom. J.* 613(1):1983–2002.
- Falk A, Fehr E, Fischbacher U (2005) Driving forces behind informal sanctions. *Econometrica* 73(6):2017–2030.
- Falk A, Becker A, Dohmen T, Enke B, Huffman D, Sunde U (2018) Global evidence on economic preferences. *Quart. J. Econom.* 133(4):313–332.
- Fan X (2003) Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational Psych. Measurement* 63(6):915–930.
- Fehr E, Fischbacher U, Rosenblatt B, Schupp J, Wagner G (2002) A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative surveys. *Schmollers Jahrbuch* 122(4):519–542.
- Festinger L (1957) *A Theory of Cognitive Dissonance* (Stanford University Press, Stanford, CA).
- Fischbacher U (2007) zTree: Zurich toolbox for ready-made economic experiments. *Experiment. Econom.* 10:171–178.
- Fouarge D, Kriechele B, Dohmen T (2014) Occupational sorting of school graduates: The role of economic preferences. *J. Econom. Behav. Organ.* 106:335–351.
- Gillen B, Snowberg E, Yariv L (2019) Experimenting with measurement error: Techniques with applications to the caltech cohort study. *J. Political Econom.* 127(4):1826–1863.
- Greiner B (2004) An online recruitment system for economic experiments. Kremer K, Macho V, eds. *Forschung und wissenschaftliches Rechnen* (Gesellschaft für Wissenschaftliche Datenverarbeitung, Göttingen, Germany), 79–93.
- Greiner B (2015) Subject pool recruitment procedures: Organizing experiments with ORSEE. *J. Econom. Sci. Assoc.* 1(1):114–125.
- Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. *J. Econom. Behav. Organ.* 3(4):367–388.
- Hardeweg B, Menkhoff L, Waibel H (2013) Experimentally-validated survey evidence on individual risk attitudes in rural Thailand. *Development Cultural Change* 61(4):859–888.
- Harrison GW, Rutström E (2008) Experimental evidence on the existence of hypothetical bias in value elicitation methods. *Handbook of Experimental Economics Results*, vol. 1 (Elsevier B.V., Amsterdam, Netherlands), 752–767.
- Harrison G, Lau M, Williams M (2002) Estimating individual discount rates for Denmark: A field experiment. *Amer. Econom. Rev.* 92(5):1606–1617.
- Hocking RR, Leslie R (1967) Selection of the best subset in regression analysis. *Technometrics* 9(4):531–540.
- Holt CA, Laury SK (2002) Risk aversion and incentive effects. *Amer. Econom. Rev.* 92(5):1644–1655.
- Jaeger DA, Dohmen T, Falk A, Huffman D, Sunde U, Bonin H (2010) Direct evidence on risk attitudes and migration. *Rev. Econom. Statist.* 92(3):684–689.
- List JA, Gallet CA (2001) What experimental protocol influences disparities between actual and hypothetical stated values? *Environ. Resource Econom.* 20(3):241–254.
- Murphy JJ, Allen PG, Stevens TH, Weatherhead D (2005) A meta-analysis of hypothetical bias in stated preference valuation. *Environ. Resource Econom.* 30(3):313–325.
- Perugini M, Gallucci M, Presaghi F, Ercolani A (2003) The personal norm of reciprocity. *Eur. J. Personality* 17(4):251–283.
- Selten R (1967) Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. Saueremann H, ed. *Beiträge zur experimentellen Wirtschaftsforschung* (J.C.B. Mohr (Paul Siebeck), Tübingen, Germany), 136–168.
- Slovic P (1995) The construction of preference. *Amer. Psychologist* 50(5):364–371.
- Spearman C (1904) The proof and measurement of association between two things. *Amer. J. Psych.* 15(1):72–101.
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. B* 58(1):267–288.

- Toubia O, Johnson E, Evgeniou T, Delquié P (2013) Dynamic experiments for estimating preferences: An adaptive method of eliciting time and risk parameters. *Management Sci.* 59(3):613–640.
- Tversky A, Simonson I (1993) Context-dependent preferences. *Management Sci.* 39(10):1179–1189.
- Vansteelandt S, Babanezhad M, Goetghebeur E (2009) Correcting instrumental variables estimators for systematic measurement error. *Statist. Sinica* 19:1223–1246.
- Vieider FM, Lefebvre M, Bouchouicha R, Chmura T, Hakimov R, Krawczyk M, Martinsson P (2015) Common components of risk and uncertainty attitudes across contexts and domains: Evidence from 30 countries. *J. Eur. Econom. Assoc.* 13(1):421–452.
- Vischer T, Dohmen T, Falk A, Huffman D, Schupp J, Sunde U, Wagner G (2013) Validating an ultra-short survey measure of patience. *Econom. Lett.* 120(2):142–145.
- Warren C, McGraw AP, Van Boven L (2011) Values and preferences: Defining preference construction. *Wiley Interdisciplinary Rev. Cognitive Sci.* 2(2):193–205.
- Weber E, Blais A-R, Betz N (2002) A domain-specific risk attitude scale: Measuring risk perceptions and risk behaviors. *J. Behav. Decision Making* 15(4):263–290.
- Ziegelmeier F, Ziegelmeier M (2012) Parenting is a risky business: Parental risk attitudes in small stakes decisions on their child's behalf. Unpublished manuscript.