Reassessing Qualitative Self-Assessments and Experimental Validation*

Jonathan Chapman
University of Bologna
jonathan.chapman@unibo.it
jnchapman.com

Pietro Ortoleva
Princeton
pietro.ortoleva@princeton.edu
pietroortoleva.com

Erik Snowberg
Utah, UBC, CESifo, NBER
snowberg@eccles.utah.edu
eriksnowberg.com

Leeat Yariv
Princeton, CEPR, CESifo, NBER
lyariv@princeton.edu
lyariv.com

Colin Camerer
Caltech
camerer@hss.caltech.edu
hss.caltech.edu/∼camerer/

March 19, 2025

Abstract

Accurate measures of economic preferences are essential for both testing economic theory and incorporating heterogeneity into macroeconomic models. Qualitative self-assessments of economic preferences have become increasingly popular as a method to elicit such preferences, especially outside of laboratory settings. The use of such self-assessments is often supported by experimental validation, a method that links them to choices in incentivized elicitations. We demonstrate theoretically that experimental validation may fail to produce reliable new measures. Empirically, analyzing data from over 13,000 participants across diverse samples, we document four key findings. First, qualitative self-assessments and traditional incentivized measures exhibit weak correlations, even when accounting for response noise. Second, qualitative self-assessments sometimes correlate more strongly with theoretically distinct incentivized elicitations than those for which they are intended to proxy. Third, relationships between qualitative self-assessments and various attributes—including geographical location, demographics, and behaviors—are unrelated to variation in incentivized elicitations. Fourth, qualitative self-assessments are no simpler for participants than incentivized elicitations: these questions show a common heuristic of extreme or midpoint responses, especially by individuals with lower cognitive ability.

JEL Classifications: C90, C91, C93, D90
Keywords: Self-Assessments, Risk Preferences, Time Preferences, Social Preferences, Preference Elicitation, Experimental Validation

*We are indebted to Travis Baseler, Roland Benabou, Anna Dreber, Simon Gächter, Andreas Grunewald, Florian Hett, Michael Kosfeld, Marta Kozakiewicz, Ted O'Donoghue, Noemi Peter, Zahra Sharafi, Charlie Sprenger, Stefan Trautmann, Jean-Robert Tyran, and Stephanie W. Wang for useful conversations and comments.

1 Introduction

Economic preferences drive individual decision-making and, in turn, shape aggregate economic outcomes. Accurately measuring these preferences is essential for explaining individual behavior (Dohmen et al., 2011), understanding national economic performance (Sunde et al., 2022), and incorporating heterogeneity into macroeconomic models (Bilbiie, forthcoming). Economists have typically elicited preferences using choices in incentivized elicitations. Over the past 15 years, however, there has been growing interest in using qualitative self-assessments of preferences as a replacement for the choice-based approach, particularly in large surveys (Falk et al., 2018).[1] This shift has been justified by experimental validations, which demonstrate that self-assessments predict choices in incentivized elicitations. For example, participants' certainty equivalents for a risky lottery are correlated with their responses to the question, "How willing are you to take risks, in general?" (Falk et al., 2023).[2] Nonetheless, it remains to be established whether experimental validation effectively identifies useful alternative measures. In particular, it is not yet clear whether experimentally-validated qualitative self-assessments yield findings comparable to those derived from incentivized elicitations.

In this paper, we examine the experimental validation method and its implications for qualitative self-assessments. We demonstrate theoretically that experimental validation may fall short of producing reliable new measures. Empirically, we evaluate qualitative self-assessments for six preference domains—risk, impatience, altruism, trust, reciprocity, and willingness to punish unfair behavior—by bringing together eight datasets, covering two rep-

[1] The impact of experimental validation and qualitative self-assessments on the economics literature has been substantial: collectively, the three primary papers on the topic have been cited over 9,000 times (Dohmen et al., 2011; Falk et al., 2018, 2023).

[2] While the term "experimental validation" has roots in psychology, Falk et al. (2023) appear to have popularized it in economics, providing what we believe to be the clearest and most developed implementation of the procedure. Falk et al. (2023) develop and use qualitative self-assessments for six preference domains—risk, impatience, altruism, trust, reciprocity, and willingness to punish unfair behavior. Researchers have also developed experimentally-validated measures for ambiguity attitudes (Cavatorta and Schröder, 2019), moral universalism (Enke et al., 2022), debt aversion (Albrecht and Meissner, 2022), competitiveness (Fallucchi et al., 2020; Buser et al., 2024), and preferences for truth telling (Schudy et al., 2024).

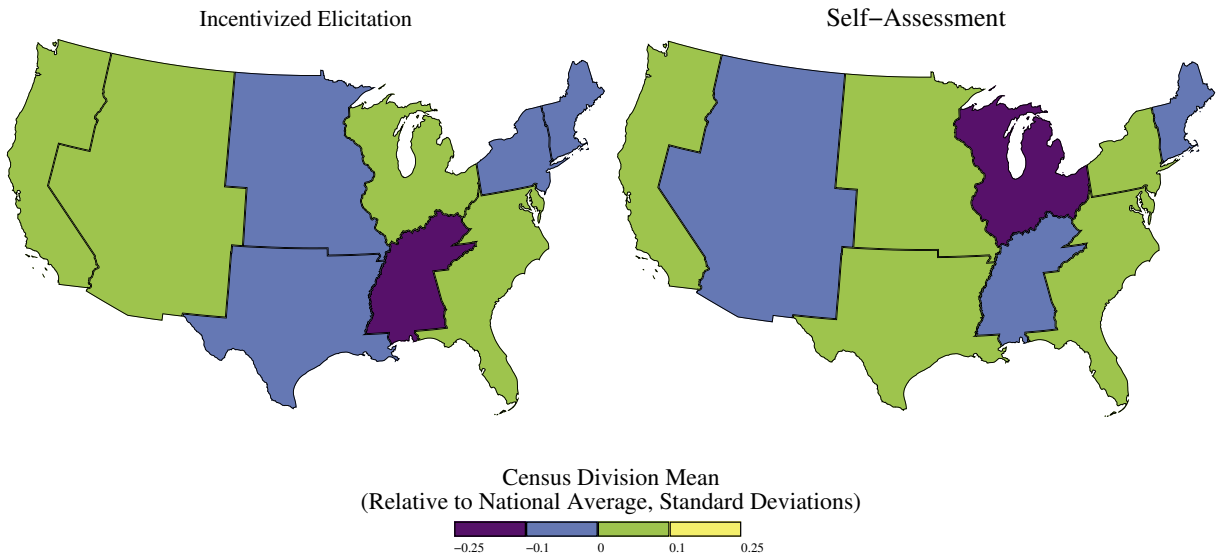
resentative samples of the U.S., three student samples, a large U.K. low-education sample, and two convenience samples (total $N \approx 13{,}000$). Each dataset contains both incentivized preference elicitations and experimentally-validated self-assessments drawn from the survey modules developed by Falk et al. (2023). We find smaller correlations between qualitative self-assessments and their incentivized counterparts than Falk et al.'s original experimental validation, in line with prior replications in student populations. We also illustrate that qualitative self-assessments are often correlated with multiple incentivized elicitations, making it difficult to determine which preferences, if any, these self-assessments are measuring. Indeed, we show that correlations between qualitative self-assessments and economic, demographic, and geographic variables are almost entirely due to confounding factors—that is, unrelated to variation in incentivized elicitations. Finally, we provide evidence that qualitative self-assessments may not be simpler for respondents than incentivized elicitations. In fact, the reliance on response heuristics is equally prevalent in qualitative self-assessments as in their traditional, choice-based counterparts.

The implications of these findings are evident from Figure 1. The left of Panel (a) displays how risk tolerance varies geographically in the U.S., using an incentivized elicitation. The map on the right of Panel (a) shows the geographic variation of the Preference Survey Module's (PSM; Falk et al., 2023) qualitative self-assessment of risk tolerance in our data.[3] The geographic patterns vary considerably between the two measures. For example, according to self-assessments, Upper Midwest residents appear to be highly risk-averse, while choices in incentivized elicitations suggest they are relatively risk-tolerant. This divergence is not unique to risk or to specific geographic units, as shown in Panel (b). The scatterplots in this panel compare U.S. states' rankings based on citizens' responses to traditional elicitations and qualitative self-assessments across the six domains of the PSM. As the plots reveal, the

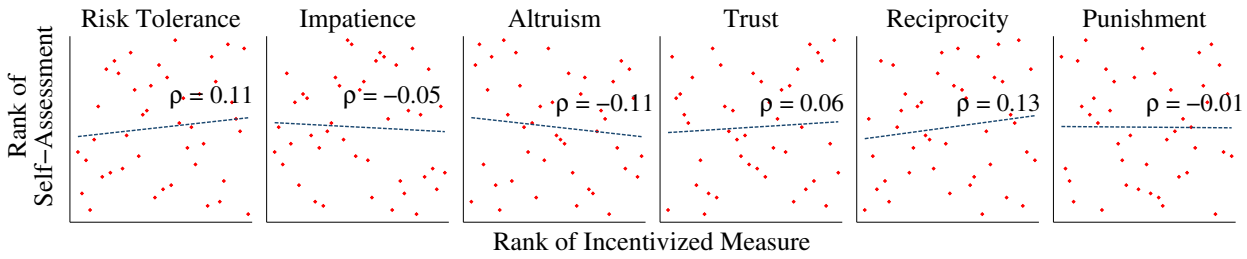
[3]To create these maps, we take the approach used by Falk et al. (2018) to plot the geographical distribution of preference indices, and apply it separately to incentivized elicitations and qualitative self-assessments. Specifically, we standardize each variable to a mean of 0 and a standard deviation of 1, then calculate the average of this standardized measure within each U.S. census division. Values below 0 indicate that the division's average is below the national mean. Appendix A includes similar maps to those in Panel (a) for impatience, trust, altruism, reciprocity, and punishment.

Figure 1: Self-assessments and incentivized measures identify different geographic distributions of risk preferences ($N = 4,950$).

(a) Estimated Risk Tolerance using Different Elicitations



(b) Ranking of preferences across U.S. states



Notes: The top panel of the figure plots the mean incentivized (left-hand side) and self-assessment (right-hand side) measures of risk tolerance for each census division, from online surveys of two representative samples of the U.S. population (U.S. Samples 1 and 2; see Table 2). The bottom panel uses the same data to plot the rank of the average self-assessment (y-axis) and incentivized measure (x-axis) at the state level. ρ is the Spearman rank correlation coefficient, and lines represent linear fits of the state-level data. States with fewer than 25 observations are excluded ($N = 43$).

correlations between these rankings are low and statistically insignificant for all six domains. For instance, knowing that residents of a state rank highly in self-reported altruism tells us little about how they would rank in actual giving behavior in a dictator game.

Why do two measures, which are highly correlated by economics standards, produce such divergent patterns? In Section 2, we show theoretically that a high correlation between two

measures does not render them interchangeable.[4] Specifically, it does not guarantee that the correlation between one measure and an auxiliary variable—such as income, education, or geographical location—will coincide, in either magnitude or sign, with the correlation between another highly correlated measure and the same auxiliary variable. This is due to the well-known fact that correlations are not transitive, a property that extends even to the signs of correlations.[5] Indeed, we show that even a correlation as high as 0.9 between two variables does not preclude the possibility that these variables could exhibit opposite-signed correlations with other variables.

We find lower correlations between incentivized elicitations and qualitative self-assessments than those reported in the PSM—in line with previous replications in student samples (Vieider et al., 2015; Kosfeld and Sharafi, 2024)—in Section 4. In contrast with earlier studies, we examine a broad range of populations, and correct for white noise measurement error in both the incentivized elicitations and qualitative self-assessments. Even with these corrections, the correlations we observe are on average around 2/3 of those of Falk et al. (2023). Further, the lower magnitudes of the correlations we find are not explained by particularly noisy responses in our data; for example, measurement error in the original data from the PSM is generally higher than in our samples.

In Section 5, we demonstrate that the correlations between qualitative self-assessments and auxiliary variables are largely due to confounding factors—they are primarily driven by variation in the qualitative self-assessment that is not associated with incentivized elicitations. First, we show that among the six qualitative self-assessments analyzed, four display stronger correlations with incentivized measures of constructs they were not designed to proxy for—even after correcting for measurement error—than with their incentivized counterpart. For example, the qualitative self-assessment of risk tolerance correlates more strongly with the incentivized altruism elicitation than with the incentivized risk elicitation.

[4]This is understood in econometrics, see McCallum (1972) and Wickens (1972). However, to the best of our knowledge, the literature on experimental validation has not drawn on these findings.

[5]For example, suppose $a = u + v$, $b = v + w$, and $c = u - w$, where u, v, and w are independent random variables. Then, $\text{Corr}[a, b] > 0$, $\text{Corr}[a, c] > 0$, while $\text{Corr}[b, c] < 0$.

Second, we illustrate that while qualitative self-assessments are robustly correlated with a range of demographic, economic, and self-reported behavioral variables, these correlations are virtually unchanged even after controlling for all the incentivized elicitations. That is, the variation in qualitative self-assessments associated with these variables is unrelated to any incentivized elicitation.[6]

One argument in favor of qualitative self-assessments is their presumed simplicity: after all, they ask participants to describe their own traits using relatively colloquial language. In Section 6, we show that qualitative self-assessments appear equally susceptible to response biases as incentivized elicitations, challenging the idea that they are inherently more intuitive for participants. Specifically, we analyze the use of response heuristics—the choice of 0, 5, or 10 on an 11-point scale—in qualitative self-assessments, and show that low cognitive ability participants tend to rely on these heuristics far more than high cognitive ability participants or those in student samples. In addition, a comparison of samples in the U.S. and the U.K. suggests that cultural differences shape interpretations of qualitative self-assessments.

We conclude, in Section 7, by discussing the implications of our findings for preference elicitation. Our data demonstrate that it is feasible to implement classical incentivized elicitations in large samples. We are hopeful that choice-based elicitations with hypothetical incentives may provide a viable alternative for researchers seeking to avoid the use of monetary incentives. However, further research is required to thoroughly assess the usefulness of such measures.

In summary, our findings underscore the need for more robust methods to validate new measures. Specifically, our results do not provide much support for the use of experimentally-validated qualitative self-assessments as proxies for choice-based preference measures. While

[6]This is in spite of the fact that incentivized elicitations are often correlated with the same variables, indicating that the self-assessments fail to capture correlations of interest. The qualitative self-assessments and incentivized elicitations yield quite different patterns of correlations between preferences and other individual characteristics—see Appendix Figure A.3. This is particularly true of correlations with cognitive ability, where self-assessments suggest that higher cognitive ability participants are less risk tolerant and more impatient, contrary to what previous literature suggests (see, for instance, Dohmen et al., 2010; Snowberg and Yariv, 2021).

qualitative self-assessments may still offer meaningful information about individuals' traits, their precise meaning remains unclear. We follow the convention in the experimental validation literature by treating incentivized elicitations as the gold standard. Incentivized elicitations are rooted in the revealed preference paradigm central to economics and directly connect to theoretical models—for instance, risk aversion is linked to the certainty equivalent of a lottery, which maps to utility curvature. Although qualitative self-assessments may capture other aspects of risk preferences, it is uncertain whether these reflect economic constructs or extraneous factors.

2 A Simple Theory of Experimental Validation

A simple theory is useful to show how the use of experimentally-validated proxies can produce misleading results. Experimental validation aims to find a proxy variable q (for qualitative self-assessment) for a preference variable p, which is measured by an incentivized elicitation. The goal is to use q to estimate the relationship between some y—for example, income, education, or responses to an experimental treatment—and p when it may be difficult to measure p directly.[7] To simplify exposition, we assume that these variables have already been standardized, that is $\mathbb{E}[p] = \mathbb{E}[y] = \mathbb{E}[q] = 0$, and $\text{Var}[p] = \text{Var}[y] = \text{Var}[q] = 1$.[8]

Experimental validation ensures that q is (perhaps highly) positively correlated with p. Both p and q may be correlated with y. Thus, we model q as:

$$q = (1-\gamma)p + \gamma\eta = (1-\gamma)p + \gamma(\delta\eta_y + (1-\delta)\eta_p)$$

in which $\gamma, \delta \in [0,1]$ are unknown parameters. Here, η is a confounding factor, which captures variation in q that is not directly linked to p. The confounding factor η can be decomposed

[7] White noise error could be added to either y or p, but we omit this for simplicity. Adding this complication simply means that in empirical applications it will be useful to have an instrument for y, which is easily obtained, see Gillen et al. (2019).

[8] As our empirical specifications use standardized versions of continuous variables, this is also consistent with our analysis.

into two components, η_p and η_y, with the property that $\text{Cov}[\eta_p, y] = \text{Cov}[\eta_y, p] = 0$. If $\text{Cov}[\eta_y, y] = 0$, then η is white noise measurement error (with respect to y). Of course, these associations are specific to a particular variable y and may not generalize: the confounding factor may represent white noise measurement error with respect to income, but not with respect to education.

If both y and p are measured, we can obtain the desired regression coefficient:

$$y = \beta_p p + \varepsilon_1 \qquad \Rightarrow \qquad \beta_p = \text{Cov}[y, p]\,.^9$$

Estimating the same regression with the experimentally-validated proxy variable yields:

$$y = \beta_q q + \varepsilon_2 \qquad \Rightarrow \qquad \beta_q = (1-\gamma)\text{Cov}[y, p] + \gamma\delta\text{Cov}[\eta_y, y]\,.$$

This equation shows both the potential and pitfalls of experimental validation.

Experimental validation produces a useful proxy when $\gamma\delta\text{Cov}[\eta_y, y] = 0$. If, in addition, $\gamma = 0$, then $\beta_q = \beta_p$. If, instead, $\gamma > 0$ (maintaining $\gamma\delta\text{Cov}[\eta_y, y] = 0$), then the confounding factor in q is orthogonal to y and acts as white noise measurement error. This measurement error attenuates the estimated coefficient from the second regression so that $0 < |\beta_q| < |\beta_p|$. This attenuation can be corrected by using a second experimentally-validated variable q' as an instrument, as long as the confounding factor in q' is orthogonal to y and to the measurement error in q.

When $\gamma\delta\text{Cov}[\eta_y, y] > 0$, experimental validation may produce a poor proxy: β_p and β_q may be of different signs, even if p and q are highly correlated. This can occur even if no part of the correlation between p and q is driven by the confounding factor in q, namely when $\delta = 1$. For example, if $q = 0.9p + 0.1\eta_y$ then the correlation between p and q is 0.9. However, if $\text{Cov}[y, p] = \sqrt{1/20} \approx 0.22$ and $\text{Corr}[\eta_y, y] = -\sqrt{19/20} \approx -0.975$, then

[9] No constant is required in the regression, as $\mathbb{E}[y] = \mathbb{E}[p] = 0$ through our standardization. Moreover, variance does not appear in the denominator of the expression for the regression coefficient due to standardization; that is, $\beta_p = \text{Cov}[y, p] = \text{Corr}[y, p]$.

$\beta_q \approx -0.22 < 0 < 0.22 \approx \beta_p$.[10]

This simple framework demonstrates that even a very high correlation between an incentivized elicitation and a qualitative self-assessment in an experimental validation is insufficient to ensure that the latter is a valid proxy. The remainder of this paper presents evidence that these issues undermine inference when employing the most popular qualitative self-assessments. We first demonstrate, in Section 4, that 65–100% of each qualitative self-assessment we examine can be attributed to confounding factors—η_y and/or η_p—that is, it is not correlated with the relevant incentivized elicitation.[11] Further, as we show in Section 5, the qualitative self-assessments we investigate are as—and sometimes more—highly correlated with other incentivized elicitations than with their target incentivized elicitation. In fact, correlations between qualitative self-assessments and variables of interest—income, cognitive ability, and so on—seem almost entirely due to confounding factors in qualitative self-assessments. That is, for all six of the qualitative self-assessments that we consider, $\gamma \approx 1$. Finally, in Section 6, we present evidence suggesting one possible confounding factor that is correlated with outcome variables (η_y). Specifically, we show that participants rely on heuristics when answering qualitative self-assessments. Lower cognitive ability participants are more likely to choose focal values at the extremes and middle of numerical response scales. Thus, responses to qualitative self-assessments are correlated with cognitive ability for reasons unrelated to economic preferences. As cognitive ability is related to many economic outcomes, this creates a plausible pathway for spurious correlations between qualitative self-assessments and economic outcomes.

[10]The value of $\text{Corr}[\eta_y, y]$ in the example is the maximum consistent with $\text{Cov}[y, p] = \sqrt{1/20}$. Due to the normalization of q, $\text{Var}[\eta_y] = 19$, and so $\text{Cov}[\eta_y, y] = \sqrt{19^2/20} \approx 4.25$. Assuming that p, y, and η_y are normally distributed, then, with only 300 observations, both $\widehat{\beta}_p$ and $\widehat{\beta}_q$ will be statistically significantly different than zero at the 0.05 level ∼97.5% of the time.

[11]The lower bound assumes that $\delta = 1$. If $\delta < 1$, then q may be entirely a combination of η_p and η_y, regardless of the correlation between p and q. For example, if $\text{Corr}[p, q] = x$, it can be rationalized by $\gamma = 1$, $\delta = 1 - x/y$, and $\text{Cov}[\eta_p, p] = y$, with $y \in [x, 1]$.

3 Measures and Datasets

Our data are drawn from a range of studies that examined economic preferences in various participant populations. Each study included both incentivized preference elicitations and qualitative self-assessments. They differ in the preference domains they considered. These studies were carried out between 2014 and 2021, and contain a total of 13,157 observations.

3.1 Measures

The central measures are qualitative self-assessments and the incentivized elicitations for which they are meant to proxy. We describe the most common way that the measures were implemented. Deviations from these descriptions are detailed dataset-by-dataset in the next subsection. Screenshots showing specific questions can be found in Appendix B.2.

Qualitative Self-Assessments: We examine qualitative self-assessments across the six domains studied in Falk et al. (2018) and Falk et al. (2023): risk tolerance (sometimes referred to simply as "risk"), impatience, altruism, trust, reciprocity, and punishment. Participants were asked to rate themselves on an 11-point scale, from 0 to 10, for each of these domains. Question wording can be found in Table 1. All the self-assessment measures were taken from an early draft of Falk et al. (2023), which was available in 2014 when we first included qualitative self-assessments in our studies. Most questions from that early draft were later incorporated into the Global Preference Survey (GPS), with the exception of the punishment item, which was split into two parts.[12] The impatience question we use is included in the

[12]The GPS dataset has been used to address a range of questions, including the origins of preferences (Becker et al., 2020; Cao et al., 2021), the relationship between patience and economic development (Sunde et al., 2022), gender gaps in preferences (Falk and Hermle, 2018); behaviors during the COVID-19 pandemic (Chan et al., 2020b,a; Campos-Mercade et al., 2021); how economic preferences affect trade outcomes (Korff and Steffen, 2022); the contribution of patience and risk preferences to human capital investments (Hanushek et al., 2022); the long-term effect of transport networks on economic integration (Flückiger et al., 2022); the effects of cultural origin on entrepreneurship (Jonsson and Ouyang, 2023); the relationship between inequality and risk preference (Pickard et al., 2024), and others. The World Bank has drawn on the GPS to evaluate national entrepreneurial characteristics (Clemente et al., 2019).

Table 1: Qualitative Self-Assessments and Incentivized Elicitations

		PSM	GPS	GPS Survey
Risk Tolerance				
Qualitative:	How do you see yourself: are you a person who is generally willing to take risks or do you try to avoid taking risks?	✓	✓	✓
Incentivized:	Certainty equivalent of a lottery (as % of expected value)	✓		
Impatience				
Qualitative:	How well does the following statement describe you as a person? "I tend to postpone things even though it would be better to get them done right away."			✓
Incentivized:	Amount needed today to forgo a payment in the future	✓		
Altruism				
Qualitative:	How would you assess your willingness to share with others without expecting anything in return, for example your willingness to give to charity?	✓	✓	✓
Incentivized:	Amount sent in a dictator game (as % endowment)	✓		
Trust				
Qualitative:	As long as I am not convinced otherwise I always assume that people have only the best intentions.	✓	✓	✓
Incentivized:	Amount sent by first mover in a Trust Game (as % endowment)	✓		
Reciprocity				
Qualitative:	How would you assess your willingness to return a favor to a stranger?		✓	✓
Incentivized:	Amount returned by second mover in a trust game (as % received)	✓		
Punishment				
Qualitative:	Are you a person who is generally willing to punish unfair behavior even if this is costly?	✓	✓	✓
Incentivized:	Amount used to punish receiver returning 0 in trust game (as % endowment)			

Notes: PSM refers to Falk et al. (2023), GPS to Falk et al. (2018), and GPS Survey to the survey documents found at https://gps.iza.org/downloads. In contrast to our measure, the GPS measure of reciprocity does not specifically reference "a stranger."

survey documents for the GPS, and perhaps the surveys themselves.[13]

Incentivized Elicitations: These measures are standard in the literature and very similar to those used by Falk et al. (2023), except for the punishment elicitation. Duplicate elicitations of risk and impatience were included in all datasets, and most datasets also included a single incentivized elicitation of the four social preferences: altruism, trust, reciprocity, and punishment. Table 1 provides brief descriptions of these elicitations, with screenshots in Appendix B.2. All variables are coded so that higher values consistently indicate greater levels of the targeted preference; for example, higher risk tolerance is reflected by either a higher self-assessed willingness to tolerate risk, or a higher certainty equivalent for a lottery. For punishment, participants were presented with a trust game setting in which the sender sends the full amount, and the receiver returns nothing. Participants then decided how much of a given stock of experimental points they would like to spend to punish the receiver, with each point spent reducing the receiver's payoff by six points. In contrast, Falk et al. (2023), measured "negative reciprocity," in which punishment is inflicted on the opponent in either a prisoner's dilemma or an ultimatum game. The only other noteworthy difference is that, in their altruism measure, dictator giving benefited a charity, whereas ours directed the gift to another survey participant.

Cognitive Ability: Each survey measured participants' cognitive ability using a set of nine questions. Six questions were taken from the International Cognitive Ability Resource (ICAR, Condon and Revelle, 2014). Three of these questions were similar to Raven's Matrices, and the other three ask participants to determine which of several images displays a rotation of a given shape. The survey also contained the Cognitive Reflection Test (CRT; Frederick, 2005), which includes three arithmetic questions with an instinctive, but incorrect,

[13] GPS survey documents were obtained from https://gps.iza.org/downloads. Specifically, the question appears to have been included as item WP13426, and described in the survey document for every language and country, but is not mentioned in either Falk et al. (2018) or the published version of Falk et al. (2023).

answer. The cognitive ability score is the sum of correct answers to these nine questions.[14]

Individual Characteristics: Our representative samples contained measures of demographic and economic characteristics. We used sex, age, education, income, and a binary measure of whether an individual owns stocks or shares. In addition, we elicited self-reported measures of participants' health behaviors and other activities for a subset of respondents in U.S. Sample 2. We use this information in Appendix A.4.

3.2 Datasets

Table 2 provides high-level descriptions of the datasets we use. We refer readers to the papers that collected these datasets for details beyond what is essential for our discussions.

Falk et al. (2023): This dataset is the point of comparison for several analyses, and taken from the replication package of Falk et al. (2023). The 409 participants in this dataset—University of Bonn students who were recruited by the experimental laboratory—were divided into two groups of roughly equal size. The first group was initially presented with incentivized elicitations of risk and time and a battery of survey questions—self-assessments and hypothetical incentivized questions—about social preferences. One week later, the same group was presented with incentivized elicitations of social preferences and survey questions regarding risk and time preferences. The second group of participants completed these tasks in reverse order, also one week apart. Due to missing values, there are 360 observations for correlations in reciprocity and punishment, and 382 for correlations in other domains. Qualitative elicitations with the strongest predictive power for incentivized measures were chosen to form part of what the authors term the Preference Survey Module (PSM), a version of which was used in the Global Preference Survey (GPS) of Falk et al. (2018).

[14]Falk et al. (2018) use a qualitative self-assessment of mathematical ability as a proxy for cognitive ability. They ask participants to rate their agreement, from 0 to 10, with the statement "I am good at math."

Table 2: Datasets

			Main Measures	
	N	Year	Preferences (# duplicates)	Other
Falk et al. (2023) (Student)	409	2010–11	All Self-Assessments (1) Incentivized Risk, Time (2) Incentivized Social (2)	
U.S. Sample 1 (Representative)	1,950	2018–19	All Self-Assessments (2) Incentivized Risk, Time (2) Incentivized Social (2)	Demographics Cognitive Ability
U.S. Sample 2 (Representative)	3,000	2015–16	All Self-Assessments (1) Incentivized Risk, Time (2) Incentivized Social (1)	Demographics Cognitive Ability Behavior SA
U.K. Sample (Low-Education)	1,984	2017	All Self-Assessments (1) Incentivized Risk, Time (2) Incentivized Social (1)	Demographics
Caltech (Student)	3,266	2014–16	Risk, Time, Altruism SA (2) Incentivized Risk, Altruism, Time (2)	
Pitt (Student)	437	2021	Altruism, Trust, Reciprocity SA (1) Incentivized Risk (2)	
UBC (Student)	202	2019	Risk, Altruism SA (2) Incentivized Risk, Altruism (2)	
Mechanical Turk (Convenience)	2,318	2016, 2019	Risk, Altruism SA (2) Incentivized Risk, Altruism (2)	

Notes: SA stands for "self-assessment."

3.2.1 General Population Datasets

All of our general population samples, as well as the Pitt student sample introduced below, were surveyed online by YouGov, using a similar survey implementation. Appendix B.1 contains further details and screenshots.

U.S. Sample 1: This representative sample of the U.S. comes from Chapman et al. (2024c). This is our main dataset as it repeatedly surveyed the same participants, making it possible to more closely mimic Falk et al. (2023). It also contained duplicate measures of all qualitative self-assessments and incentivized elicitations, allowing us to examine whether weak correlations are explained by classical measurement error.

U.S. Sample 2: There are two samples in this dataset. The first is a representative sample of the U.S. from Chapman et al. (2024a). The second is a representative sample of the U.S. from Chapman et al. (2023). These samples contain compatible versions of the measures used in this paper, so we combine them into a single dataset.

U.K.: This sample, from Barcellos et al. (2024), consists of low-education U.K. residents. Survey respondents are U.K. residents who had attended school in the U.K., left school at or before 16, and were born between September 1, 1954 and August 31, 1960.

3.2.2 Student and Convenience Sample Datasets

Caltech: The Caltech dataset combines the Fall 2014, Spring 2015, Fall 2015, and Spring 2016 waves of the incentivized surveys within the Caltech Cohort Study used in Gillen et al. (2019) and Jackson et al. (2023). It also includes the Caltech lab responses from Snowberg and Yariv (2021). Only the Spring 2016 wave includes a qualitative self-assessment of altruism ($N = 605$). Qualitative self-assessments of impatience do not appear in the Fall 2014 and Spring 2016 waves, leaving 1,778 observations for this measure. All impatience elicitations used hypothetical incentives. We treat all of these samples as a unified dataset, despite some repeated observations of individuals.

UBC: This lab sample of University of British Columbia students comes from Snowberg and Yariv (2021). The study itself is similar to those run within the Caltech Cohort Study.

Pitt: This is an incentivized survey of participants recruited from the Pittsburgh Experimental Economics Laboratory (PEEL) taken from Chapman et al. (2024b). The survey is similar to those in U.S. Sample 1 and 2.

Mechanical Turk: These two samples of Mechanical Turk workers are taken from Snowberg and Yariv (2021). The latter sample used incentive levels set at half those of the earlier sample, but the two are otherwise identical, and we analyze them as a combined dataset.

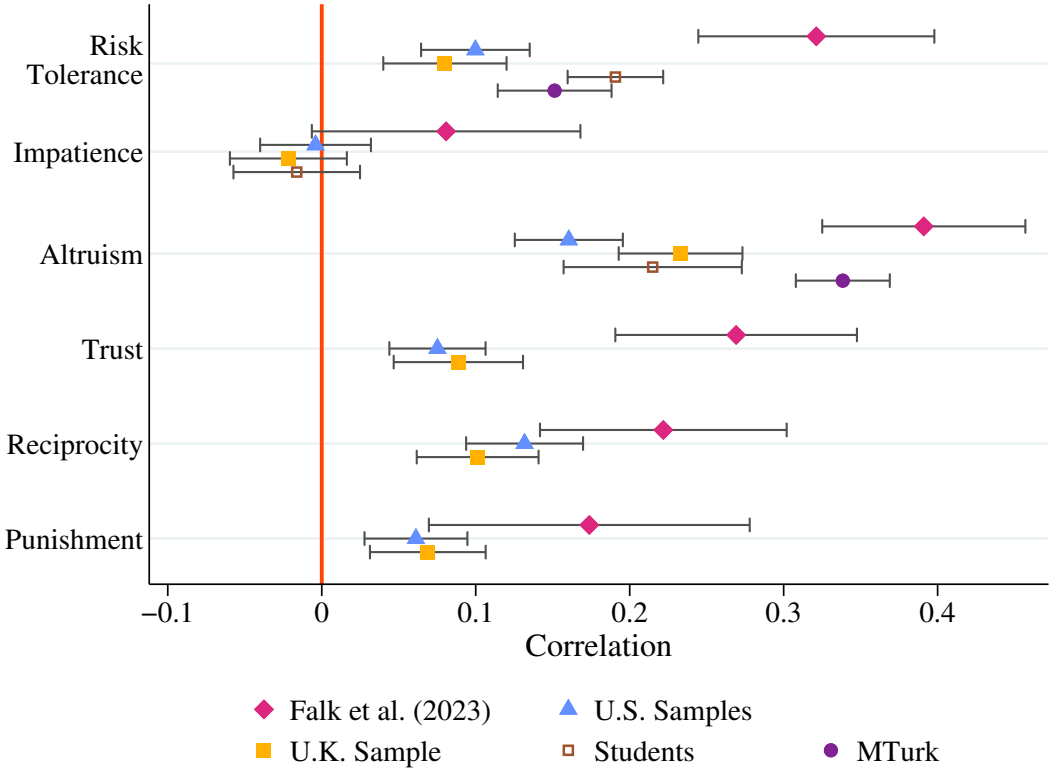
4 Experimental Validation in General Populations

In line with previous experimental validation exercises, we find significant positive correlations between qualitative self-assessments and incentivized elicitations in five of six preference domains examined in the PSM/GPS. However, the magnitudes are consistently smaller than found in Falk et al. (2023), indicating that qualitative self-assessments have relatively little explanatory power for choices in incentivized elicitations. Further, this pattern holds in all our datasets, across a range of subgroups within our representative samples, and does not appear to stem from measurement error.

Our datasets consistently reveal positive, yet modest, correlations between most qualitative self-assessments and incentivized elicitations, as shown in Figure 2. This implies that a relatively large share of the variation in qualitative self-assessments can be attributed to confounding factors. In the language of the model presented in Section 2, $1-\gamma = \text{Corr}[p, q]$, so the confounding factor's share of q is $1 - \text{Corr}[p, q]$, translating to roughly 60-90% for the Falk et al. (2023) data, and 80–100% in our U.S. samples. In fact, we observe smaller correlations than those in Falk et al. (2023) in each of our datasets. There is some evidence that correlations may be slightly higher in convenience samples than in broader samples, particularly for risk tolerance, but the observed differences are, in any case, small. This suggests that small correlations are not specific to general population samples.[15]

[15]Appendix A.2 shows that the magnitude of the correlations is similar when estimating Spearman cor-

Figure 2: Qualitative self-assessments and incentivized elicitations exhibit low correlations in all datasets.



Notes: The figure compares correlations between qualitative self-assessments and incentivized elicitations across different samples. To ensure comparability, the correlations for each sample, except those for Falk et al. (2023), use a single self-assessment question, a corresponding incentivized elicitation for social preferences, and an average of two incentivized elicitations for risk tolerance and impatience. Correlations for Falk et al. (2023) also use the average of two incentivized elicitations for trust, reciprocity, and punishment. Bars represent 90% confidence intervals.

The smaller correlations we observe are unlikely to be explained by attenuation due to classical measurement error, as data from our samples are not unusually noisy. Table 3 estimates the extent of classical measurement error by utilizing the duplicate measures of each preference available in our datasets.[16] As demonstrated in Gillen et al. (2019), the proportion of variation in a given measure within a dataset that is due to measurement error can be calculated as one minus the correlation between its duplicate measures. We report

relations, as well as when using alternative incentivized elicitations, or when focusing on subgroups of our representative samples.

[16]The duplicate incentivized elicitations differed in incentive levels and varied in other fine details. The duplicate qualitative self-assessments were identical, but asked at different points in the same survey.

this proportion for different measures and samples in Table 3.[17]

The level of measurement error in U.S. Sample 1 compares favorably to those of North American university students. In particular, U.S. Sample 1 has a lower level of measurement error than university students for the risk tolerance and impatience incentivized elicitations, and higher for the altruism incentivized elicitation.[18] Moreover, U.S. Sample 1 appears to have lower levels of measurement error than Falk et al. (2023) for qualitative questions, as shown in the last two columns of Table 3. U.S. Sample 1 repeats the same qualitative self-assessments about 20 minutes apart. Falk et al. (2023) ask nearly identical self-assessments in a single survey. From a large number of similar self-assessments, we selected a pair that seemed most alike, based on their variable labels. For example, for impatience, we use v158: "I often regret decisions that I make" and v162: "I make decisions I later regret."[19] Table 3 reveals another important pattern: qualitative self-assessments exhibit measurement error levels comparable to those of incentivized elicitations, challenging the notion that qualitative self-assessments are simpler to complete. We return to this point in Section 6.

The difference between our results and those of Falk et al. (2023) also does not appear to stem from differences in implementation, as evidenced in Figure 3. We mimic Falk et

[17]Formally, suppose X^a and X^b are two measures of the same underlying preference x^*. Classical measurement error implies that $X^a = X^* + \nu_X^a$ and $X^b = X^* + \nu_X^b$, with ν_X^a, ν_X^b i.i.d. random variables, and $\mathbb{E}\big[\nu_X^a \nu_X^b\big] = 0$. If we assume that $\frac{\text{Var}[\nu_X^a]}{\text{Var}[X^a]} = \frac{\text{Var}\big[\nu_X^b\big]}{\text{Var}[X^b]} \equiv \frac{\text{Var}[\nu_X]}{\text{Var}[X]}$, then

$$\widehat{\text{Corr}}[X^a, X^b] \to_p \text{Corr}[X^a, X^b] = \frac{\sigma^2_{X^*}}{\sigma^2_{X^*} + \sigma^2_{\nu_X}}.$$

Thus, $1 - \widehat{\text{Corr}}[X^a, X^b]$ captures the proportion of the measure's variation that is due to measurement error.

[18]Measurement error was similar in U.S. Sample 2—29% (s.e. = 2.2%) for risk tolerance, and 25% (3.0%) for impatience. Similarly, in the altruism domain, we observe measurement error in the qualitative self-assessments of 40% (1.7%) in the MTurk sample, and 44% (3.4%) among Caltech students. The duplicate qualitative self-assessment in these two samples asks if they would "go out of [their] way to do something nice for a stranger." The replication dataset of Falk et al. (2023) does not include the duplicate incentivized measures underlying the included average.

[19]For other domains, the questions are as follows. For risk, v104: "I like risky things," and v105: "I like taking risks." For altruism, v10: "Willingness to give to charities," and v12: "Willing to spend for charity even if I don[']t benefit." For trust: v72 "Willingness to trust," v83 "Compared to others, I easily trust people." For reciprocity, v53: "Willingness to reward a favor," and v64: "If someone does me a favor, I am willing to return it." For punishment, v36: "If someone puts me in a diff[icult] position, I'll do the same," and v38: "If someone intentionally harms me, I'll do the same to him/her."

Table 3: Measurement Error in Different Samples

	Incentivized Measures					Self-Assessments	
			Students				
	Sample 1	MTurk	Caltech	UBC	Pitt	Sample 1	Falk et al.
Risk Tolerance	25% (2.2)	35% (1.9)	26% (1.6)	31% (5.3)	29% (3.6)	18% (1.6)	35% (3.9)
Impatience	16% (1.9)		20% (2.7)			13% (1.4)	36% (4.0)
Altruism	44% (3.0)	20% (1.2)	14% (3.3)	26% (4.8)		27% (2.1)	35% (3.5)
Trust	39% (2.6)					18% (1.5)	35% (3.3)
Reciprocity	28% (1.9)					25% (3.2)	60% (4.8)
Punishment	33% (1.9)					33% (2.8)	28% (3.0)
N	1,950	2,318	3,001*	202	437	1,950	397†

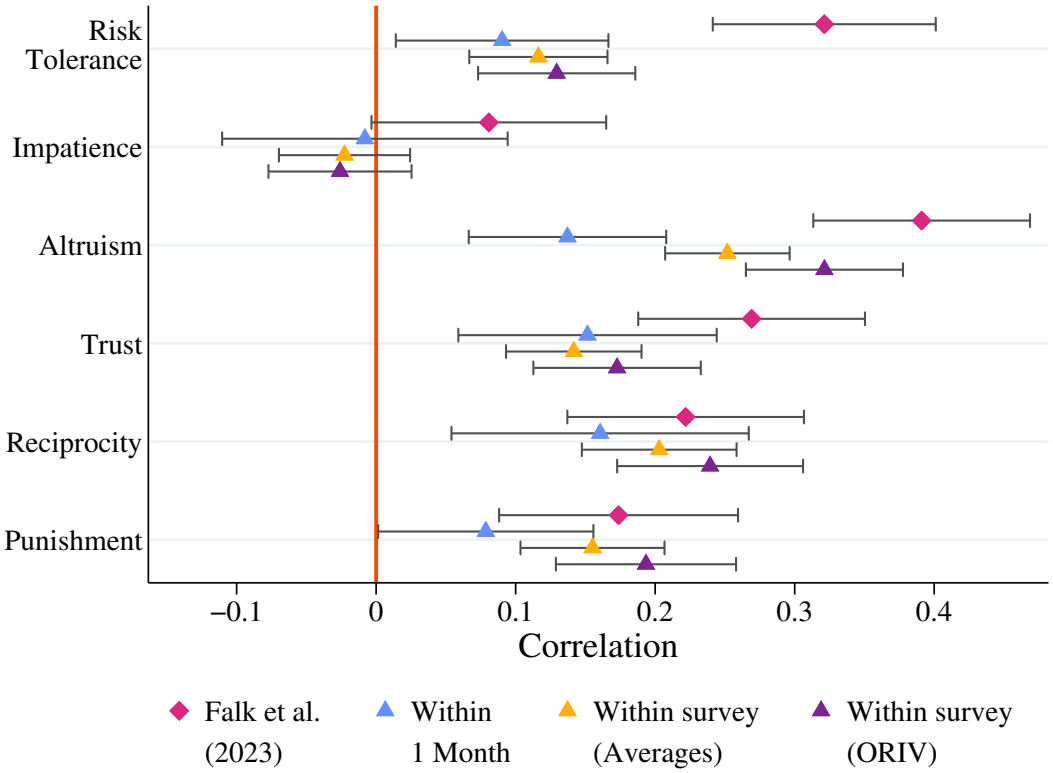
Notes: U.S. Sample 1 included two identical elicitations of each self-assessment. For Falk et al. (2023) we estimate measurement error using two self-assessments which appear most similar based on the variable labels in their replication dataset. *: $N = 3,001$ for risk tolerance, $N = 1,778$ for impatience, and $N = 605$ for altruism. †: the number of observations ranges from $N = 373$ (for reciprocity and punishment) to $N = 397$ (for trust and altruism). Bootstrapped standard errors in parentheses.

al.'s (2023) design using our U.S. Sample 1, the dataset most similar to theirs. Falk et al. (2023) collect responses to the same qualitative self-assessments and incentivized elicitations measured one week apart, with duplicate elicitations for five of the six incentivized elicitations. We approximate this design by correlating the responses received in the initial survey in U.S. Sample 1, and those received at any point in the following five weeks—a total of 480 respondents.[20] Both the initial and follow-up surveys contained qualitative self-assessments and incentivized measures. Thus, each participant generates two data points, and we cluster standard errors at the individual level.[21] The correlations are markedly smaller than those

[20]U.S. Sample 1 is a multi-wave study in which the first survey had 1,950 participants, and then, each week, 150 randomly-chosen participants from the first survey were invited to complete the same survey.

[21]That is, we correlate qualitative self-assessments in the first survey with incentivized elicitations in the latter survey, and incentivized elicitations in the first survey with qualitative self-assessments in the latter

Figure 3: Correlations are weak even after accounting for measurement error.



Notes: The "Within 1 month" series mimics the procedure used by Falk et al. (2023) as closely as possible in our datasets in U.S. Sample 1 ($N = 480$). As in Falk et al. (2023), this series uses one elicitation of each self-assessment, and—with the exception of altruism—the average of two elicitations for each incentivized measure. The "Within Survey (Averages)" and "Within Survey (ORIV)" series present correlations within the initial survey from U.S. Sample 1, first by averaging measures, and then using the ORIV technique to correct for measurement error. Bars represent 90% confidence intervals.

reported by Falk et al. (2023) and, as might be expected, smaller than the within survey correlations displayed above. The impatience correlation is close to zero, and for punishment, it is statistically indistinguishable from zero at the 5% level.

Figure 3 also shows that the correlations between qualitative self-assessments and incentivized elicitations remain small even after accounting for possible attenuation due to classical measurement error. This figure includes, as the final set of estimates, the correlations derived by instrumenting one duplicate with the other, using ORIV (Gillen et al., 2019). This approach eliminates the attenuating effects of classical measurement error on

survey, while clustering standard errors by individual.

correlations. Correspondingly, for altruism, reciprocity, and punishment—but not risk, time, and trust—this brings our estimated coefficients substantially closer to those found in Falk et al. (2023). However, this comparison likely underestimates the discrepancy between our results and those of Falk et al. (2023), as their data does not allow for a similar correction for measurement error—which appears to be just as prevalent in their dataset as in ours.[22] More importantly, the magnitude of correlations implies that qualitative self-assessments can explain, at best, around one-third of the variation in the incentivized elicitations. That is, adjusting for white noise measurement error suggests that confounding factors (η) explain 65–100% of the variation in qualitative self-assessments.

In sum, this section broadly replicates the experimental validation of Falk et al. (2023), and other papers, within a general population sample. Except for patience, we observe small, positive correlations between qualitative self-assessments and incentivized elicitations in both general populations and convenience samples. However, in contrast to those earlier studies, we demonstrate that small correlations are not explained by classical measurement error. This poses a major threat to inference, as most of the variation in qualitative self-assessments is not explained by choices in the corresponding incentivized elicitation.

5 Experimentally-Validated Measures May Not be Valid

Across all six domains, qualitative self-assessments are correlated with multiple incentivized elicitations. In only two of these domains is the largest correlation with the analogous incentivized elicitation. In addition, qualitative self-assessments are robustly correlated with demographic and economic variables in our data, as in some prior studies (for example, Dohmen et al., 2011). Yet, these correlations are not explained by the variation in a qualita-

[22]The larger correlations observed between qualitative self-assessments and incentivized elicitations in Falk et al. (2023) are likely due to their approach. Specifically, their method selects the linear combination of survey items that best predicts choices in incentivized elicitations. The questions chosen using this procedure often do not exhibit statistically significantly larger correlations with incentivized elicitations than many that are not chosen. Thus, it is unsurprising that questions selected using this approach show smaller correlations with incentivized elicitations in other datasets.

tive self-assessment associated with *any* of the six incentivized elicitations. Thus, it is challenging to interpret the correlations between qualitative self-assessments and demographic or economic variables.

5.1 Strong Correlations with Multiple Incentivized Elicitations

Qualitative self-assessments are often related to multiple incentivized behaviors, including many that are theoretically orthogonal to the concept they are trying to capture, as shown in Table 4. This table displays the complete ORIV correlation table between all incentivized elicitations and qualitative self-assessments in U.S. Sample 1. As can be seen, the qualitative self-assessment of risk tolerance is more highly correlated with the incentivized elicitations of altruism, trust, and punishment than with the incentivized elicitation of risk tolerance. Similar patterns hold for other qualitative self-assessments. Perhaps most disturbing, the qualitative self-assessment of impatience is slightly negatively correlated with the incentivized elicitation for impatience, but statistically significantly correlated with altruism and trust. This smorgasbord of correlations means that inferences that associate qualitative self-assessments with specific preference domains may well be misleading.[23]

5.2 Associations with Demographic and Economic Characteristics

Nearly all of the variation in qualitative self-assessments associated with demographic and economic characteristics is independent of the variation in incentivized elicitations. To interpret the findings that demonstrate this, it is helpful to revisit our model from Section 2. In that section, p denoted an incentivized elicitation, y a demographic (or other) variable. We modeled qualitative self-assessments as $q = (1 - \gamma)p + \gamma\eta$, in which the confounding factor

[23]The experimental data of Falk et al. (2023) also exhibit considerable variation in correlations across domains. As they do not present an analogue for Table 4, we provide it in Appendix Table A.2. In Appendix Tables A.4 and A.5 we also show that correlations between incentivized elicitations, or between qualitative self-assessments, display different patterns than the cross-correlations, suggesting the patterns in Table 4 are not due to inherent linkages between attributes.

Table 4: Correlations between Qualitative Self-Assessments and Incentivized Elicitations, using ORIV

		Incentivized Elicitations					
		Risk Tolerance	Impatience	Altruism	Trust	Reciprocity	Punishment
Qualitative Self-Assessment	Risk Tolerance	0.13*** (.034)	0.01 (.034)	0.19*** (.038)	0.15*** (.036)	0.07* (.035)	0.14*** (.032)
	Impatience	−0.01 (.033)	−0.03 (.031)	0.11*** (.036)	0.12*** (.035)	0.01 (.031)	0.02 (.033)
	Altruism	−0.01 (.037)	−0.09** (.036)	0.32*** (.035)	0.25*** (.036)	0.24*** (.038)	0.07** (.033)
	Trust	0.07** (.031)	0.02 (.032)	0.21*** (.037)	0.17*** (.037)	0.08** (.033)	0.01 (.033)
	Reciprocity	−0.00 (.039)	−0.09** (.036)	0.30*** (.035)	0.29*** (.036)	0.24*** (.040)	0.12*** (.035)
	Punishment	0.02 (.036)	−0.03 (.036)	0.10** (.041)	0.08** (.038)	0.08** (.035)	0.19*** (.040)

Notes: Data from U.S. Study 1, Week 0 ($N = 1,950$). Correlations are using ORIV, with bootstrapped standard errors. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

$\eta = \delta\eta_y + (1-\delta)\eta_p$ is decomposed such that $\text{Cov}[\eta_p, y] = \text{Cov}[\eta_y, p] = 0$.[24]

The validity of q as a proxy for p cannot be assessed by comparing β_p—the coefficient from a regression of y on p—to β_q—the coefficient from a regression of y on q. In particular, even if $\beta_q = \beta_p$, the coefficient β_q may be entirely due to confounding factors. For example, suppose that $\gamma = 1$ and $\text{Cov}[\eta_y, y] = \beta_p/\delta$. In this case, $\beta_q = \beta_p$, despite the fact that q contains none of the relevant variation in p. Thus, $\beta_q = \beta_p$ for one auxiliary variable y would not translate to other variables of interest.

[24]This decomposition is generally not unique. Hence, it is not possible to identify δ, η_p, and η_y without further assumptions. In any decomposition, however, $\text{Cov}[\eta_i, i] \neq 0 \iff \text{Cov}[\eta, i] \neq 0$, for $i \in \{p, y\}$.

To evaluate the validity of q (with respect to y), the following regression is helpful:

$$q = \beta_{y|p} y + \beta_{p|y} p + \varepsilon_3 \qquad \Rightarrow \qquad \beta_{y|p} = \frac{\gamma(\delta \text{Cov}[\eta_y, y] - (1-\delta)\beta_p \text{Cov}[\eta_p, p])}{1 - \beta_p^2},$$

in which the expression for $\beta_{y|p}$ follows from an application of the Frisch–Waugh–Lovell theorem. If there is only white noise error in q (with respect to both y and p)—that is, if $\text{Cov}[\eta_y, y] = \text{Cov}[\eta_p, p] = 0$—then $\beta_{y|p} = 0$. We use this insight to evaluate qualitative self-assessments vis à vis a variety of demographic and economic variables.

The variation in qualitative self-assessments and demographic and economic characteristics appears to be independent of the variation in qualitative self-assessments that is associated with incentivized elicitation(s), as shown in Table 5. This table regresses the qualitative self-assessments of risk tolerance and altruism on cognitive ability, gender, age, education, income, and an indicator for stock ownership.[25] The first column for each measure shows that these self-assessments are generally strongly correlated with demographic and economic variables. The second column for each measure introduces the incentivized measure as a control. As shown above, if qualitative self-assessments differ from their counterparts only in the addition of white noise measurement error, the coefficients on demographic and economic variables would go to zero. However, the coefficients are largely unchanged, indicating that almost *none* of the correlations between self-assessments and demographics or economic variables are driven by variation in the incentivized elicitations. This observation is not due to measurement error in the incentivized elicitations: we use the duplicate incentivized elicitations as instruments for one another, ensuring that attenuation bias in the coefficients does not allow the demographic and economic variables to absorb excess variation (see Table 3, and surrounding text, in Gillen et al., 2019). Further, this pattern is in spite of the fact that the incentivized elicitations themselves are often statistically significantly correlated with the demographic and economic variables (see Appendix Figure A.3).

[25] Appendix A.3 shows similar patterns for impatience, trust, reciprocity, and punishment. For risk tolerance and impatience, we can also incorporate data from U.S. Sample 2, and obtain similar results.

Table 5: Relationships between qualitative self-assessments and demographics are unrelated to variation in incentivized elicitations (N = 1,950).

	Dependent Variable = Qualitative Self-Assessment					
	Risk Tolerance			Altruism		
Cognitive Ability	-0.12^{***}	-0.13^{***}	-0.12^{***}	-0.04	-0.06^{**}	-0.06^{**}
	(.028)	(.028)	(.030)	(.028)	(.027)	(.030)
Male	0.34^{***}	0.35^{***}	0.33^{***}	-0.20^{***}	-0.22^{***}	-0.21^{***}
	(.055)	(.054)	(.055)	(.056)	(.055)	(.054)
Age	-0.09^{***}	-0.10^{***}	-0.10^{***}	0.17^{***}	0.17^{***}	0.16^{***}
	(.028)	(.028)	(.027)	(.028)	(.028)	(.027)
Education	0.05	0.05	0.06^{**}	0.03	0.03	0.03
	(.031)	(.031)	(.031)	(.029)	(.029)	(.029)
Income (Log)	0.00	0.01	0.00	-0.02	-0.02	-0.01
	(.035)	(.035)	(.035)	(.034)	(.033)	(.033)
Stock Investor	0.07	0.07	0.08	0.06	0.05	0.04
	(.065)	(.064)	(.065)	(.061)	(.062)	(.061)
Incentivized Elicitation:						
Risk Tolerance		0.15^{***}	0.14^{***}			-0.05
		(.037)	(.037)			(.039)
Impatience			0.04			-0.06
			(.035)			(.037)
Altruism			0.29^{**}		0.35^{***}	0.46^{***}
			(.139)		(.048)	(.130)
Trust			-0.06			-0.20
			(.159)			(.145)
Reciprocity			-0.04			0.11^{**}
			(.058)			(.055)
Punishment			0.12^{***}			-0.02
			(.037)			(.037)

Notes: All columns use data from U.S. Sample 1. Incentivized elicitations are instrumented to eliminate the effect of classical measurement error. Coefficients and standard errors (in parentheses) on all continuous measures are standardized. All specifications include an indicator variable for missing income. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

We find similar results when examining the relationship between qualitative self-assessments and self-reported behaviors, including health behaviors, community engagement, and political interest; see Appendix A.4. Consistent with previous studies, we find that qualitative self-

assessments, as well as incentivized elicitations, are correlated with a range of self-reported behaviors, although the interpretation of these correlations is not always clear. For instance, self-assessed risk tolerance is positively correlated with both unhealthy behaviors, like smoking and binge drinking, and healthy behaviors, like exercise and eating fruit and vegetables. These correlations are not driven by the variation in incentivized elicitations. Further, controlling for self-reported behaviors has little effect on estimated correlations between demographics and qualitative self-assessments. Together, these findings suggest that qualitative self-assessments are capturing something unrelated to preferences as typically understood by economists.

5.3 Bounds on Correlations Driven by Confounding Factors

Most of the correlations between qualitative self-assessments and demographics stem from confounding factors. To demonstrate this, we derive estimable bounds on the share of these correlations attributable to confounding factors.

We restate two regression specifications discussed in Section 2, with two additional specifications that are useful for further analysis:

$$
\begin{aligned}
y &= \beta_p p + \varepsilon_1 && \Rightarrow && \beta_p = \mathrm{Cov}[y, p] \\
q &= \beta_{qp} p + \varepsilon_2 && \Rightarrow && \beta_{qp} = (1-\gamma) + \gamma(1-\delta)\mathrm{Cov}[\eta_p, p] \\
y &= \beta_q q + \varepsilon_3 && \Rightarrow && \beta_q = (1-\gamma)\beta_p + \gamma\delta\mathrm{Cov}[\eta_y, y] \\
y &= \beta_{q|p} q + \beta_{p|q} p + \varepsilon_4 && \Rightarrow && \beta_{q|p} = \frac{\gamma\delta\mathrm{Cov}[\eta_y, y] - \gamma(1-\delta)\beta_p\mathrm{Cov}[\eta_p, p]}{1-\beta_{qp}^2}
\end{aligned}
$$

Similar to our discussion of $\beta_{y|p}$, if η is pure white noise error (with respect to both y and p), so that $\mathrm{Cov}[\eta_y, y] = \mathrm{Cov}[\eta_p, p] = 0$, then $\beta q|p = 0$.

When $\mathrm{Cov}[\eta_y, y] \neq 0$, there is a risk of a correlation between q and y that stems from confounding factors. We now bound the segment of β_q that is driven by correlation that is attributable to confounding factors, namely $\Psi \equiv \gamma\delta\mathrm{Cov}[\eta_y, y]$.

We can use the equalities above to write:

$$\beta_{q|p} = \frac{\Psi - \gamma(1-\delta)\beta_p \text{Cov}[\eta_p, p]}{1-\beta_{qp}^2},$$

$$\Psi = \beta_{q|p}(1-\beta_{qp}^2) + \gamma(1-\delta)\beta_p \text{Cov}[\eta_p, p]\,.$$

We now derive lower and upper bounds on Ψ. For both, it is useful to note that, by construction, the variance of η_y is bounded by 1. It follows that $(1-\delta)|\text{Cov}[\eta_p, p]\,| \leq (1-\delta)\sqrt{\text{Var}[\eta_p]} \leq 1$. This is a direct consequence of the correlation between η_p and p falling in the $[-1, 1]$ interval. We assume that $\beta_p \geq 0$. Analogous arguments follow for $\beta_p < 0$.

Since $\beta_p \geq 0$, Ψ is bounded above by β_q. Furthermore, since $(1-\delta)|\text{Cov}[\eta_p, p]\,| \leq 1$, Ψ is maximized when $\beta_{qp} = \gamma(1-\delta)\text{Cov}[\eta_p, p]$. In other words, Ψ is maximized when $\gamma = 1$, implying that $\Psi = \beta_q$. Thus, we can never rule out the possibility that the correlation between y and q is entirely due to confounding factors.

The minimum value of Ψ is achieved when $\gamma(1-\delta)\text{Cov}[\eta_p, p]$ is minimized (given β_{qp} and $\beta_{q|p}$). Since $|\text{Cov}[\eta_p, p]\,| \leq \sqrt{\text{Var}[\eta_p]}$, the variance of η_p must satisfy

$$\text{Var}[q] = 1 = (1-\gamma)^2 + \gamma^2(1-\delta)^2\text{Var}[\eta_p] + \gamma^2\delta^2\text{Var}[\eta_y]\,.$$

The value of $\text{Var}[\eta_p]$ is maximized when $\text{Var}[\eta_y] = 0$. As this implies that η_y is a constant, this is also consistent with $\delta\text{Cov}[\eta_y, y] = 0$. In this case, $|\gamma(1-\delta)\text{Cov}[\eta_p]\,| \leq \sqrt{1-(1-\gamma)^2}$. In particular, it follows that $\gamma(1-\delta)\text{Cov}[\eta_p] \geq -\sqrt{1-(1-\gamma)^2}$. To get a lower bound on Ψ, we therefore need to minimize $-\sqrt{1-(1-\gamma)^2}$, subject to $\gamma, \delta \in [0, 1]$ and $\beta_{qp} = (1-\gamma) - \sqrt{1-(1-\gamma)^2}$. That minimum is achieved at $-\frac{1}{\sqrt{2}}\sqrt{1-\beta_{qp}\sqrt{(2-\beta_{qp}^2)}}$.

Using symmetric arguments for $\beta_p < 0$, we have the following:

Proposition (Bounds on Correlation Component due to Confounding Factors). *The com-*

ponent Ψ *of the coefficient* β_q *that is due to compounding factors satisfies*

$$\mathcal{B}^- = \beta_{q|p}(1-\beta_{qp}^2) - \beta_p\tilde{\beta}_{qp} \leq \Psi \leq \beta_q = \mathcal{B}^+ \qquad \text{if } \beta_p \geq 0$$
$$\mathcal{B}^- = \beta_q \leq \Psi \leq \beta_{q|p}(1-\beta_{qp}^2) - \beta_p\tilde{\beta}_{qp} = \mathcal{B}^+ \qquad \text{if } \beta_p < 0,$$

in which $\tilde{\beta}_{qp} = \frac{1}{\sqrt{2}}\sqrt{1-\beta_{qp}\sqrt{(2-\beta_{qp}^2)}}$.

Assuming again that $\beta_p \geq 0$, a few implications follow. First, an estimated $\beta_{q|p} = 0$ does not ensure that $\Psi \approx 0$. Indeed, the regression estimation corresponding to $\beta_{q|p}$ does not allow us to rule out the possibility that the component Ψ that is due to confounding factors is "canceled out" by the correlation due to confounding factors between q and p. Nevertheless, if $\beta_{q|p} = 0$, then the interval specified in the proposition will always contain 0. That is, zero correlation due to confounding factors ($\Psi = 0$) cannot be ruled out. As such, checking whether $\beta_{q|p} = 0$ can provide a helpful (minimal) test for future experimental validations. Second, if β_p and β_q have different signs, then we can be sure that $|\Psi| > 0$. Third, if $\beta_{qp} = 1$ then $\Psi \in [0, \beta_q]$. That is, the fraction of β_q could be 0 or could capture the whole of β_q.

Corollary (Minimum Coefficient Share due to Confounding Factors)**.** *The minimum share* Ψ *of the coefficient* β_q *that is driven by confounding factors satisfies:*

$$\Psi \geq \begin{cases} 0 & \textit{if } 0 \in [\mathcal{B}^-, \mathcal{B}^+] \\ \frac{\min\{|\mathcal{B}^-|, |\mathcal{B}^+|\}}{|\beta_q|} & \textit{otherwise.} \end{cases}$$

Table 6 displays the portion of the coefficients β_q that remain after removing the minimum portion that is explained by confounding factors Ψ, using the corollary. Nearly all the coefficients are close to zero, and few are statistically significant. That is, even in the best case scenario, nearly all the predictive power of these measures is due to confounding factors.

Table 6: After accounting for confounding factors, there is limited evidence that qualitative self-assessments are correlated with economic or demographic characteristics.

	Cognitive Ability	Male	Age	Income (Log)	Education	Stock Investor
Risk Tolerance	−0.01	0.00	0.00	0.00	0.00	0.00
	(.032)	(.031)	(.030)	(.037)	(.034)	(.031)
Impatience	0.00	0.01	−0.06*	−0.04	−0.04	−0.00
	(.031)	(.031)	(.032)	(.033)	(.031)	(.032)
Altruism	0.00	0.00	0.02	0.01	0.01	0.05
	(.033)	(.034)	(.033)	(.039)	(.035)	(.034)
Trust	0.00	0.00	0.00	0.04	0.02	0.09***
	(.030)	(.032)	(.031)	(.033)	(.032)	(.031)
Reciprocity	0.09***	0.03	0.07**	0.03	0.02	0.05
	(.033)	(.035)	(.035)	(.039)	(.035)	(.034)
Punishment	0.00	0.05	0.00	0.04	0.00	0.00
	(.037)	(.037)	(.041)	(.044)	(.040)	(.039)

Notes: $N = 1,950$. The table displays univariate correlations between qualitative self-assessments and individual characteristics (β_q), using the corollary to remove the minimum portion that is explained by confounding factors. Incentivized elicitations and qualitative self-assessments are instrumented to eliminate the effect of classical measurement error. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

6 Response Heuristics in Qualitative Self-Assessments

Earlier sections highlight limitations of qualitative self-assessments as proxies for traditional economic preference measures. However, their use could also be justified by their perceived simplicity. For example, Dohmen et al. (2018, p. 126) contend that the "simplicity of [the] general risk question....has the advantage of being easy to understand, thereby limiting the problem of decision errors or noise." The analyses in this section reveal that, in fact, qualitative self-assessments exhibit response biases comparable to those of incentivized elicitations, challenging the notion that they are inherently straightforward for participants. Moreover, we provide evidence of cross-cultural variation in how self-assessments are interpreted.

A simple example illustrates how qualitative self-assessments may be as—if not more—difficult to answer than more objective measures. Imagine visiting an optometrist and being

asked to rate your eyesight on a scale from 0 to 10. Now, consider being asked to assess your skills as an economist using the same scale. How would you interpret the meaning of each number in these contexts? Would you be confident that your family and/or colleagues would use the same interpretation? Would assigning a score be easier than taking a vision test or collecting a citation count?

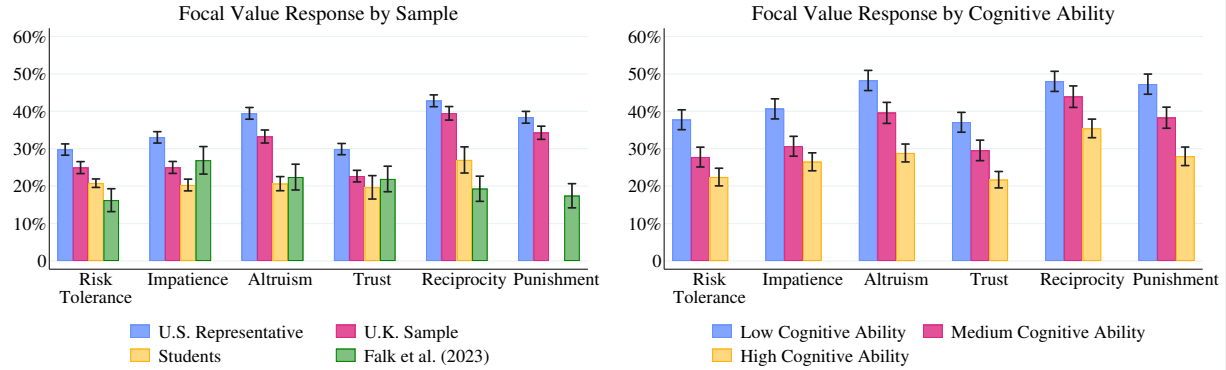
The qualitative self-assessments we study likely pose similar challenges to participants, but with added conceptual complexity. To respond carefully to "How willing are you to take risks, in general?", for example, a participant should i) determine the conceptual content of the question—for instance, which types of risks are being referred to; ii) consider how that concept applies to their own behavior and experiences; and then iii) decide how to aggregate that information and project it onto an 11-point numerical scale. Consequently, while responses to self-assessments might result from preferences, they could also reflect a combination of participants' reference groups, self-perceptions, social expectations, and their understanding of numerical scales, much of which may vary independently of preferences.[26] The coefficients relating to self-assessments in Table 5 may be capturing exactly such factors.

Qualitative self-assessments appear more challenging to answer for those with lower cognitive ability, as evidenced by Figure 4. This figure shows the rate of focal value response (FVR; Chapman et al., 2024a)—defined as selecting 0, 5, or 10—across different samples (left panel) and among cognitive ability terciles within our representative samples of the U.S. (right panel).[27] FVR levels are notably higher in broad population samples than in student samples, whether from the U.S. or from Germany (using data from Falk et al., 2023). This disparity appears to stem from participants in the representative sample who rank in the bottom tercile of cognitive ability. The pattern is consistent with lower cognitive ability individuals "rounding" their responses when faced with the perceived complexity of

[26] Arslan et al. (2020) report that individuals, when completing the risk qualitative self-assessment, think about behaviors they engage in.

[27] High levels of FVR in general populations have been observed in other studies as well. In the GPS, 40% of responses to the trust qualitative self-assessment were at focal values, compared to 30% in our representative samples of the U.S. (based on the replication dataset of Falk et al. (2018)). Similarly, Dohmen et al. (2011, see Figure 1) report high FVR rates for the risk tolerance self-assessment.

Figure 4: Focal Value Response (FVR) is more common in the general population, particularly among participants with low cognitive ability.



Notes: The figure displays the percentage of responses that were focal values—0, 5, or 10—in each qualitative self-assessment. U.S. Representative includes U.S. Samples 1 and 2 (combined $N = 4,950$). Students combines Caltech, Pittsburgh, and UBC samples. Low, medium, and high cognitive ability refer to terciles, from the combination of U.S. Samples 1 and 2. Bars represent 90% confidence intervals.

translating their preferences into an 11-point response scale (Barrington-Leigh, 2024).[28] The relationship between FVR and cognitive ability also suggests that heuristics may be one source of confounding factors (η_y) leading to the spurious correlations between qualitative self-assessments and other individual characteristics documented in the previous section.

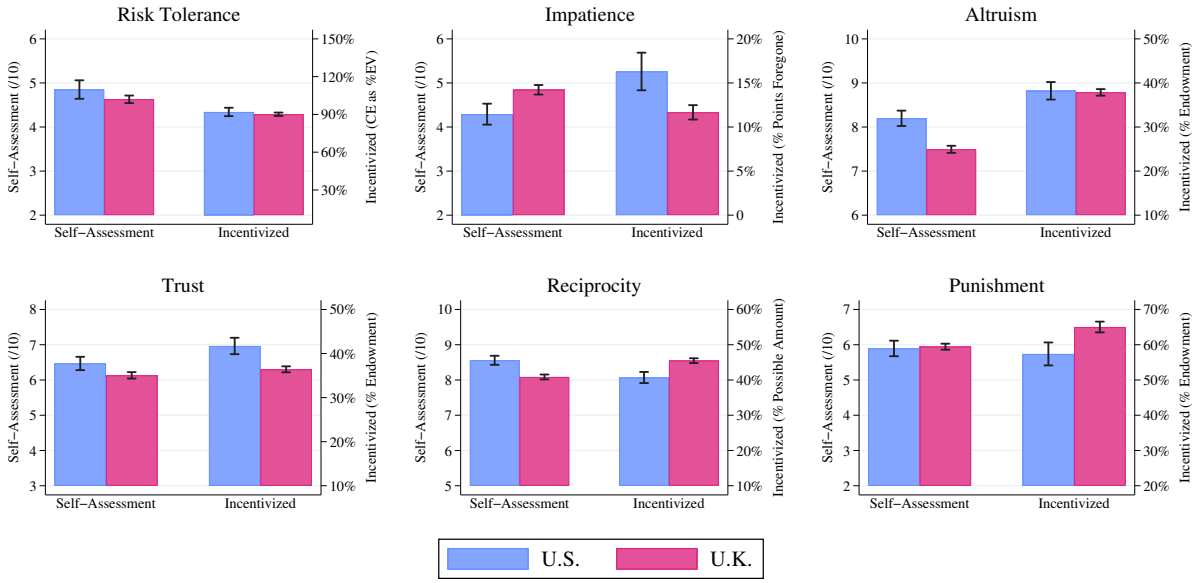
The observed pattern of FVR implies that qualitative self-assessments suffer from similar measurement issues as established incentivized elicitations, negating a purported advantage of the method. Chapman et al. (2024a; Figure 8) report comparable levels and patterns of FVR for both qualitative self-assessments and multiple price lists (MPLs) used to elicit risk tolerance and impatience.[29] Further, the use of heuristics such as FVR is particularly troubling as, unlike classical measurement error, they may lead to biased estimates, rather than simple attenuation.

A more detailed comparison of our U.K. and U.S. samples, displayed in Figure 5, suggests

[28]See Krosnick and Presser (2010) for a general discussion of theoretical challenges associated with numerical rating scales. Correlations between incentivized elicitations and qualitative self-assessments of risk and time preferences also vary by cognitive ability (see Chapman et al., 2024a, Figure 9). This variation may stem from participants with lower cognitive ability interpreting terms such as "willingness to take risk" differently or employing distinct approaches (or heuristics) to map their responses onto a numerical scale.

[29]For MPLs eliciting risk tolerance, Chapman et al. (2024a) report that 41% of participants with low cognitive ability gave focal value responses, compared to 31% and 30% for participants with medium and high cognitive ability.

Figure 5: Comparison of the preferences of low-education individuals in the U.S. and U.K.



Notes: The figure compares the average preference within different samples. The U.K. sample includes all participants in the U.K. low education sample ($N = 1,984$). These individuals are contrasted with a comparable subsample of U.S. Samples 1 and 2—those aged over 55 and with at most high school education ($N = 803$). Conclusions are similar when comparing to the entire U.S. Sample—see Appendix Figure A.4.

that cultural background may play a role in respondents' interpretation of qualitative self-assessments.[30] This figure compares the average responses to qualitative self-assessments and incentivized elicitations in our sample of older, low-education, adults in the U.K., to those from a comparable subsample of U.S. participants. While cultural differences affect responses to both types of elicitations, they are at least as pronounced for qualitative self-assessments. Furthermore, in line with our discussion in previous sections, in only one of the six domains—trust—are the differences between the U.S. and the U.K. in the same direction (and statistically significant) for both the qualitative self-assessment and the incentivized elicitation. The differences are particularly striking for impatience and reciprocity.

[30]Differences in interpretation across groups, within or between countries, would indicate that qualitative self-assessments lack measurement invariance, and preclude meaningful comparisons between those groups. See Mellenbergh (1989) and the more recent survey by Dong and Dumas (2020).

7 Discussion

We examine both the experimental validation method and the validity of qualitative self-assessments for core economic preferences, using data from over 13,000 participants in a number of representative, convenience, and student samples. Across these samples, correlations between qualitative self-assessments and incentivized elicitations are consistently small, even after adjusting for measurement error. Moreover, multiple qualitative self-assessments exhibit stronger correlations with incentivized measures of other constructs. Notably, the links between various attributes—geography, demographics, and behaviors—and qualitative self-assessments differ from the links between these same attributes and incentivized elicitations. Experimental validation is thus insufficient to ensure valid inference in broad populations. Finally, while qualitative self-assessments are often promoted as simpler than incentivized elicitations, our findings indicate that participants frequently rely on heuristics when answering qualitative questions. This tendency is particularly prevalent among participants with lower cognitive ability, who disproportionately select salient options, such as the extremes or midpoints of numerical scales.

In what follows, we discuss the implications of our findings for the use of qualitative self-assessments and for preference elicitation. This discussion rests on the premise that incentivized elicitations are the gold standard for preference measurement. For instance, Falk et al. (2023, p. 1946) emphasize that "the guiding methodology for developing the [preference survey] modules is identifying survey items that can predict well the choices in incentivized experiments." Measures that do not reflect choices in incentivized experiments may still be of value. Future research could delve deeper into the constructs captured by qualitative self-assessments, and develop models that align with these constructs. Alternatively, future work could design richer qualitative measures (Kosfeld et al., 2025), or develop tools to address issues associated with numerical response scales (Benjamin et al., 2023). The Domain-Specific Risk-Taking (DOSPERT) scale, for example, employs 30 questions that focus primarily on behaviors, rather than self-assessments (Blais and Weber, 2006). Expanding

qualitative measures into longer, more nuanced question sequences could help disentangle preferences from behaviors, perceptions, and differences in interpretation.

Implications for Findings Based on Self-Assessments: Our findings call into question the conclusions drawn from studies that rely on qualitative self-assessments, such as the GPS. Our results are consistent with previous research indicating that qualitative self-assessments may capture differences in understanding, perceptions, or habits (Paulhus and Vazire, 2007; Brenner and DeLamater, 2016; Arslan et al., 2020). As such, using qualitative self-assessments as proxies for preferences may yield misleading results. When both qualitative and choice-based responses are available, it may be informative to analyze the relationship between the two, and to understand when the different measures generate disparate findings. Making responses to the individual questions underlying the GPS indices—separating responses to the qualitative self-assessments from those to quantitative questions—accessible to the research community would facilitate this process and enhance credibility.[31]

Implications for Preference Elicitation: The appeal of qualitative self-assessments largely stems from their perceived ability to facilitate preference measurement across broad populations. However, our results suggest that these measures do not accurately capture preferences reflected in incentivized elicitations. Instead, qualitative self-assessments may be informative about how individuals perceive, say, the riskiness of their behavior(Schonberg et al., 2011), or how they conceptualize their "ideal self" (Brenner and DeLamater, 2016). While these perceptions may be important in their own right, they may bear little connection to the preferences economists typically study or to the utility parameters that incentivized elicitations are designed to capture. Another path would be to pursue further exploration

[31]Trust and negative reciprocity in the GPS are assessed solely through qualitative self-assessments. Measures for risk, impatience, and altruism include experimental tasks with hypothetical incentives, while reciprocity is evaluated using a hypothetical scenario involving a "thank you gift." Falk et al. (2018) combine the multiple questions for each preference into a single index, and so their replication dataset does not include responses to the underlying questions. The authors of Falk et al. (2018) have not been willing to share the disaggregated responses.

of whether monetary incentives are essential for traditional choice-based elicitations (Stango and Zinman, 2020; Brañas-Garza et al., 2023).

Advances in methodology and online survey platforms have made it feasible to conduct incentivized studies with broad population samples, such as those studied in this paper. While incentivized elicitations have been criticized for their limited predictive power, this concern has been mitigated by improved techniques for addressing measurement error (Beauchamp et al., 2017; Gillen et al., 2019; Chapman et al., 2024b; Jagelka, 2024). For instance, the incentivized elicitations in this paper demonstrate meaningful correlations with a range of individual characteristics (Appendix Figure A.3) and self-reported behaviors (Appendix Figure A.6). Despite these advances, there remains room to improve incentivized elicitations; see, for example, Chapman et al. (2024a); Gerhardt and Suchy (2024).[32] We believe that efforts should focus on refining choice-based elicitations rather than moving away from them.

[32]Chapman et al. (2024a) use the Dynamically Optimized Sequential Experimentation (DOSE) method to elicit risk and time preferences in a representative online survey. FVR rates in DOSE stand at around 6%, with minimal variation across cognitive ability levels, suggesting the method is simple to understand.

References

Albrecht, David and Thomas Meissner, "The Debt Aversion Survey Module: An Experimentally Validated Tool to Measure Individual Debt Aversion," 2022. arXiv:2211.02742.

Arslan, Ruben C, Martin Brümmer, Thomas Dohmen, Johanna Drewelies, Ralph Hertwig, and Gert G Wagner, "How People Know their Risk Preference," *Scientific Reports*, 2020, *10* (1), 15365.

Barcellos, Silvia, Leandro Carvalho, Pietro Ortoleva, Francisco Perez-Arce, and Erik Snowberg, "Why did a Compulsory Schooling Law Raise Earnings but Lower Life Satisfaction?," 2024. Mimeo.

Barrington-Leigh, Christopher, "The Econometrics of Happiness: Are we Underestimating the Returns to Education and Income?," *Journal of Public Economics*, 2024, *230*, 105052.

Beauchamp, Jonathan P, David Cesarini, and Magnus Johannesson, "The psychometric and empirical properties of measures of risk preferences," *Journal of Risk and Uncertainty*, 2017, *54*, 203–237.

Becker, Anke, Benjamin Enke, and Armin Falk, "Ancient Origins of the Global Variation in Economic Preferences," in "AEA Papers and Proceedings," Vol. 110 2020, pp. 319–323.

Benjamin, Daniel J, Kristen Cooper, Ori Heffetz, Miles S Kimball, and Jiannan Zhou, "Adjusting for Scale-Use Heterogeneity in Self-Reported Well-Being," 2023. NBER Working Paper #31,728.

Bilbiie, Florin O., "Monetary Policy and Heterogeneity: An Analytical Framework," *Review of Economic Studies*, forthcoming.

Blais, Ann-Renée and Elke Weber, "A Domain-Specific Risk-Taking (DOSPERT) Scale for Adult Populations," *Judgment and Decision Making*, 2006, *1* (1), 33–47.

Brañas-Garza, Pablo, Diego Jorrat, Antonio M Espín, and Angel Sánchez, "Paid and hypothetical time preferences are the same: Lab, field and online evidence," *Experimental Economics*, 2023, *26* (2), 412–434.

Brenner, Philip S and John DeLamater, "Lies, damned lies, and survey self-reports? Identity as a cause of measurement bias," *Social Psychology Quarterly*, 2016, *79* (4), 333–354.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek, "Can Competitiveness Predict Education and Labor Market Outcomes? Evidence from Incentivized Choice and Survey Measures," *Review of Economics and Statistics*, 2024, pp. 1–45.

Campos-Mercade, Pol, Armando Meier, Florian Schneider, and Erik Wengström, "Prosociality Predicts Health Behaviors during the COVID-19 Pandemic," *Journal of Public Economics*, 2021, *195*, 104367.

Cao, Yiming, Benjamin Enke, Armin Falk, Paola Giuliano, and Nathan Nunn, "Herding, warfare, and a culture of honor: Global evidence," 2021. NBER Working Paper #29,250.

Cavatorta, Elisa and David Schröder, "Measuring Ambiguity Preferences: A New Ambiguity Preference Survey Module," *Journal of Risk and Uncertainty*, 2019, *58*, 71–100.

Chan, Ho Fai, Ahmed Skali, David Savage, David Stadelmann, and Benno Torgler, "Risk Attitudes and Human Mobility during the COVID-19 Pandemic," *Nature*

Scientific Reports, 2020, *10* (1), 19931.

_ , **Martin Brumpton, Alison Macintyre, Jefferson Arapoc, David A Savage, Ahmed Skali, David Stadelmann, and Benno Torgler**, "How Confidence in Health Care Systems Affects mMbility and Compliance during the COVID-19 Pandemic," *PLOS one*, 2020, *15* (10), e0240644.

Chapman, Jonathan, Erik Snowberg, Stephanie Wang, and Colin Camerer, "Dynamically Optimized Sequential Experimentation (DOSE) for Estimating Economic Preference Parameters," 2024. NBER Working Paper #33,013.

_ , _ , _ , **and Colin F. Camerer**, "Looming Large or Seeming Small? Attitudes Towards Losses in a Representative Sample," *The Review of Economic Studies*, 2024. rdae093.

_ , **Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer**, "Econographics," *Journal of Political Economy Microeconomics*, 2023, *1* (1), 115–161.

_ , **Pietro Ortoleva, Erik Snowberg, and Colin Camerer**, "Time Stability of Behavioral (and other) Measures," 2024. Mimeo.

Charness, Gary, Thomas Garcia, Theo Offerman, and Marie Claire Villeval, "Do Measures of Risk Attitude in the Laboratory Predict Behavior under Risk in and Outside of the Laboratory?," *Journal of Risk and Uncertainty*, 2020, *60* (2), 99–123.

Clemente, Eva Miranda, Michael Ehst, Timothy John Charles Kelly, Doreen Oppan, Jana Kunicova, Max William Mattern, Kaoru Kimura, Ander Alcalde Odriozola, Claudia Ivette Garcia Romero, Daniel Enrique Querejazu, and Minita Mary Varghese, "Ghana Digital Economy Diagnostic: Stock-Taking Report," 2019.

Condon, David M. and William Revelle, "The International Cognitive Ability Resource: Development and Initial Validation of a Public-Domain Measure," *Intelligence*, 2014, *43*, 52–64.

Dohmen, Thomas, Armin Falk, Armin Huffman, and Uwe Sunde, "Are Risk Aversion and Impatience Related to Cognitive Ability?," *The American Economic Review*, 2010, *100* (3), 1238–1260.

_ , _ , **David Huffman, and Uwe Sunde**, "On the Relationship between Cognitive Ability and Risk Preference," *Journal of Economic Perspectives*, 2018, *32* (2), 115–34.

_ , _ , _ , _ , **Jürgen Schupp, and Gert G. Wagner**, "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences," *Journal of the European Economic Association*, 2011, *9* (3), 522–550.

Dong, Yixiao and Denis Dumas, "Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age," *Personality and Individual Differences*, 2020, *160*, 109956.

Enke, Benjamin, Ricardo Rodriguez-Padilla, and Florian Zimmermann, "Moral Universalism: Measurement and Economic Relevance," *Management Science*, 2022, *68* (5), 3590–3603.

Falk, Armin and Johannes Hermle, "Relationship of gender differences in preferences to economic development and gender equality," *Science*, 2018, *362* (6412), eaas9899.

_ , **Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde**, "Global Evidence on Economic Preferences," *The Quarterly Journal of Economics*, 2018, *133* (4), 1645–1692.

_ , _ , _ , **David Huffman, and Uwe Sunde**, "The Preference Survey Module: A

Validated Instrument for Measuring Risk, Time, and Social Preferences," *Management Science*, 2023, *69* (4), 1935–1950.

Fallucchi, Francesco, Daniele Nosenzo, and Ernesto Reuben, "Measuring Preferences for Competition with Experimentally-Validated Survey Questions," *Journal of Economic Behavior & Organization*, 2020, *178*, 402–423.

Flückiger, Matthias, Erik Hornung, Mario Larch, Markus Ludwig, and Allard Mees, "Roman transport network connectivity and economic integration," *The Review of Economic Studies*, 2022, *89* (2), 774–810.

Frederick, Shane, "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, 2005, *19* (4), 25–42.

Gerhardt, Holger and Rafael Suchy, "Estimating preference parameters from strictly concave budget restrictions," 2024. ECONtribute Discussion Paper.

Gillen, Ben, Erik Snowberg, and Leeat Yariv, "Experimenting with Measurement Error: Techniques and Applications from the Caltech Cohort Study," *Journal of Political Economy*, 2019, *127* (4), 1826–1863.

Hanushek, Eric A, Lavinia Kinne, Philipp Lergetporer, and Ludger Woessmann, "Patience, risk-taking, and human capital investment across countries," *The Economic Journal*, 2022, *132* (646), 2290–2307.

Jackson, Mathew, Stephen Nei, Erik Snowberg, and Leeat Yariv, "The Dynamics of Networks and Homophily," 2023. NBER Working Paper #30,815.

Jagelka, Tomáš, "Are economists' preferences psychologists' personality traits? A structural approach," *Journal of Political Economy*, 2024, *132* (3), 910–970.

Jonsson, Sara and Qinglin Ouyang, "Effects of cultural origin on entrepreneurship," *Journal of Economic Behavior & Organization*, 2023, *216*, 308–319.

Korff, Alex and Nico Steffen, "Economic Preferences and Trade Outcomes," *Review of World Economics*, 2022, *158* (1), 253–304.

Kosfeld, Michael and Zahra Sharafi, "The Preference Survey Module: Evidence on Social Preferences from Tehran," *Journal of the Economic Science Association*, 2024, *10* (1), 152–164.

_ , _ , Maria Sontag Gonzalez, and Na Zou, "Measuring Economic Preferences with Surveys and Behavioral Experiments," 2025. CEPR Discussion Paper No. 19,845.

Krosnick, Jon and Stanley Presser, "Question and Questionnaire Design," in Peter Marsden and James Wright, eds., *Handbook of Survey Research*, 2010, pp. 263–313.

McCallum, Bennett T, "Relative asymptotic bias from errors of omission and measurement," *Econometrica*, 1972, *40* (4), 757–758.

Mellenbergh, Gideon J, "Item bias and item response theory," *International Journal of Educational Research*, 1989, *13* (2), 127–143.

Paulhus, Delroy L and Simine Vazire, "The Self-Report Method," in Richard W. Robins, Chris R. Fraley, and Robert F. Krueger, eds., *Handbook of Research Methods in Personality Psychology*, Vol. 1, Guilford, 2007, pp. 224–239.

Pickard, Harry, Thomas Dohmen, and Bert Van Landeghem, "Inequality and risk preference," *Journal of Risk and Uncertainty*, 2024, *69* (2), 191–217.

Schonberg, Tom, Craig R Fox, and Russell A Poldrack, "Mind the gap: bridging economic and naturalistic risk-taking with cognitive neuroscience," *Trends in Cognitive Sciences*, 2011, *15* (1), 11–19.

Schudy, Simeon, Susanna Grundmann, and Lisa Spantig, "Individual Preferences for Truth-Telling," 2024. CESifo Working Paper No. 11,521.

Snowberg, Erik and Leeat Yariv, "Testing the Waters: Behavior across Participant Pools," *American Economic Review*, 2021, *111* (2), 687–719.

Stango, Victor and Jonathan Zinman, "Behavioral biases are temporally stable," Technical Report, National Bureau of Economic Research 2020.

Sunde, Uwe, Thomas Dohmen, Benjamin Enke, Armin Falk, David Huffman, and Gerrit Meyerheim, "Patience and comparative development," *The Review of Economic Studies*, 2022, *89* (5), 2806–2840.

Vieider, Ferdinand, Mathieu Lefebvre, Ranoua Bouchouicha, Thorsten Chmura, Rustamdjan Hakimov, Michal Krawczyk, and Peter Martinsson, "Common Components of Risk and Uncertainty Attitudes across Contexts and Domains: Evidence from 30 Countries," *Journal of the European Economic Association*, 2015, *13* (3), 421–452.

Wickens, Michael R, "A note on the use of proxy variables," *Econometrica*, 1972, pp. 759–761.

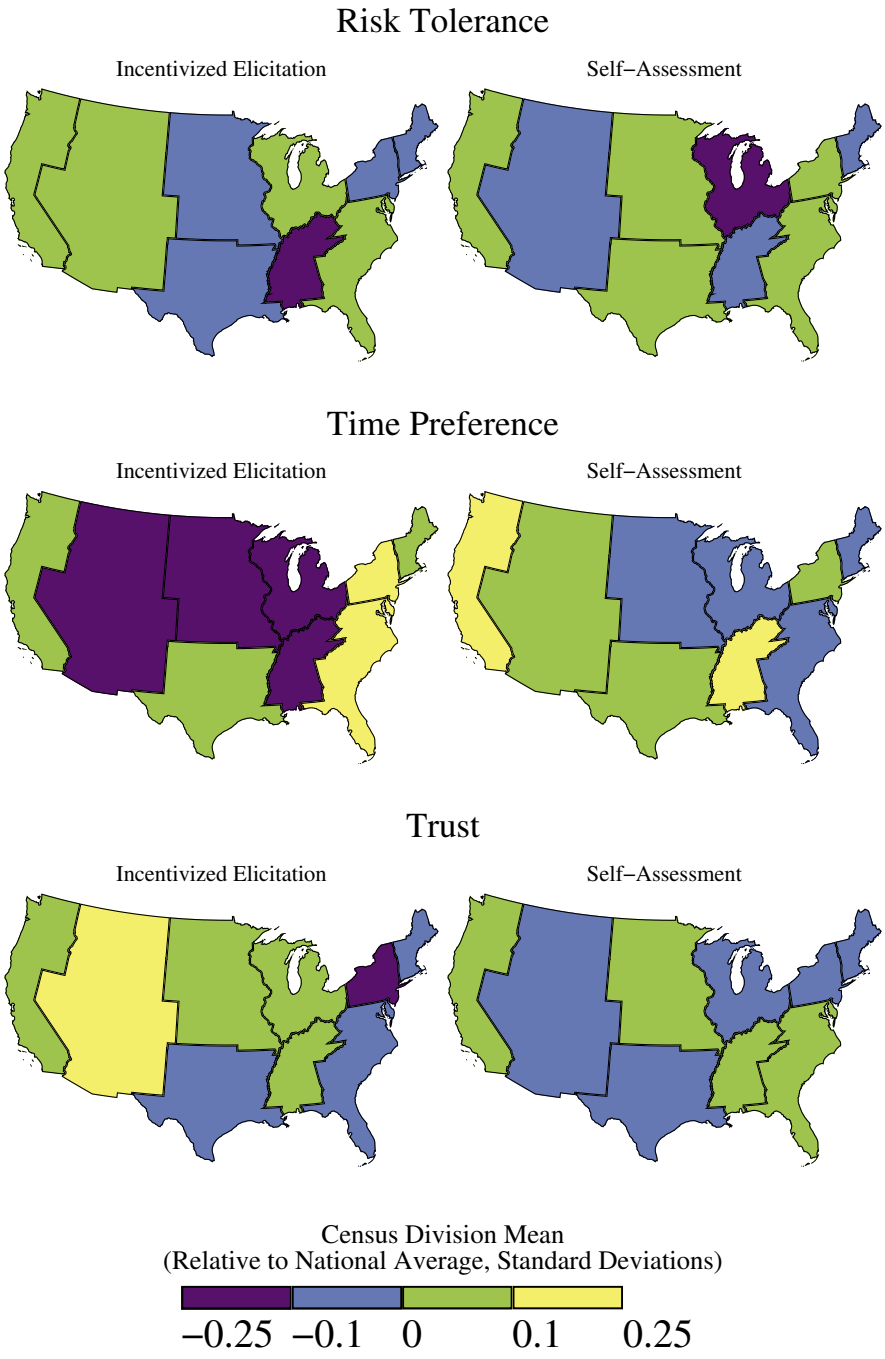
Online Appendix—Not Intended for Publication

A Additional Results

A.1 Distribution of Preferences within the United States

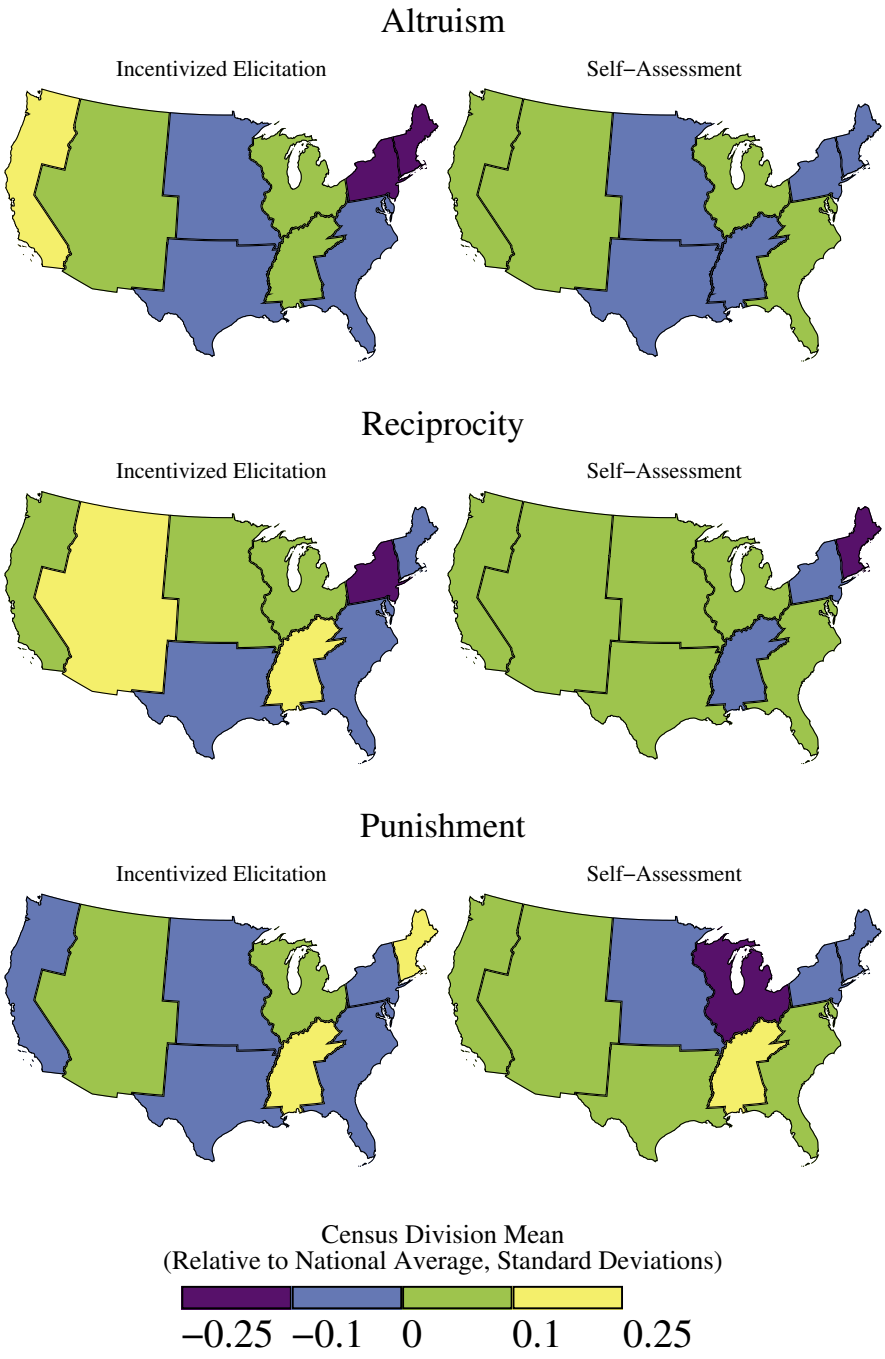
Figures A.1 and A.2 generalize Figure 1 in the main text and display the distribution of each preference at the census division level. The left-hand side of each figure displays incentivized elicitations, while the right-hand side displays the corresponding qualitative self-assessments. In each case, there are marked differences between patterns corresponding to qualitative self-assessments and the respective incentivized elicitations.

Figure A.1: Risk tolerance, time preferences, and trust in the United States.



Notes: Data from combined U.S. Samples 1 and 2. Alaska and Hawaii are included in the Pacific division. Each variable is standardized, then collapsed to state level. The scale is in standard deviations, relative to a mean of 0.

Figure A.2: Altruism, reciprocity, and punishment in the United States.



Notes: Data from combined U.S. Samples 1 and 2. Alaska and Hawaii are included in the Pacific division. Each variable is standardized, then collapsed to state level. The scale is in standard deviations, relative to a mean of 0.

A.2 Additional Correlations Between Qualitative Self-Assessments and Incentivized Elicitations

Table A.1 displays correlations between qualitative self-assessments and alternative incentivized elicitations of risk tolerance and punishment within U.S. Sample 1. Our preferred measure of risk tolerance elicited a participant's Willingness-to-Accept payment for a lottery. This measure has the virtue of being similar to Falk et al. (2023)'s measure, and also being available in U.S. Sample 2 and the U.K. Sample. The survey also included a number of other elicitations of risk preferences, including participants' Willingness-to-Pay for a lottery, their Certainty Equivalent for two lotteries, and their Certainty Equivalent for a draw from a risky urn and a draw from an ambiguous urn. For punishment, our preferred measure involves punishing the person that sent nothing back in a trust game, after receiving the maximum possible amount from the sender. We also elicited participants' willingness to engage in "anti-social punishment"—punishing the person that sent the full amount in the same trust game (who then received nothing in return). Each of these measures was elicited twice. As can be seen, the correlations between our preferred measures and these alternative elicitations are of a similar magnitude.

Table A.2 presents the correlations between qualitative self-assessments and incentivized elicitations using the replication dataset from Falk et al. (2023). As in our representative sample (Table 4), qualitative self-assessments are often statistically significantly correlated with incentivized elicitations other than those they are intended to proxy for. Table A.3 compares the correlations in Falk et al. (2023) to those in the subsample of our data that most closely resembles Falk et al. (2023)'s design—participants in U.S. Study 1 that completed two surveys within one month (see Figure 3 and surrounding discussion). Both Spearman and Pearson correlations are significantly smaller in our data than in Falk et al. (2023). Tables A.6 and A.7 present correlations between qualitative self-assessments and incentivized elicitations within various sub-groups of our U.S. general population samples. In each case, we use one qualitative self-assessment, the average of two incentivized elicitations for risk

Table A.1: Correlations with Alternative Elicitations

	Within	Within Survey	
	1 Month	Averages	ORIV
	Correlations with Risk Self-Assessment		
Preferred Measure	0.09*	0.12***	0.13***
	(.046)	(.030)	(.034)
Willingness To Pay	0.13**	0.17***	0.19***
	(.051)	(.031)	(.035)
Risk Aversion Urn	0.09*	0.10***	0.11***
	(.046)	(.031)	(.034)
Ambiguous Urn	0.15***	0.17***	0.19***
	(.044)	(.030)	(.032)
	Correlations with Punishment Self-Assessment		
Preferred Measure	0.08*	0.07**	0.09**
	(.047)	(.031)	(.038)
Anti-Social Punishment	0.09*	0.12***	0.13***
	(.046)	(.030)	(.034)
N	480	1,950	1,950

Notes: Data from U.S. Sample 1. Bootstrapped standard errors in parentheses. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

tolerance and impatience, and a single incentivized elicitation of social preferences.

Table A.2: Correlations between Qualitative Self-Assessments and Incentivized Elicitations in Falk et al. (2023)

		Incentivized Elicitations					
		Risk Tolerance	Impatience	Altruism	Trust	Reciprocity	Punishment
Qualitative Self-Assessment	Risk Tolerance	0.32*** (.047)	−0.02 (.052)	−0.03 (.051)	0.09* (.049)	−0.10** (.052)	0.03 (.052)
	Impatience	0.09* (.051)	0.08 (.053)	0.08 (.049)	0.06 (.048)	−0.00 (.050)	0.01 (.047)
	Altruism	−0.02 (.054)	0.13** (.052)	0.39*** (.039)	0.16*** (.052)	0.24*** (.051)	−0.01 (.062)
	Trust	0.03 (.046)	0.19*** (.051)	0.17*** (.050)	0.27*** (.048)	0.23*** (.050)	−0.09* (.052)
	Reciprocity	−0.03 (.052)	0.12** (.050)	0.11** (.052)	0.23*** (.049)	0.22*** (.049)	−0.05 (.055)
	Punishment	0.05 (.053)	−0.00 (.052)	0.00 (.057)	0.05 (.059)	0.15*** (.053)	0.17*** (.062)

Notes: Data from Falk et al. (2023), $N = 360$ for measures of reciprocity and punishment, and $N = 382$ for all other domains. Correlations are Pearson correlations, with bootstrapped standard errors. ***, **, * denote statistical significance at the 1%, 5%, and 10% level. Color increases in intensity at each 0.05 of magnitude.

Table A.3: Comparison of Pearson and Spearman Correlations

	Correlation Between Self-Assessment and Incentivized Elicitation			
	Pearson		Spearman	
	Replication	Falk et al. (2023)	Replication	Falk et al. (2023)
Risk Tolerance	0.09^{*}	0.32^{***}	0.10^{**}	0.35^{***}
	(.046)	(.047)	(.043)	(.046)
Impatience	-0.01	0.08	0.03	0.05
	(.062)	(.053)	(.052)	(.052)
Altruism	0.14^{***}	0.39^{***}	0.13	0.38^{***}
	(.043)	(.039)	(.045)	(.044)
Trust	0.15^{***}	0.27^{***}	0.13^{**}	0.28^{***}
	(.056)	(.047)	(.054)	(.048)
Reciprocity	0.16^{**}	0.22^{***}	0.13^{**}	0.21^{***}
	(.065)	(.049)	(.053)	(.049)
Punishment	0.08^{*}	0.17^{***}	0.08	0.16^{***}
	(.047)	(.063)	(.049)	(.053)

Notes: Replication refers to correlations within one month, as in Figure 3 ($N = 480$). Estimates for Falk et al. (2023) are based on their replication dataset. Bootstrapped standard errors in parentheses. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

Table A.4: Correlations between Incentivized Elicitations, using ORIV

	Risk Tolerance	Impatience	Altruism	Trust	Reciprocity	Punishment
Risk Tolerance		-0.15^{***} (.038)	0.05 (.041)	0.07^{*} (.042)	-0.00 (.040)	0.04 (.035)
Impatience	-0.15^{***} (.037)		-0.13^{***} (.039)	-0.17^{***} (.037)	-0.11^{***} (.037)	-0.07^{**} (.034)
Altruism	0.05 (.041)	-0.13^{***} (.040)		$1.02^{***,\dagger}$ (.030)	0.54^{***} (.040)	0.19^{***} (.042)
Trust	0.07^{*} (.041)	-0.17^{***} (.038)	$1.02^{***,\dagger}$ (.030)		0.67^{***} (.034)	0.18^{***} (.040)
Reciprocity	-0.00 (.040)	-0.11^{***} (.038)	0.54^{***} (.040)	0.67^{***} (.034)		0.22^{***} (.035)
Punishment	0.04 (.036)	-0.07^{**} (.034)	0.19^{***} (.041)	0.18^{***} (.040)	0.22^{***} (.035)	

Notes: Data from U.S. Study 1, Week 0 ($N = 1,950$). Correlations are using ORIV, with bootstrapped standard errors. ***, **, * denote statistical significance at the 1%, 5%, and 10% level. † unlike a standard correlation coefficient, correlations estimated by ORIV do not have an upper bound of 1.

Table A.5: Correlations between Qualitative Self-Assessments, using ORIV

	Risk Tolerance	Impatience	Altruism	Trust	Reciprocity	Punishment
Risk Tolerance		0.16^{***} (.035)	0.19^{***} (.034)	0.27^{***} (.037)	0.20^{***} (.035)	0.30^{***} (.037)
Impatience	0.16^{***} (.035)		-0.03 (.033)	0.06^{*} (.034)	-0.02 (.032)	0.17^{***} (.036)
Altruism	0.19^{***} (.034)	-0.03 (.033)		0.44^{***} (.028)	0.72^{***} (.033)	0.15^{***} (.036)
Trust	0.27^{***} (.036)	0.06^{*} (.034)	0.44^{***} (.028)		0.39^{***} (.032)	0.14^{***} (.037)
Reciprocity	0.20^{***} (.035)	-0.02 (.032)	0.72^{***} (.033)	0.39^{***} (.032)		0.22^{***} (.037)
Punishment	0.30^{***} (.036)	0.17^{***} (.037)	0.15^{***} (.036)	0.14^{***} (.038)	0.22^{***} (.038)	

Notes: Data from U.S. Study 1, Week 0 ($N = 1,950$). Correlations are using ORIV, with bootstrapped standard errors. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

Table A.6: Correlations between Qualitative Self-Assessments and Incentivized Elicitations, by Subgroup

	Risk	Impatience	Altruism	Trust	Reciprocity	Punishment
All	0.10^{***}	-0.00	0.16^{***}	0.08^{***}	0.13^{***}	0.06^{***}
	(.022)	(.022)	(.021)	(.019)	(.023)	(.020)
			$N = 4{,}950$			
ICAR: Above Median	0.16^{***}	-0.04	0.16^{***}	0.08^{***}	0.15^{***}	0.12^{***}
	(.026)	(.029)	(.029)	(.026)	(.029)	(.025)
			$N = 2{,}816$			
ICAR: Top $\sim 10\%$	0.16^{***}	0.01	0.26^{***}	0.16^{***}	0.18^{***}	0.18^{***}
	(.056)	(.061)	(.056)	(.061)	(.054)	(.054)
			$N = 561$			
ICAR: Top $\sim 5\%$	0.20^{**}	0.12	0.38^{***}	0.17^{**}	0.15^{**}	0.26^{***}
	(.098)	(.090)	(.073)	(.074)	(.075)	(.067)
			$N = 235$			
CRT: No Questions Correct	0.07^{**}	0.01	0.14^{***}	0.06^{**}	0.11^{***}	0.05^{*}
	(.029)	(.028)	(.028)	(.026)	(.032)	(.027)
			$N = 2{,}725$			
CRT: One or More Questions Correct	0.15^{***}	-0.00	0.19^{***}	0.11^{***}	0.17^{***}	0.10^{***}
	(.029)	(.035)	(.030)	(.029)	(.029)	(.029)
			$N = 2{,}225$			
CRT: All Three Questions Correct	0.18^{***}	-0.06	0.22^{***}	0.18^{***}	0.17^{***}	0.05
	(.053)	(.083)	(.058)	(.055)	(.049)	(.071)
			$N = 500$			
High School or Less	0.08^{**}	0.04	0.15^{***}	0.09^{***}	0.14^{***}	0.04
	(.038)	(.037)	(.034)	(.031)	(.040)	(.035)
			$N = 1{,}689$			
Some College or College Degree	0.12^{***}	-0.06^{**}	0.18^{***}	0.08^{***}	0.13^{***}	0.07^{***}
	(.025)	(.027)	(.028)	(.024)	(.028)	(.025)
			$N = 2{,}690$			
Advanced Degree	0.12^{**}	0.08	0.08	-0.03	0.10^{*}	0.19^{***}
	(.054)	(.054)	(.074)	(.071)	(.059)	(.049)
			$N = 571$			
Response Time: Not Fastest 10%	0.11^{***}	-0.02	0.16^{***}	0.07^{***}	0.13^{***}	0.06^{***}
	(.023)	(.023)	(.022)	(.020)	(.023)	(.021)
			$N = 4{,}504$			
Response Time: Not Slowest or Fastest 25%	0.11^{***}	-0.05	0.15^{***}	0.05^{**}	0.11^{***}	0.07^{***}
	(.030)	(.033)	(.030)	(.025)	(.031)	(.028)
			$N = 2{,}495$			

Notes: Data from combined U.S. Samples 1 and 2 ($N = 4,950$). Bootstrapped standard errors in parentheses. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

Table A.7: Correlations between Qualitative Self-Assessments and Incentivized Elicitations, by Subgroup (Continued)

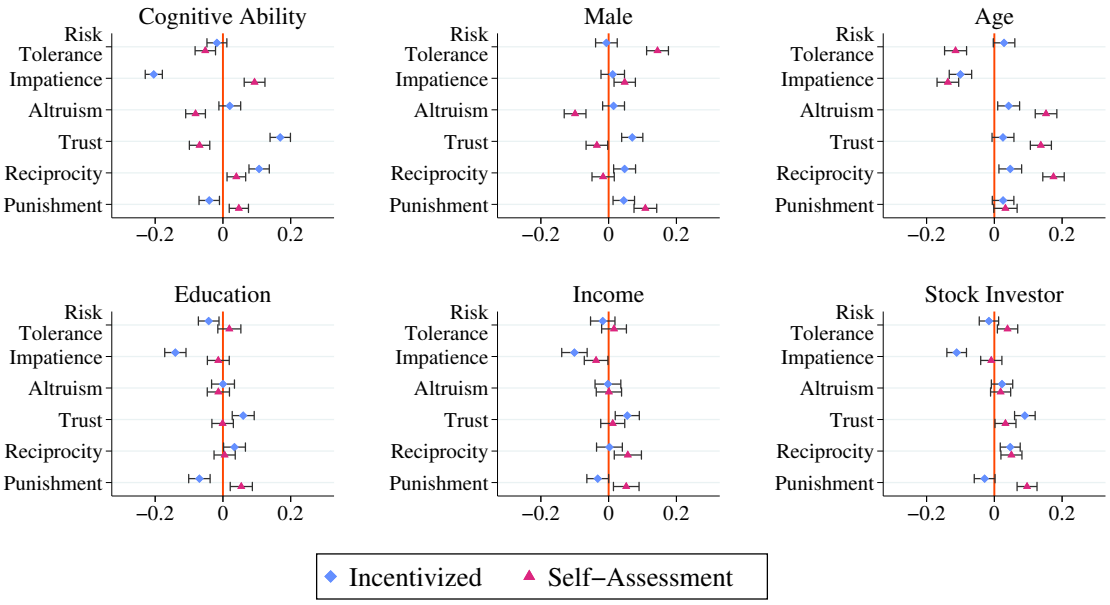
	Risk	Impatience	Altruism	Trust	Reciprocity	Punishment
All	0.10***	−0.00	0.16***	0.08***	0.13***	0.06***
	(.022)	(.022)	(.021)	(.019)	(.023)	(.020)
				$N = 4{,}950$		
Female	0.10***	−0.02	0.11***	0.06***	0.11***	0.04
	(.029)	(.028)	(.030)	(.025)	(.027)	(.026)
				$N = 2{,}688$		
Male	0.11***	0.02	0.21***	0.09***	0.15***	0.08**
	(.032)	(.034)	(.029)	(.029)	(.037)	(.032)
				$N = 2{,}262$		
Investor	0.10***	−0.03	0.09**	0.03	0.10***	0.08**
	(.034)	(.035)	(.039)	(.032)	(.036)	(.033)
				$N = 1{,}720$		
Not Investor	0.10***	0.00	0.19***	0.09***	0.14***	0.06**
	(.027)	(.027)	(.024)	(.023)	(.029)	(.025)
				$N = 3{,}230$		
Age: Under 40	0.06	0.03	0.20***	0.06*	0.10**	0.07*
	(.038)	(.039)	(.033)	(.033)	(.042)	(.038)
				$N = 1{,}694$		
Age: 40 to 60	0.14***	−0.06*	0.10**	0.10***	0.18***	0.06*
	(.035)	(.036)	(.038)	(.033)	(.032)	(.033)
				$N = 1{,}682$		
Age: Over 60	0.11***	−0.02	0.16***	0.07**	0.10***	0.05*
	(.036)	(.029)	(.038)	(.029)	(.034)	(.032)
				$N = 1{,}574$		
Above Median Income	0.11***	−0.05	0.13***	0.05*	0.13***	0.05*
	(.030)	(.029)	(.030)	(.029)	(.026)	(.027)
				$N = 2{,}491$		
Above 90% Income	0.15***	0.00	0.03	−0.00	0.17***	0.06
	(.054)	(.059)	(.071)	(.059)	(.048)	(.047)
				$N = 646$		
Above 95% Income	0.10	0.03	−0.06	−0.08	0.17***	0.09
	(.063)	(.041)	(.086)	(.077)	(.065)	(.064)
				$N = 395$		

Notes: Data from combined U.S. Samples 1 and 2 ($N = 4{,}950$). Bootstrapped standard errors in parentheses. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

A.3 Correlations with Demographics

Figure A.3 displays univariate correlations between preferences and various individual characteristics.

Figure A.3: Incentivized elicitations and qualitative self-assessments exhibit different correlations with individual characteristics.



Notes: Data from combined U.S. Samples 1 and 2 ($N = 4,950$). Each panel displays univariate correlations between a demographic characteristic and each preference measure. Bars represent 90% confidence intervals.

Tables A.8 and A.9 display results for the same regressions as in Table 5 for the other four preference domains. In each case, the estimated relationships between qualitative self-assessments and demographic factors are essentially unchanged after controlling for incentivized elicitations. The only exception is the coefficient relating to cognitive ability and reciprocity, which becomes smaller and statistically insignificant.

Table A.10 shows results similar to those in Table 5 for risk tolerance and impatience using our two combined U.S. samples. For these two domains, we have duplicate elicitations in both samples. The results are similar to those reported in the main text. The only noteworthy difference is that the correlation between qualitative risk tolerance and stock investment is now statistically significant, as documented by Dohmen et al. (2011). However, as with the

Table A.8: Relationships between self-assessed impatience and trust and demographics are unrelated to variation in incentivized elicitations (N = 1,950).

	Dependent Variable = Qualitative Self-Assessment					
	Impatience			Trust		
Cognitive Ability	0.08***	0.08***	0.06**	−0.04	−0.08***	−0.06*
	(.029)	(.029)	(.030)	(.028)	(.027)	(.029)
Male	0.06	0.06	0.05	−0.05	−0.07	−0.06
	(.056)	(.056)	(.056)	(.056)	(.056)	(.056)
Age	−0.14***	−0.15***	−0.14***	0.16***	0.16***	0.17***
	(.029)	(.029)	(.029)	(.029)	(.028)	(.028)
Education	−0.05	−0.05	−0.05	−0.01	−0.01	−0.00
	(.031)	(.031)	(.031)	(.033)	(.032)	(.031)
Income (Log)	−0.05	−0.05	−0.07**	0.01	0.00	0.01
	(.035)	(.035)	(.034)	(.033)	(.034)	(.035)
Stock Investor	0.06	0.06	0.05	0.10	0.09	0.10
	(.069)	(.068)	(.067)	(.067)	(.066)	(.066)
Incentivized Elicitation:						
Risk Tolerance			−0.03			0.05
			(.037)			(.036)
Impatience		−0.01	−0.00			0.07*
		(.036)	(.036)			(.036)
Altruism			−0.05			0.21*
			(.116)			(.120)
Trust			0.26**		0.21***	0.11
			(.132)		(.047)	(.133)
Reciprocity			−0.14**			−0.09
			(.056)			(.054)
Punishment			0.01			−0.05
			(.039)			(.038)

Notes: All columns use data from U.S. Sample 1. Incentivized elicitations are instrumented to eliminate the effect of classical measurement error. Coefficients and standard errors (in parentheses) on all continuous measures are standardized. All specifications include an indicator variable for missing income. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

other demographic factors, this relationship is not driven by preferences captured by the corresponding incentivized elicitation.

Figure A.4 compares the average preferences within our low-education U.K. Sample to

Table A.9: Relationships between demographics and self-assessed reciprocity and willingness-to-punish are unrelated to variation in incentivized elicitations (N = 1,950).

	Dependent Variable = Qualitative Self-Assessment					
	Reciprocity			Punishment		
Cognitive Ability	0.09***	0.06**	0.04	−0.01	−0.00	−0.00
	(.029)	(.028)	(.031)	(.027)	(.027)	(.029)
Male	−0.02	−0.04	−0.05	0.26***	0.25***	0.25***
	(.057)	(.057)	(.056)	(.059)	(.059)	(.059)
Age	0.20***	0.18***	0.18***	−0.02	−0.03	−0.03
	(.029)	(.029)	(.028)	(.031)	(.030)	(.030)
Education	−0.05	−0.05	−0.04	0.02	0.03	0.03
	(.030)	(.029)	(.029)	(.032)	(.032)	(.032)
Income (Log)	0.03	0.04	0.02	−0.02	−0.03	−0.03
	(.035)	(.033)	(.034)	(.036)	(.035)	(.036)
Stock Investor	0.08	0.08	0.06	0.12*	0.12*	0.12*
	(.063)	(.060)	(.058)	(.067)	(.069)	(.069)
Incentivized Elicitation:						
Risk Tolerance			−0.04			0.03
			(.039)			(.041)
Impatience			−0.06			0.00
			(.037)			(.037)
Altruism			0.24**			0.06
			(.111)			(.125)
Trust			0.07			0.02
			(.124)			(.143)
Reciprocity		0.21***	0.04			−0.04
		(.046)	(.053)			(.058)
Punishment			0.05		0.15***	0.14***
			(.036)		(.042)	(.041)

Notes: All columns use data from U.S. Sample 1. Incentivized elicitations are instrumented to eliminate the effect of classical measurement error. Coefficients and standard errors (in parentheses) on all continuous measures are standardized. All specifications include an indicator variable for missing income. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

those across the whole U.S. population.

Table A.10: Findings regarding the relationships between qualitative self-assessments and demographics are similar using combined U.S. Samples ($N = 4{,}950$).

	Dependent Variable = Qualitative Self-Assessment			
	Risk Tolerance		Impatience	
Incentivized Measure		0.13^{***}		-0.01
		(.026)		(.026)
Cognitive Ability	-0.12^{***}	-0.12^{***}	0.09^{***}	0.09^{***}
	(.019)	(.019)	(.019)	(.020)
Male	0.31^{***}	0.31^{***}	0.05	0.05
	(.039)	(.039)	(.039)	(.039)
Age	-0.13^{***}	-0.14^{***}	-0.13^{***}	-0.13^{***}
	(.021)	(.021)	(.020)	(.020) ’
Education	0.04^{**}	0.05^{**}	-0.03	-0.03
	(.021)	(.021)	(.021)	(.021)
Income (Log)	0.01	0.01	-0.04^{*}	-0.04^{*}
	(.024)	(.024)	(.023)	(.023)
Stock Investor	0.12^{***}	0.13^{***}	0.05	0.05
	(.043)	(.043)	(.044)	(.044)

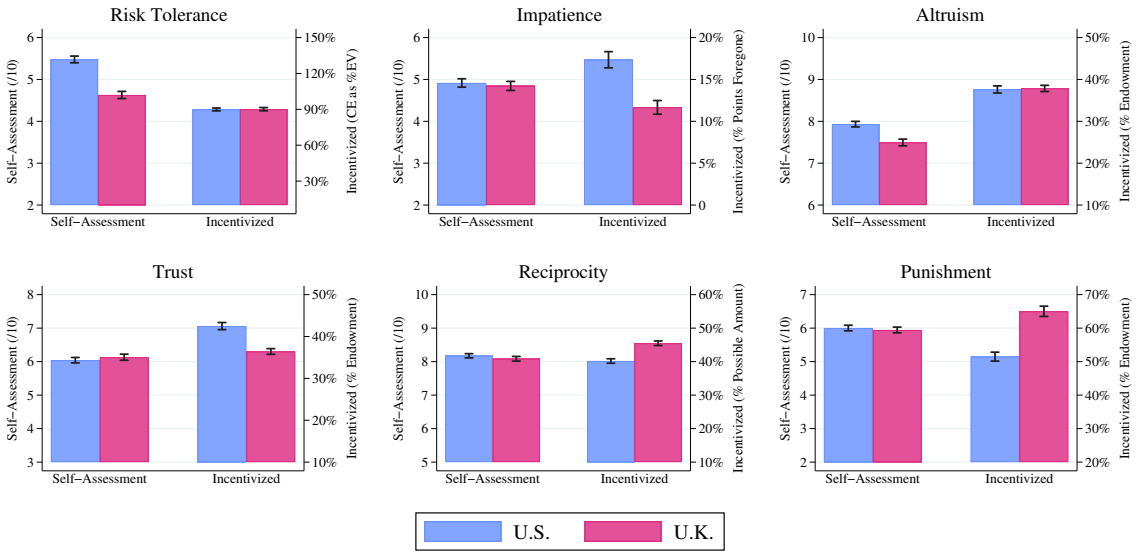
Notes: All columns use data from U.S. Samples 1 and 2. Incentivized elicitations are instrumented to eliminate the effect of classical measurement error. Coefficients and standard errors (in parentheses) on all continuous measures are standardized. All specifications include an indicator variable for missing income. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

A.4 Correlations with Self-Reported Behaviors

In this section we examine correlations between qualitative self-assessments and self-reported behaviors, for a subset of U.S. Sample 2. Data for part of this sample was collected as part of a five-wave survey. The main text analyzes the first wave. In waves 3 to 5 of the survey, we elicited self-reported behaviors relating to charitable giving, community engagement, health behaviors, and political interest.[1] Some previous studies have suggested that self-assessments may be valuable based on their ability to predict such activities. Here, we examine the relationships between these behaviors, qualitative self-assessments, and incentivized preference measures in this survey.

[1]YouGov provides separate probability weights for each survey wave. We use sample weights from the first survey in waves 3 to 5 that a participant completed.

Figure A.4: Comparison of low-education U.K. Sample to U.S. general population.



Notes: The figure compares the average preference within the two U.S. representative samples ($N = 4,950$) and the U.K. Sample ($N = 1,984$).

Table A.11: Relationships between qualitative self-assessments and behaviors.

	Risk	Impatience	Altruism	Trust	Reciprocity	Punishment
Smoke	0.30***	0.08	0.16**	−0.04	0.21***	0.07
	(.078)	(.091)	(.079)	(.102)	(.074)	(.076)
Binge Drink	0.43***	0.20**	−0.02	0.18**	0.08	0.17**
	(.091)	(.082)	(.085)	(.079)	(.069)	(.078)
Eat 5 Fruit/Veg A Day	0.32***	−0.12*	0.07	0.11*	0.01	0.01
	(.064)	(.063)	(.064)	(.063)	(.065)	(.060)
Exercise	0.38***	−0.19***	0.15**	0.13**	0.13**	0.08
	(.063)	(.062)	(.066)	(.060)	(.065)	(.059)
Attended Local Political Meeting	0.32***	0.18**	0.27***	0.18**	0.22***	0.27***
	(.093)	(.078)	(.077)	(.085)	(.072)	(.078)
Put Up Political Sign	0.19**	0.15**	0.26***	0.08	0.30***	0.23***
	(.079)	(.077)	(.068)	(.082)	(.064)	(.073)
Worked for Political Campaign	0.33***	0.30***	0.28***	0.17	0.13	0.18*
	(.110)	(.111)	(.098)	(.119)	(.102)	(.094)
Donated to Campaign	0.30***	0.18***	0.29***	0.14**	0.33***	0.24***
	(.068)	(.065)	(.066)	(.068)	(.052)	(.058)
Donated Blood	0.40***	0.08	0.30***	0.20**	0.26***	0.10
	(.097)	(.101)	(.086)	(.096)	(.089)	(.095)
Worked for Community	0.36***	0.17***	0.41***	0.09	0.26***	0.18***
	(.069)	(.065)	(.057)	(.071)	(.060)	(.067)
Contacted Govt Official	0.19***	0.23***	0.34***	0.03	0.26***	0.22***
	(.063)	(.063)	(.056)	(.069)	(.056)	(.051)
Community Meeting	0.45***	0.19***	0.37***	0.19**	0.28***	0.19***
	(.074)	(.068)	(.058)	(.074)	(.058)	(.070)
Volunteer Work	0.00	0.00	0.01***	0.00	0.01***	0.00
	(.002)	(.002)	(.003)	(.002)	(.003)	(.002)
Church/Charity Contribution	0.23***	−0.08	0.59***	0.31***	0.44***	−0.03
	(.066)	(.061)	(.061)	(.058)	(.062)	(.058)
Church Attendance	0.12***	−0.00	0.25***	0.18***	0.12***	−0.01
	(.033)	(.033)	(.030)	(.031)	(.028)	(.030)
# Organizations Member of	0.17***	0.08***	0.15***	0.09***	0.11***	0.09***
	(.026)	(.025)	(.029)	(.026)	(.029)	(.025)
#Days Talked Politics	0.07**	0.02	0.20***	0.09***	0.23***	0.06**
	(.030)	(.029)	(.034)	(.030)	(.027)	(.026)
Read Blog	0.14**	0.22***	0.13**	0.05	0.20***	0.08
	(.063)	(.061)	(.061)	(.063)	(.059)	(.059)
Watched TV News	0.08	−0.11*	0.35***	0.31***	0.31***	0.11*
	(.065)	(.062)	(.072)	(.065)	(.070)	(.063)
Read Newspaper	0.26***	−0.03	0.21***	0.24***	0.28***	0.02
	(.060)	(.060)	(.058)	(.058)	(.057)	(.053)
Listened to Radio	0.19***	0.12**	0.23***	0.14**	0.13**	0.12**
	(.064)	(.057)	(.059)	(.054)	(.064)	(.053)

Notes: All columns use data from U.S. Sample 2, waves 3–5 ($N = 4,134$). Coefficients and standard errors (in parentheses) on all continuous measures are standardized. Standard errors are clustered by participant. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

First, Table A.11 presents simple correlations between each of the qualitative self-assessments and different self-reported activities. Most variables in this table are binary, taking a value of one if a participant reported an activity, and zero otherwise. Exceptions include church attendance, the (square-root of the) number of organizations a participant is a member of, and the number of days a participant spent discussing politics, where these variables are all standardized. Since some individuals responded to multiple waves of the survey, we cluster standard errors by participant.

As can be seen, qualitative self-assessments are correlated with a wide range of behaviors, sometimes in unexpected ways. For instance, we replicate the finding of Dohmen et al. (2011) that self-reported willingness to take risks is correlated with smoking.[2] However, it is also correlated with almost every other behavior in our dataset, including the seemingly low risk behavior of eating five portions of fruit and vegetables per day. We also see that the social preference measures are correlated with a wide range of activities relating to charitable activity, engagement with local politics, and political interest.

To aid interpretation of this large battery of variables, we carry out two principal components analyses—see Figure A.5 and Tables A.12–A.13. In each analysis, we determine the number of components to retain using parallel analysis. The first principal components analysis includes the four variables, at the top of Table A.11, related to health behaviors. This analysis, reported in Table A.12, identifies two intuitive components—*Unhealthy Behaviors* (primarily capturing smoking and drinking more than five alcoholic drinks) and *Healthy Behaviors* (loading on exercise, eating five fruit and/or vegetables a day). The second analysis, reported in Table A.13, includes the remaining variables, which relate to involvement in a participant's local community, volunteering or giving to charity, political activity, and media use. This identifies three components. *Political Engagement* primarily relates to political activities such as working for a political candidate or working to tackle issues in the local community. *Charity/Community* relates primarily to volunteering, religious attendance, and

[2]See Arslan et al. (2020, Supplementary Materials S1) for a general review of studies assessing relationships between the qualitative self-assessment of risk and behaviors.

Figure A.5: Scree Plots for Principal Components Analyses.

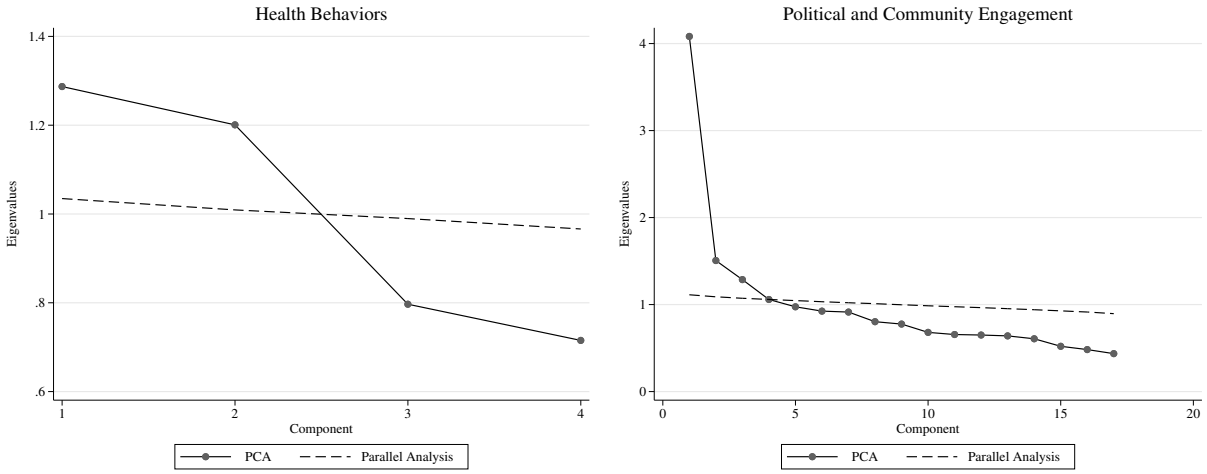


Table A.12: Principal Components Analysis for Health Behaviors.

	Unhealthy Behaviors	*Healthy Behaviors*	Unexplained
Do you currently...			
Smoke cigarettes, cigars, or pipes	0.72	−0.12	0.34
Drink five or more alcoholic drinks occasionally	0.70	0.13	0.36
Exercise	0.00	0.69	0.42%
Eat five or more servings of fruits and/or vegetables a day	−0.01	0.70	0.40
Percent of Variation	32%	31%	38%

Notes: The principal components analysis used data from waves 3–5 of U.S. Sample 2 ($N = 4,134$). Each question offered participants a choice of yes or no.

contributing to church and charity. *Political Interest* captures variance relating to the use of media, and discussing politics with friends and family.

Figure A.6 shows correlations between these five principal components and both incentivized elicitations and qualitative self-assessments. Both types of measures are correlated with several behavioral variables, but the patterns are markedly different, particularly for

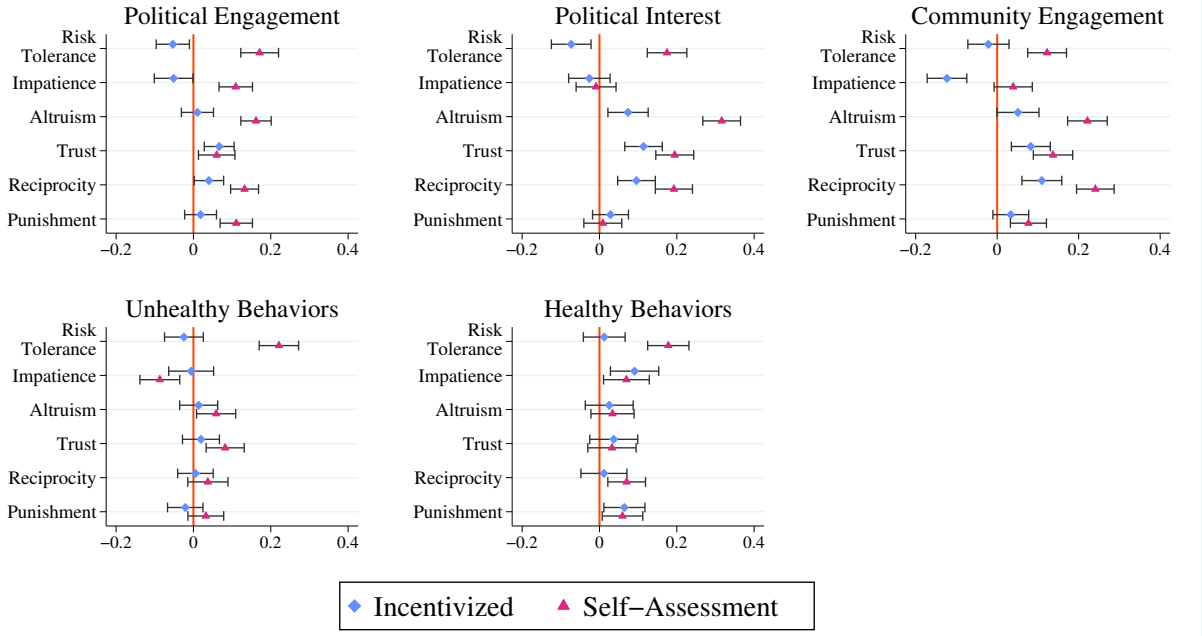
Table A.13: Principal Components Analysis for Political and Community Activities.

	Political Engagement	*Charity/ Community*	*Political Interest*	Unexplained
During the past year did you...				
Attend local political meetings	0.43	−0.03	−0.05	0.48
Put up a political sign	0.33	0.07	−0.16	0.65
Work for a candidate or campaign	0.40	−0.07	−0.08	0.57
Donate money to a candidate, campaign, or political organization	0.32	0.15	−0.14	0.60
Donate blood	0.14	−0.12	0.21	0.85
Indicate whether in past 12 months you...				
Worked with other people to deal with some issue facing your community	0.35	−0.01	0.14	0.52
Telephoned, wrote a letter to, or visited a government official to express your views on a public issue	0.28	0.16	−0.01	0.61
Attended a meeting about an issue facing your community or schools	0.36	−0.05	0.12	0.54
In past 12 months devoted time or made contributions...				
Did volunteer work	−0.04	0.14	0.08	0.95
Contributed to church and charity	0.01	0.10	0.56	0.36
How often attend church	−0.04	−0.05	0.65	0.33
# of organizations a member of	0.25	−0.02	0.31	0.55
In past 24 hours have you.....				
Read a blog	−0.03	0.34	0.13	0.71
Watched TV news	0.05	0.40	−0.12	0.66
Read a newspaper in print or online	−0.00	0.47	0.01	0.55
Listened to a radio news program or talk radio	−0.16	0.41	0.08	0.69
# Days in last week discussed politics with friends/family	0.06	0.47	−0.01	0.50
Percent of Variation	18%	12%	10%	60%

Notes: The principal components analysis used data from waves 3–5 of U.S. Sample 2 ($N = 4,134$).

risk and impatience. Interestingly, however, both incentivized and self-assessment measures of altruism, trust, and reciprocity have predictive power for behaviors related to political engagement, political interest, and charitable giving/community engagement. Bigger dif-

Figure A.6: Correlations between principal components of self-reported behaviors and preferences.



Notes: The figure uses data from U.S. Sample 2, waves 3–5 ($N = 4,134$). Standard errors are clustered by participant. Bars represent 90% confidence intervals.

ferences appear in the realm of risk and, to a lesser extent, impatience. This indicates that incentivized measures may have better predictive power than might be suggested by examining risk measures alone (Charness et al., 2020).

Finally, in Table A.14, we present regression results similar to those in Table 5. That table indicated whether correlations between demographic characteristics and qualitative self-assessments capture behavior on incentivized elicitations. Here, we investigate whether the correlations with demographics are explained by a participant's behavior outside of the survey—if so, this could suggest that self-assessments are capturing some other latent factor that predicts behavior, but is not captured in incentivized elicitations. We also test whether the correlations between qualitative self-assessments and self-reported behaviors are explained by the incentivized elicitation. The surveys that included questions about behaviors included only a single incentivized elicitation of altruism (as with the initial survey in U.S. Study 2 that we study in the main text). As such, to account for measurement error,

we use the incentivized elicitation of trust as an instrument for altruism. The first stage F-statistic is 96.2, suggesting that this is appropriate.[3]

The results in Table A.14 show that, consistent with the correlations in Table A.11, the risk and altruism measures are correlated with a range of behaviors, not only those most obviously related to each preference. Second, with the possible exception of education, the coefficients related to demographic variables are little affected by the inclusion of self-reported behaviors. This suggests that the correlations between demographics and self-assessments are not reflected in different behaviors outside the survey. For instance, if women appear more altruistic based on self-assessments, this does not translate into increased giving in the dictator game or higher (self-reported) charitable contributions. Third, the correlations between self-reported behaviors and self-assessments are close to unchanged after controlling for the incentivized measures. That is, the variation between self-reported behaviors and qualitative self-assessments is not correlated with variation in the incentivized elicitations.

B Data Sources and Screenshots

B.1 General Population Datasets

Our general population datasets are drawn from a number of incentivized online surveys. Each survey was conducted by YouGov, a commercial survey company. Participants in the three representative samples were drawn from YouGov's two-million-person survey panel. YouGov obtains nationally representative samples by using targeted quota sampling and then constructing sample weights to account for the fact that some demographic groups are underrepresented. See Chapman et al. (2024b) for further discussion of the composition of the YouGov survey panel and payment procedure. Typically, each survey took around 45 minutes to one hour to complete, with payment approximately three times the average

[3]The incentivized risk tolerance elicitation included in these surveys was slightly different from the elicitation in the original survey of U.S. Sample 2, which we analyze in the main text.

Table A.14: Relationships between qualitative self-assessments, demographics, and behaviors.

	Dependent Variable = Qualitative Self-Assessment					
	Risk Tolerance			Altruism		
Cognitive Ability	-0.12^{***}	-0.10^{***}	-0.09^{***}	0.01	0.01	0.00
	(.031)	(.028)	(.028)	(.030)	(.029)	(.029)
Male	0.23^{***}	0.18^{***}	0.17^{***}	-0.20^{***}	-0.22^{***}	-0.23^{***}
	(.069)	(.063)	(.063)	(.067)	(.060)	(.059)
Age	-0.08^{**}	-0.10^{***}	-0.10^{***}	0.13^{***}	0.05	0.04
	(.033)	(.033)	(.033)	(.031)	(.031)	(.031)
Education	0.10^{***}	0.04	0.04	-0.01	-0.09^{***}	-0.08^{***}
	(.035)	(.035)	(.035)	(.032)	(.031)	(.030)
Income (Log)	0.07^{*}	0.03	0.03	0.11^{***}	0.06^{*}	0.07^{**}
	(.037)	(.033)	(.033)	(.037)	(.034)	(.034)
Behaviors:						
Unhealthy Activities		0.15^{***}	0.15^{***}		-0.02	-0.02
		(.031)	(.031)		(.028)	(.028)
Healthy Activities		0.16^{***}	0.16^{***}		0.05^{*}	0.05
		(.031)	(.031)		(.030)	(.030)
Political Engagement		0.08^{***}	0.08^{***}		0.05^{**}	0.06^{**}
		(.023)	(.023)		(.025)	(.025)
Political Interest		0.12^{***}	0.12^{***}		0.27^{***}	0.26^{***}
		(.029)	(.029)		(.030)	(.031)
Charity / Community		0.06^{*}	0.05^{*}		0.12^{***}	0.11^{***}
		(.031)	(.031)		(.035)	(.036)
Incentivized Measure			0.06^{*}			0.13^{***}
			(.033)			(.046)

Notes: All columns use data from U.S. Sample 2, waves 3–5 ($N = 4,134$). Incentivized measures are instrumented to eliminate the effect of classical measurement error. Coefficients and standard errors (in parentheses) on all continuous measures are standardized. All specifications include an indicator variable for missing income. ***, **, * denote statistical significance at the 1%, 5%, and 10% level.

for similar surveys. In each survey, participants were paid for either one or two randomly-selected questions.

B.2 Data Sources and Screenshots

In this section, we provide illustrative screenshots from our various samples. These are broken into two groups. One contains the U.S. Samples, U.K. Sample, and Pitt student sample, which all used the same survey platform, and thus had a similar look and feel, as well as identical incentive levels. The Caltech, UBC, and MTurk samples were on a different survey platform, and those measures thus had a different look and feel.

In the U.S., U.K., and Pitt samples, rewards for incentivized questions were expressed via "points", an internal YouGov currency used to pay panel members. These points can be converted into monetary compensation or prizes, using the approximate rate of $0.001 per point. YouGov allows points to be converted to awards at specific point values, which leads to a slightly convex payoff schedule. This convexity does not appear to distort behavior—see Chapman et al. (2024b, Appendix C.6) for a detailed discussion. In the Caltech, UBC, and MTurk samples, rewards for incentivized questions were expressed via "tokens." At Caltech, each 100 tokens translated to $1, while at UBC, each 100 tokens translated to $1CAD. On MTurk, where payments are generally lower, 300 tokens translated to $1.

Figure B.1: Qualitative Self-Assessment of Risk Tolerance (with 8 selected, U.S. and U.K.)

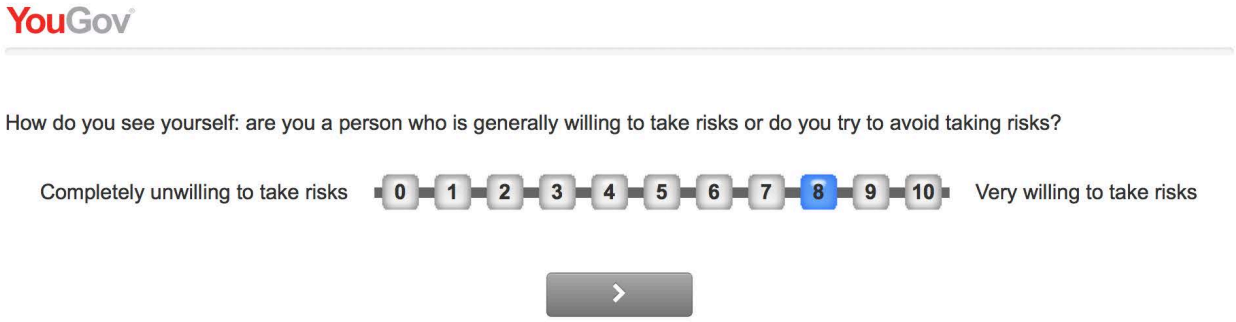


Figure B.2: Qualitative Self-Assessment of Impatience (U.S. and U.K.)

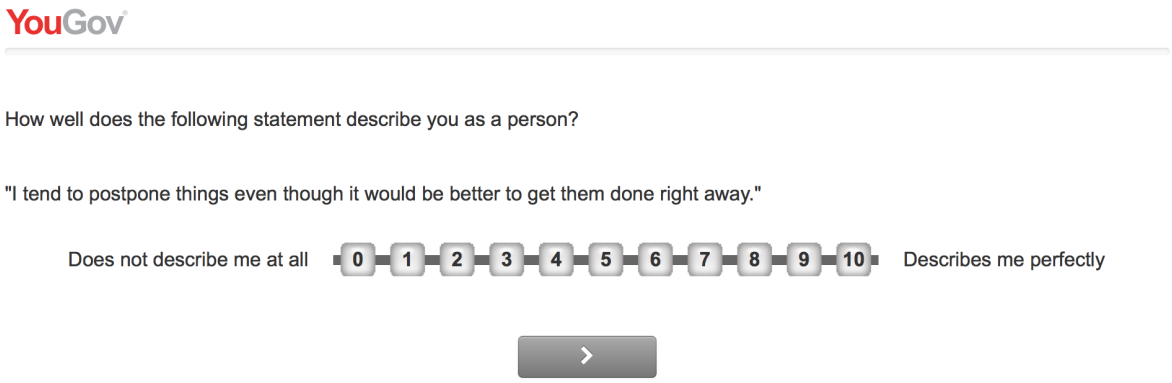


Figure B.3: Qualitative Self-Assessment of Altruism (U.S., U.K., and Pitt)

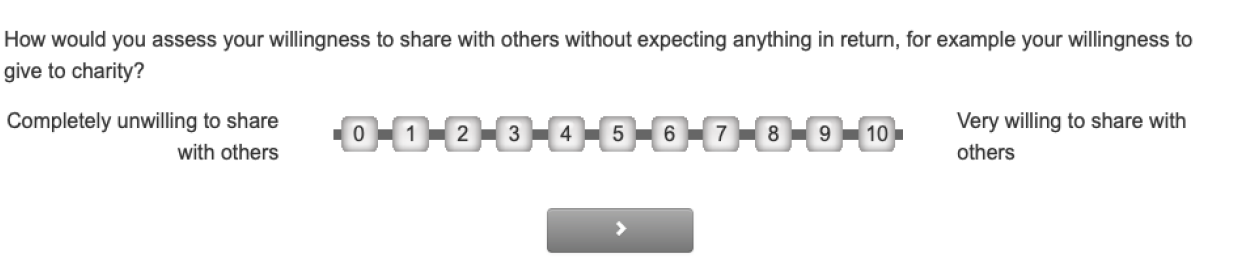


Figure B.4: Qualitative Self-Assessment of Trust (U.S., U.K., and Pitt)

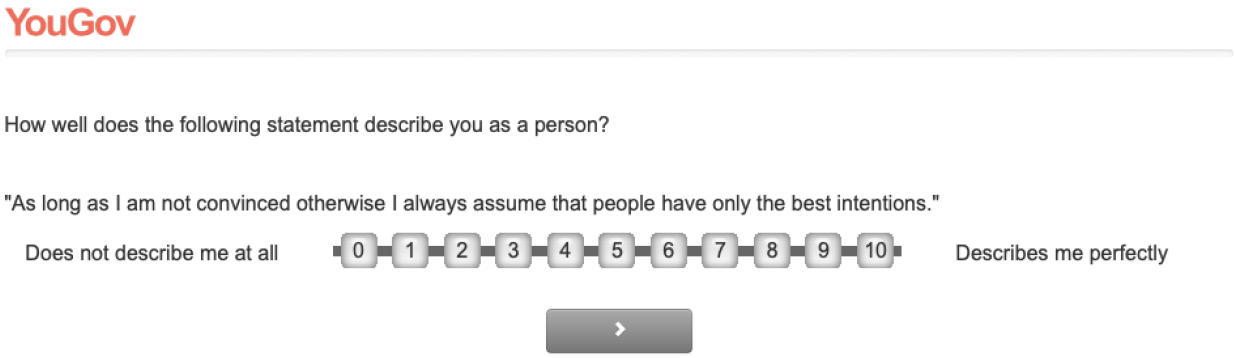


Figure B.5: Qualitative Self-Assessment of Reciprocity (U.S., U.K., and Pitt)

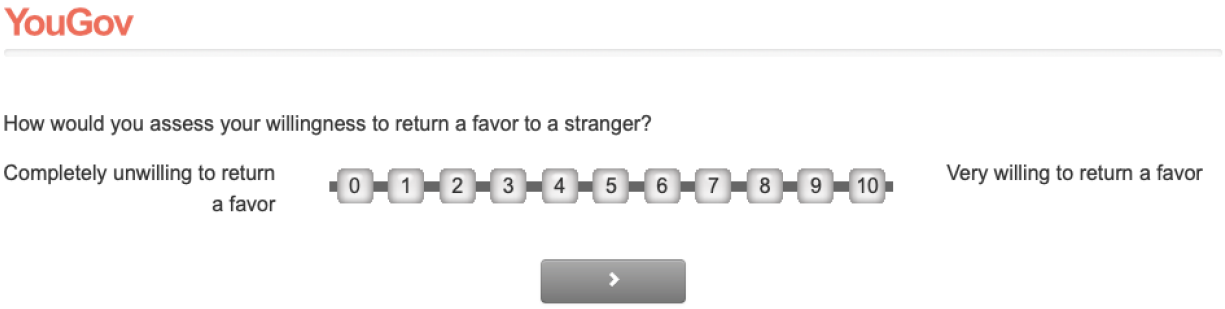


Figure B.6: Qualitative Self-Assessment of Punishment (U.S. and U.K.)

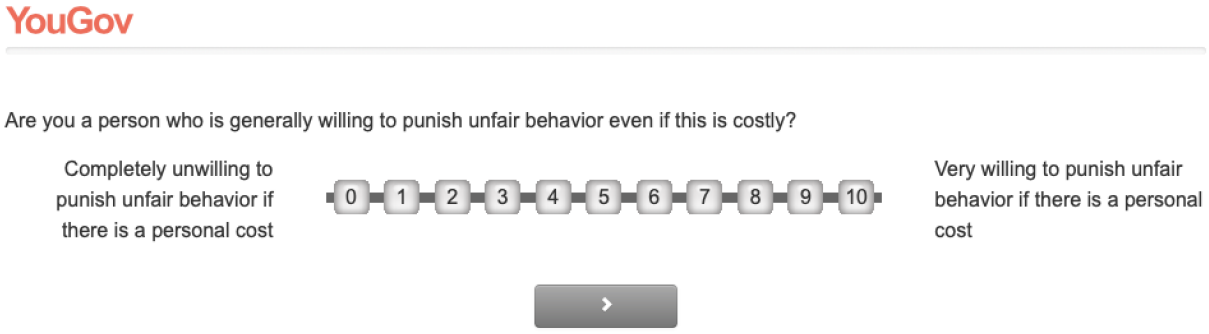


Figure B.7: Incentivized Elicitation of Risk Tolerance (U.S., U.K., and Pitt)

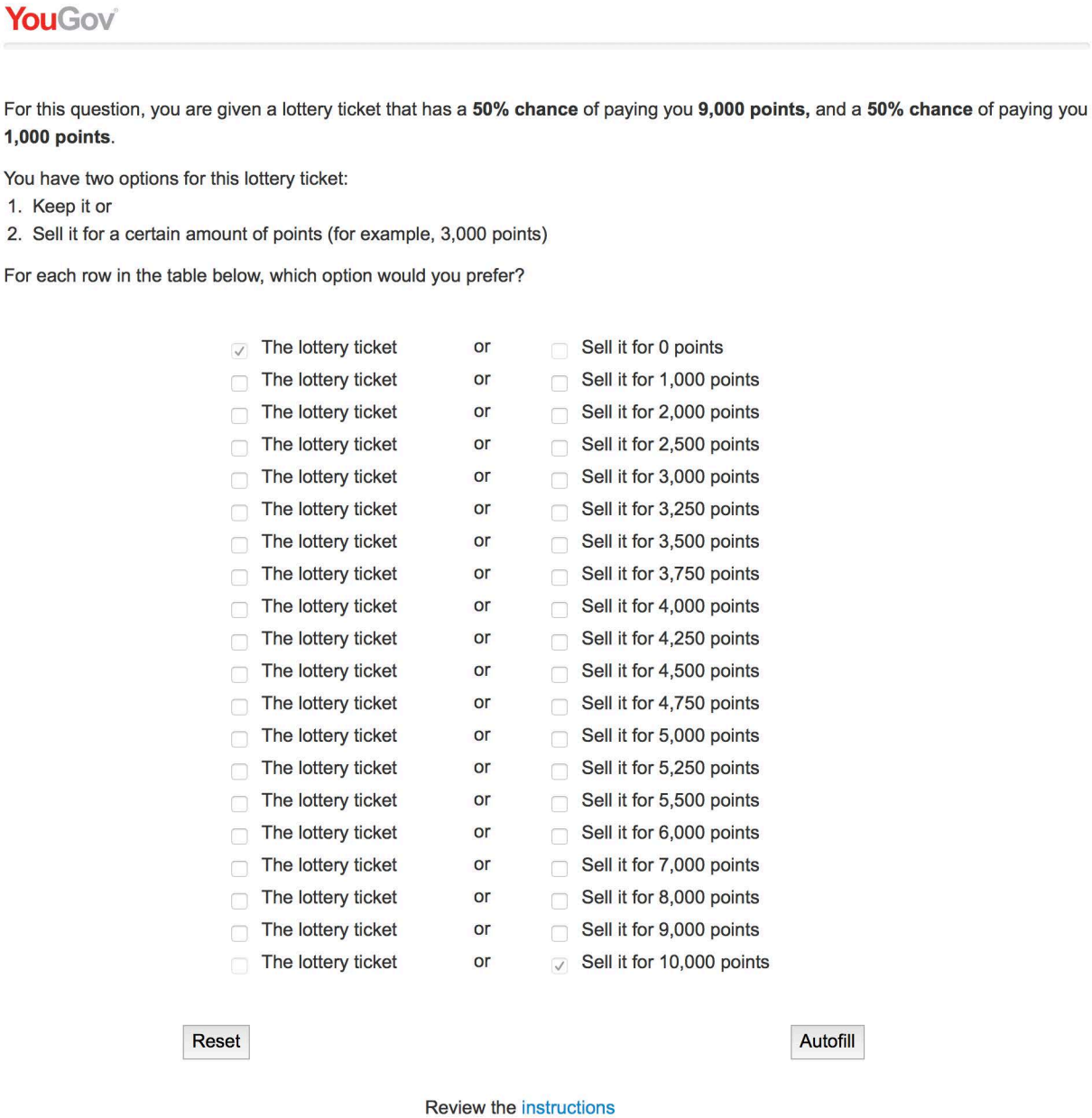


Figure B.8: Incentivized Elicitation of Impatience (U.S. and U.K.)

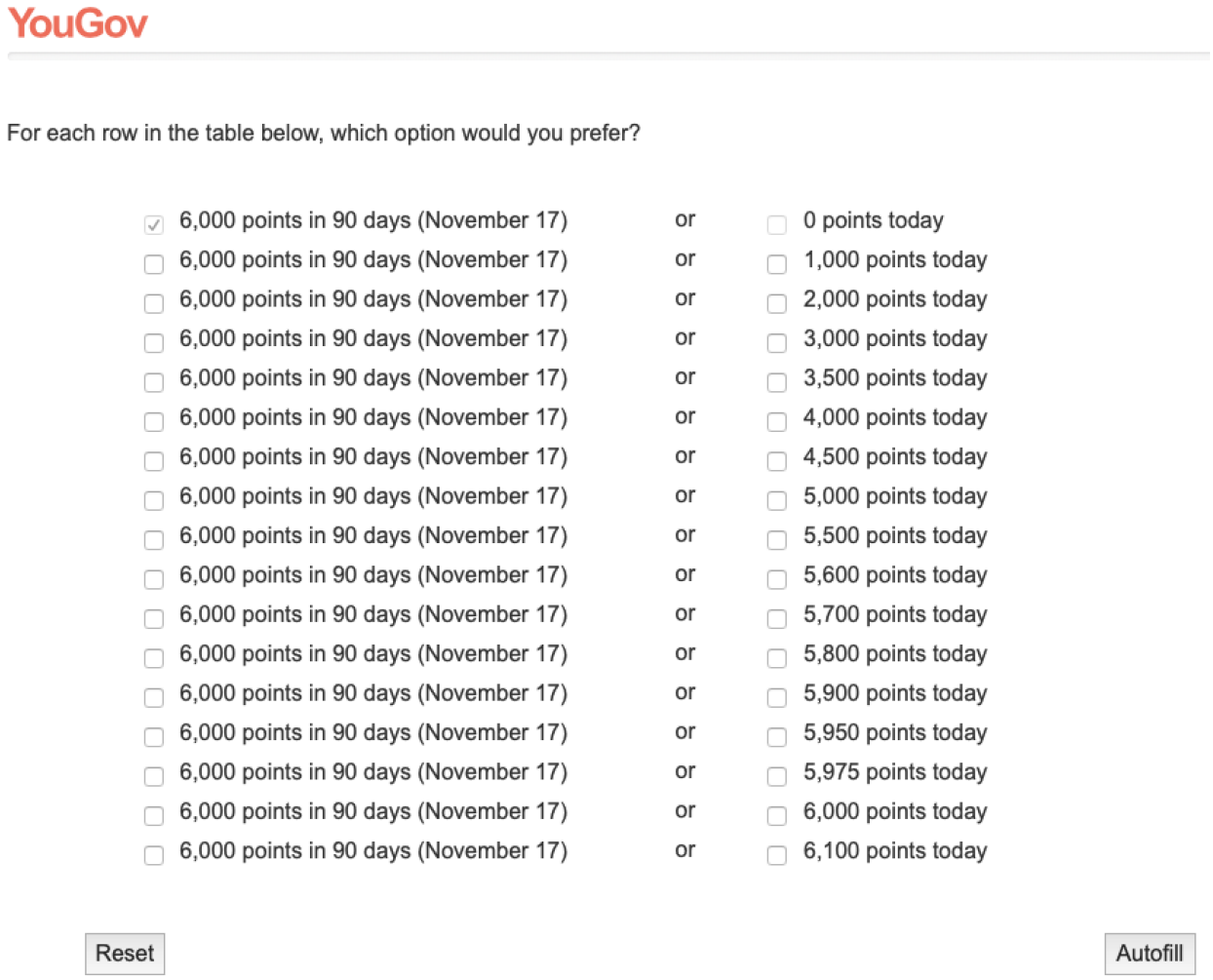


Figure B.9: Incentivized Elicitation of Altruism (U.S. and U.K.)

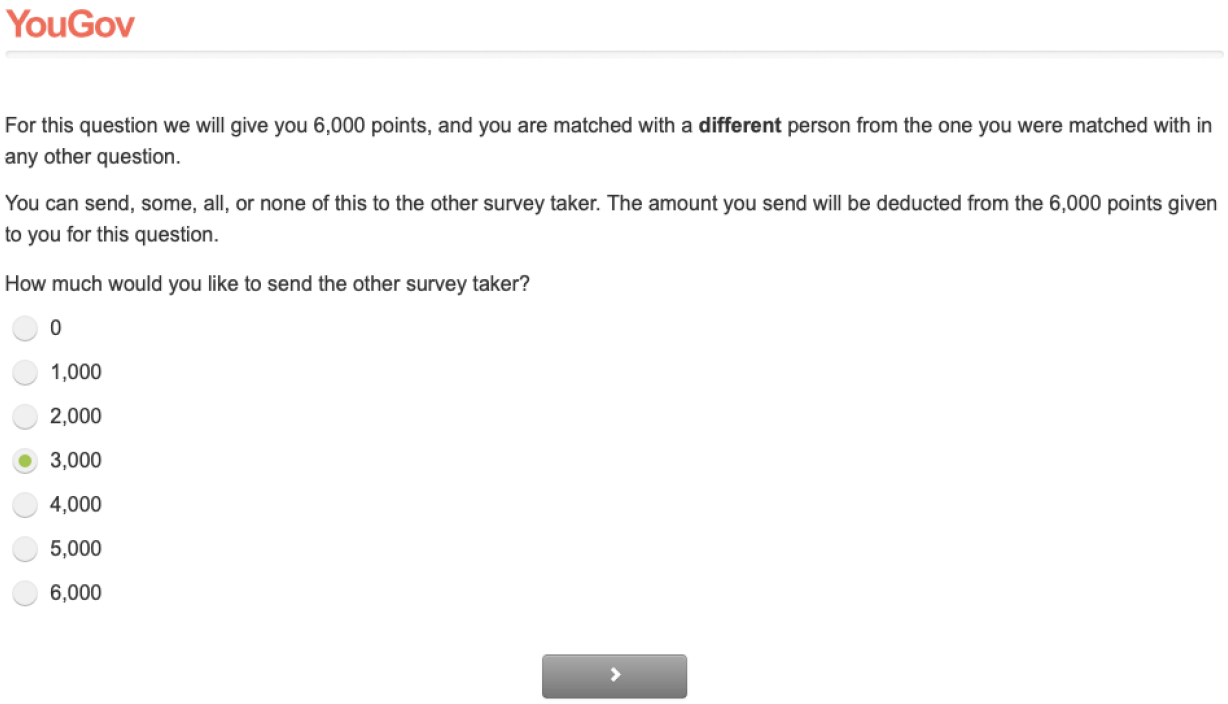


Figure B.10: Incentivized Elicitation of Trust (U.S. and U.K.)

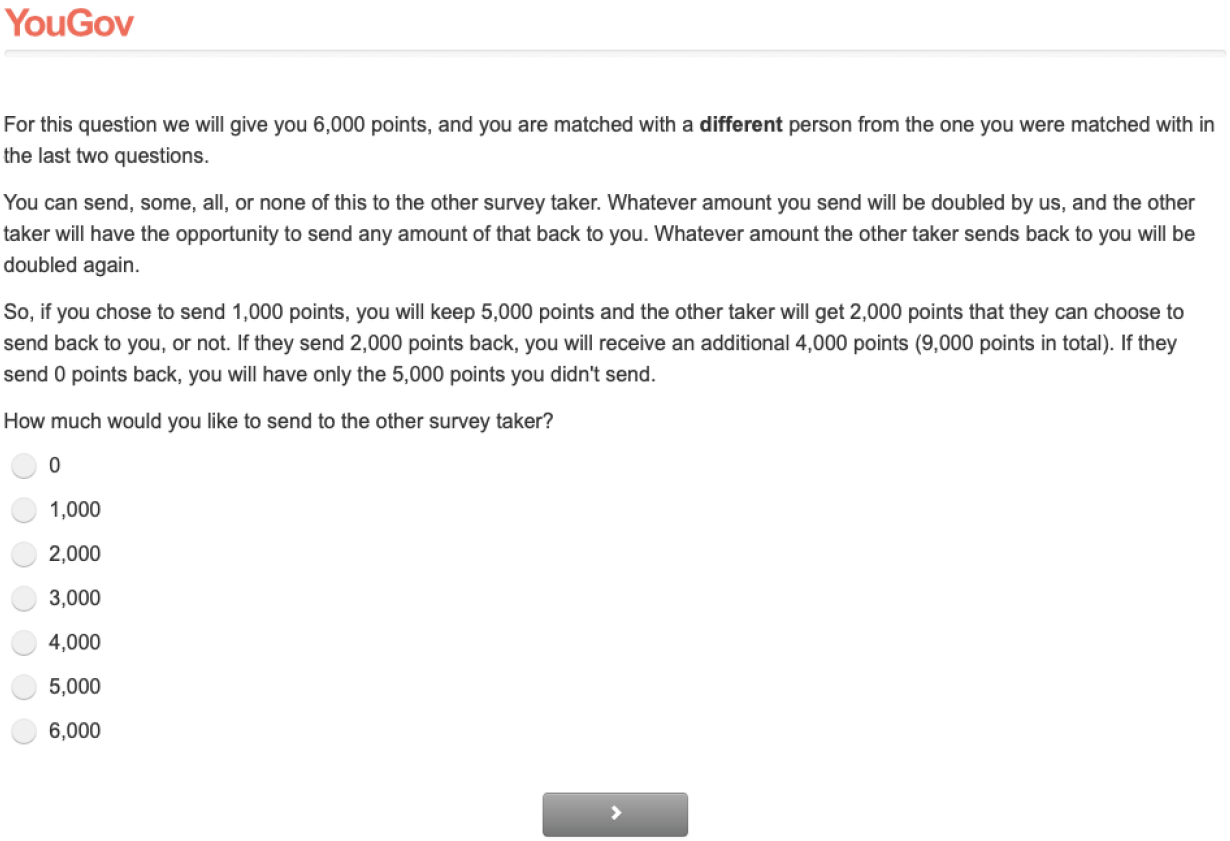


Figure B.11: Incentivized Elicitation of Reciprocity (U.S. and U.K.)

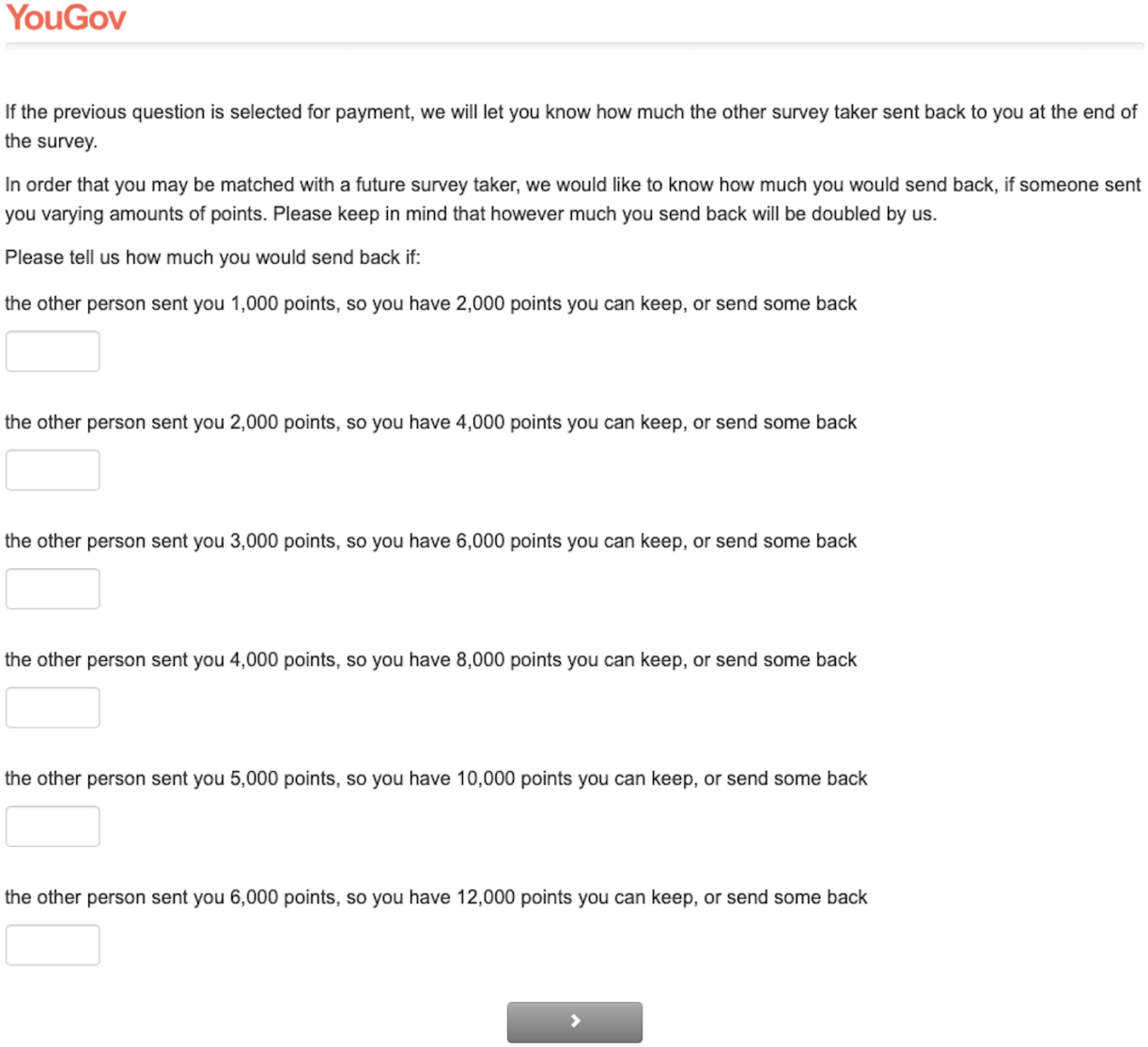


Figure B.12: Incentivized Elicitation of Punishment (U.S. and U.K.)

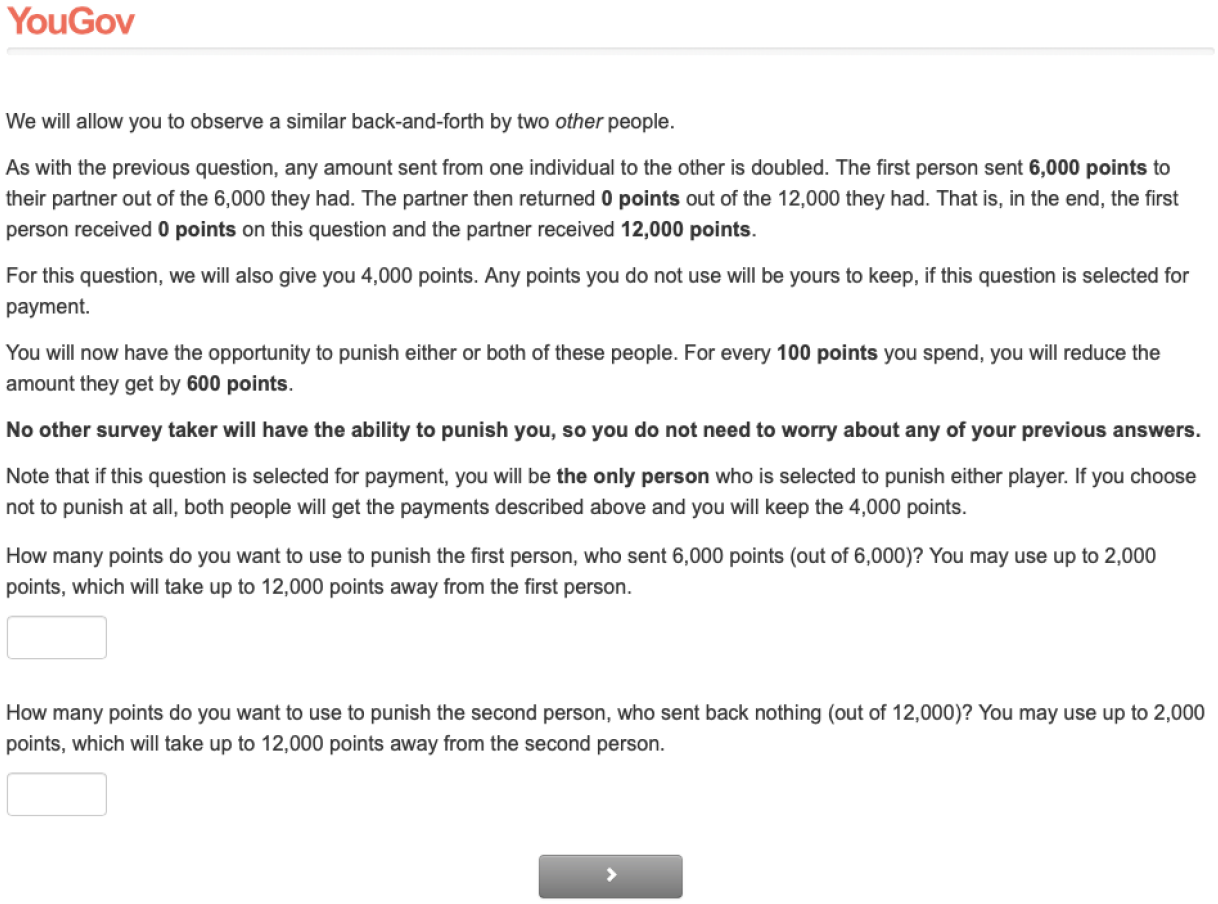


Figure B.13: Qualitative Self-Assessment of Risk Tolerance (Caltech, UBC, and MTurk)

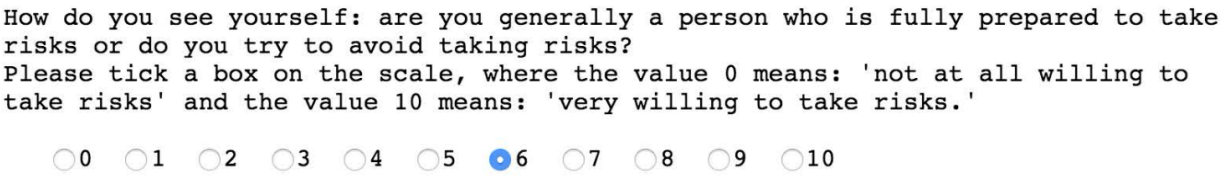


Figure B.14: Qualitative Self-Assessment of Impatience (Caltech)

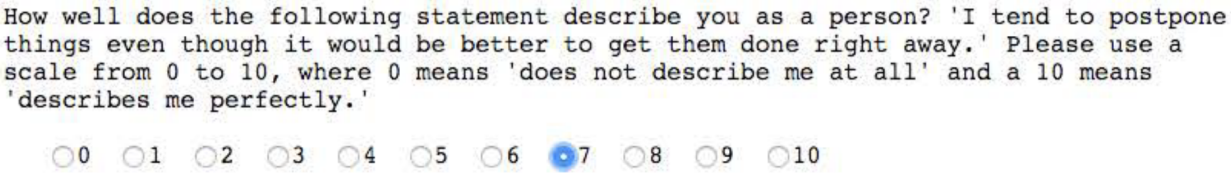


Figure B.15: Qualitative Self-Assessment of Altruism (Caltech, UBC, and MTurk)

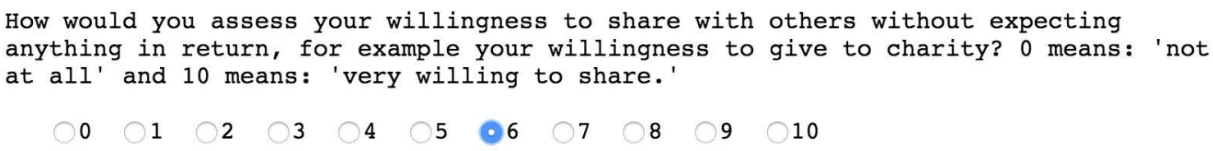


Figure B.16: Incentivized Elicitation of Risk Tolerance (Caltech, UBC, and MTurk)

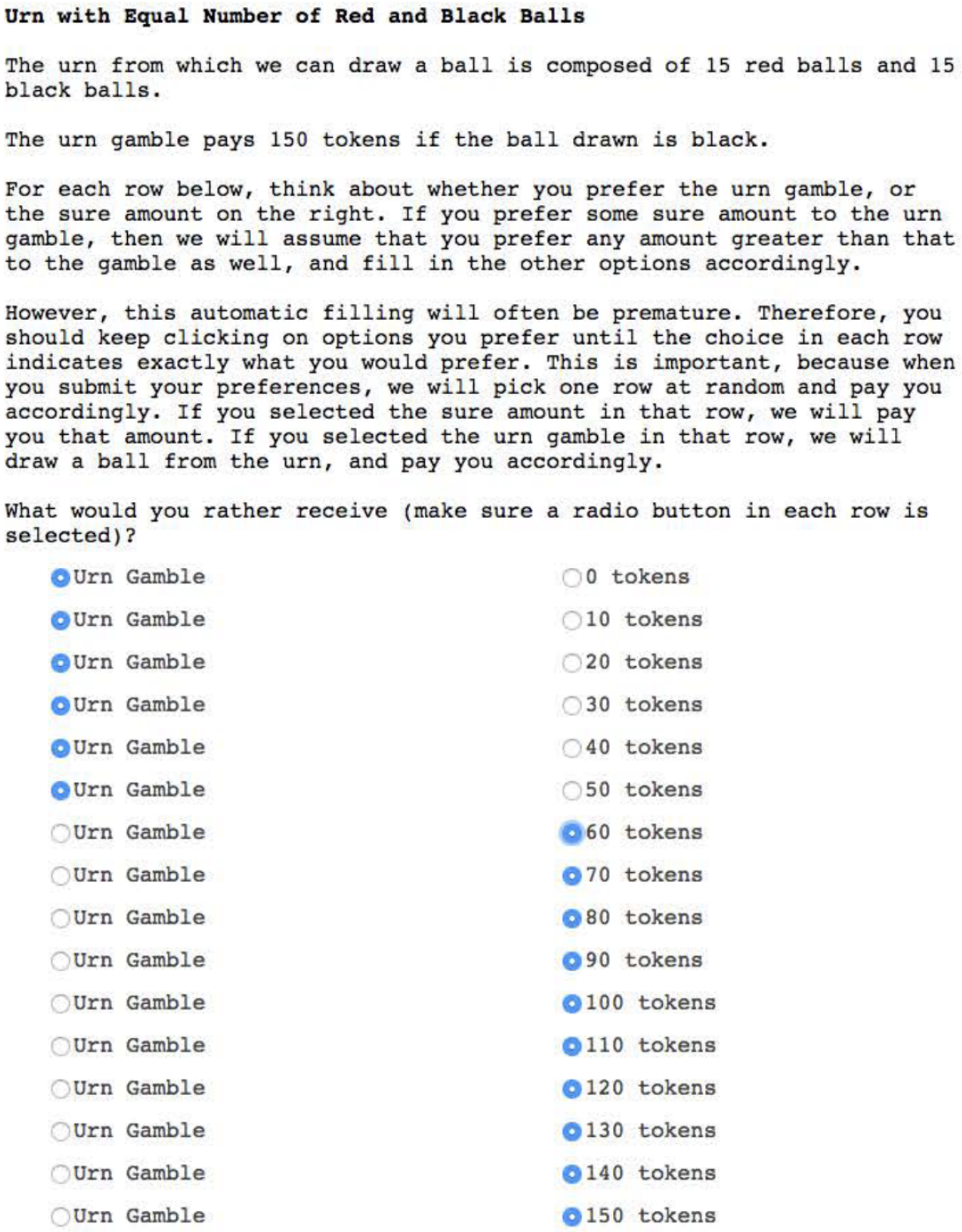


Figure B.17: Incentivized Elicitation of Impatience (Caltech)

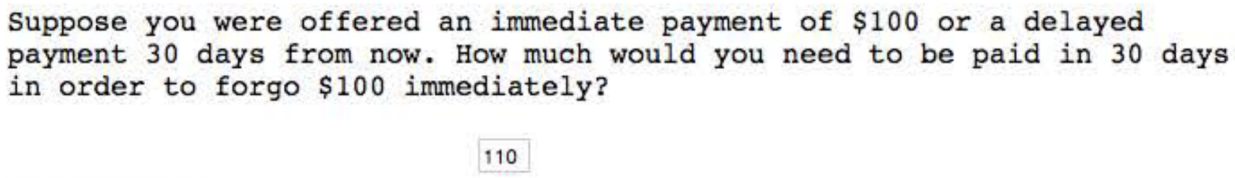


Figure B.18: Incentivized Elicitation of Altruism (Caltech, UBC, and MTurk)

