European Economic Review 127 (2020) 103493

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/euroecorev

Using survey questions to measure preferences: Lessons from an experimental validation in Kenya ☆

Michal Bauer [b], Julie Chytilová [a,*], Edward Miguel [c]

[a] Institute of Economic Studies, Faculty of Social Sciences, Charles University, Opletalova 26, 110 00 Prague, Czech Republic, and CERGE-EI (a joint workplace of Charles University in Prague and the Economics Institute of the Czech Academy of Sciences), Politických vězňů 7, 111 21 Prague, Czech Republic

[b] CERGE-EI and Institute of Economic Studies, Faculty of Social Sciences, Charles University, Czech Republic

[c] University of California, Berkeley, Department of Economics, United States of America

ARTICLE INFO

Article history:
Received 24 June 2019
Revised 25 December 2019
Accepted 4 June 2020
Available online xxx

Keywords:
Preference measurement
Experiment
Survey
Validation

ABSTRACT

Can a short survey instrument reliably measure a range of fundamental economic preferences across diverse settings? We focus on survey questions that systematically predict behavior in incentivized experimental tasks among German university students (Becker et al. 2016) and were implemented among representative samples across the globe (Falk et al. 2018). This paper presents results of an experimental validation conducted among low-income individuals in Nairobi, Kenya. We find that *quantitative* survey measures – hypothetical versions of experimental tasks – of time preference, attitude to risk and altruism are good predictors of choices in incentivized experiments, suggesting these measures are broadly experimentally valid. At the same time, we find that *qualitative* questions – self-assessments – do not correlate with the experimental measures of preferences in the Kenyan sample. Thus, caution is needed before treating self-assessments as proxies of preferences in new contexts.

1. Introduction

Fundamental preferences in the economic domain, such as time discounting and risk preferences, and in the social domain, such as altruism, reciprocity and spitefulness, constitute key elements of individual decision-making. Figuring out ways to accurately measure these preferences among large samples in the field holds considerable promise since doing so may shed light on the sources of vast differences in preferences observed across individuals and societies, and their role in fundamental economic choices and societal trajectories. While measuring preferences using incentivized tasks is generally considered the gold standard,[1] implementing incentivized tasks among large samples outside of the controlled environment

☆ We thank Livia Alfonsi, Ben Enke, Armin Falk, Johannes Haushofer, Karen Macours, Eric Ochieng, Matthias Sutter and Chaning Yang for helpful conversations and advice. Darya Korlyakova, Irene Nginga, Pauline Wangjeri, Debra Opiyo, Joseph Kimani and Innovations for Poverty Action (IPA) field team in Kenya provided excellent research assistance. We also thank Ivana Burianová for administrative assistance. Michal Bauer and Julie Chytilová thank the Czech Science Foundation for funding of the data collection (17-13869S) and for support of further work on the project (20-11091S). The research has been covered by IRB approval obtained by Busara Center for Behavioral Economics (MSU/DRPC/MUERC000011/13).

* Corresponding author.
E-mail addresses: bauer@cerge-ei.cz (M. Bauer), chytilova@fsv.cuni.cz (J. Chytilová), emiguel@berkeley.edu (E. Miguel).

[1] Experimental measures of preferences have been shown to predict a wide range of real-life behavior (e.g., Ashraf, Karlan, & Yin, 2006; Burks, Carpenter, Goette, & Rustichini, 2009; Meier & Sprenger, 2010; Rustagi, Engel, & Kosfeld, 2010; Sutter, Kocher, Glätzle-Rützler, & Trautmann, 2013).

https://doi.org/10.1016/j.euroecorev.2020.103493
0014-2921/

of an experimental laboratory is often infeasible, given that they are relatively expensive and time consuming. Consequently, a potentially attractive alternative is to employ survey questions instead of incentivized experiments, but there has long been widespread concern that non-incentivized self-reported survey measures of preferences may not reliably capture real life choices.

To tackle this important methodological trade-off, Falk et al. (2018) have recently developed an innovative short (7-8 minutes) survey module, designed to measure a wide range of economic preferences. It has been implemented among representative samples of subjects in more than seventy countries (Falk et al., 2018), creating the most comprehensive global data set with comparable measures of preferences, namely, the Global Preference Survey (GPS). Measures of preferences in each domain are constructed as a weighted average based on one objective quantitative item – a hypothetical version of an experimental task – and one subjective qualitative item that measures self-reported willingness to act in a certain way.

To establish the validity of the survey preference measures, Becker et al. (2016) perform a careful experimental validation of the survey questions, and document that survey measures of preferences do predict choices in incentivized decisions. The validation was conducted among students at the University of Bonn, Germany. Given the wide coverage of the existing GPS data set and the convenience of the survey module in terms of implementation,[2] it has the potential to become a widely adopted instrument for (i) studying differences in preferences across societies and their relationships with economic outcomes, (ii) employing preference measures as control variables when identifying causal effects of other factors correlated with preferences, and (iii) as outcome variables in new randomized controlled trials aiming to uncover the effects of various interventions on individual preferences.[3]

This paper adds to these efforts and aims to be useful in three ways. First, we test the experimental validity of the survey questions outside of a sample of university students from a rich country, by focusing on a sample from the other end of the global distribution of income and education. Our experimental subjects are residents of working class neighborhoods (sometimes referred to as "slums") in Nairobi, Kenya, a setting with a different set of institutions and economic constraints. The participants are aged between 20-46, with average income of around USD 3 per day, and 54% are unemployed. Establishing the experimental validity of the measures among this subject pool is important for several reasons. Most of humanity lives in low and middle income countries, outside of Western, Educated, Industrialized, Rich and Democratic societies (Henrich et al. 2010), in which the original GPS validation was conducted. Next the GPS module is particularly suitable to be integrated into large-scale follow-up surveys in randomized control trials, which are routinely implemented by development economists (Banerjee and Duflo 2012), often in Africa, and thus knowledge of whether the survey preference measures predict incentivized behavior among low-income individuals in Kenya is a useful input for scholars considering the adoption of these measures.[4]

Second, comparing the results of analogous validations conducted in Germany and Kenya is methodologically interesting, because measures of economic preferences in GPS are derived from both objective quantitative tasks as well as subjective qualitative questions, based on self-assessments.[5,6] There is a legitimate concern that subjective self-assessments might be understood and interpreted in different ways across countries, which can attenuate their ability to uncover personality traits and complicate cross-country comparisons. For example, the Big Five measures of personality traits, the most widely-used method to measure and classify personality traits in psychology, are based on self-assessments, and recent attempts to validate the Big Five measures have failed to reliably predict the intended personality traits in low- or middle-income countries, in contrast to samples from the wealthy countries for which they were originally developed (Laajaj et al. 2019; Gurven et al. 2013). An advantage of GPS is that, besides self-assessments, it also contains quantitative questions that are arguably less subject to this issue, because they directly define the parameters and nature of the decision and more closely mirror the incentivized experimental task. Thus, we can test whether quantitative questions are relatively more robust predictors of actual incentivized behavior across two diverse settings, as compared to qualitative self-assessments.

Third, we place additional emphasis on the types of preferences that are likely to be especially important in settings with low social capital and a history of inter-group conflict, issues that are particularly pressing in low-income countries (Blattman and Miguel 2010). While pro-social preferences, such as altruism and positive reciprocity, help to establish and maintain cooperative and fair group outcomes even in situations with limited scope for reputation-building (Bowles 2006; Fehr and Fischbacher 2003), anti-social preferences (such as spitefulness and aggressive competitiveness) can contribute to

[2] There is a laudable public good element in the GPS project. The global data set, as well as the survey instrument – and its 116 versions for 70 countries and 78 different languages – are readily available to researchers at https://www.briq-institute.org/global-preferences/home. Our validation experiment benefited greatly from this transparency, as we build on the Swahili translation of the survey module for Kenya.

[3] To date, the GPS measures have been used to explore global variations of preferences and their relationships with country-level and individual-level characteristics (Falk et al., 2018), deep historical origins of variation of preferences (Becker, Enke, and Falk 2018) and the relationships between economic development and gender differences in preferences (Falk & Hermle, 2018).

[4] Indeed, this experimental validation itself took place as a part of a larger project that aims to estimate the long-term effects on individual preferences of a randomized public health intervention (a school-based deworming program) which took place in Western Kenya starting in the late 1990s (Baird et al. 2016; Miguel and Kremer 2004). We used lessons from the current validation exercise in the design of a preference survey module that is integrated into the most recent round of follow-up data collection (Kenyan Life Panel Survey, KLPS, round 4).

[5] An example of a qualitative question from GPS would be "Please tell me, in general, how willing or unwilling you are to take risks, using a scale from 0 to 10", or "How willing are you to give to a charity without expecting anything in return?"

[6] For recent discussions about the importance of replications and other methods that aim to foster credibility of research findings see, for example, Maniadis, Tufano, and List (2014) andChristensen and Miguel (2018).

the deterioration of co-operation (Falk, Fehr, &Fischbacher, 2005; Herrmann, Thoni, &Gachter, 2008).[7] Furthermore, ethnic biases in social preferences – in-group favoritism and out-group hostility – create fertile ground for violent inter-group conflict. While the GPS focuses on measuring preferences relevant for explaining positive aspects of human social behavior, such as generalized altruism and reciprocity, we also assess the experimental validity of survey questions designed to measure the dark side of human social behavior. Specifically, we test the validity of questions designed to uncover anti-social preferences, such as spite, and distinguish between generalized, in-group, and out-group preferences, along both prosocial and anti-social dimensions.

2. Experimental design

The sample in our study are 123 subjects from the Kibera neighborhood in Nairobi, Kenya. The participants come from a low-income environment, are between 20 and 46 years of age, more than half are unemployed, half are women and, on average, they have two children (Table A1). The experiments were implemented in a state-of-the-art experimental economics laboratory in the Busara Center for Behavioral Economics (Haushofer et al. 2014).

Subjects were invited to the lab twice, for visits one week apart, where the time gap was introduced in order to minimize spillovers between the survey and experimental measures. During one visit, they made choices in a set of incentivized experiments, while during the other, they answered non-incentivized survey questions. The order of experiments/survey was randomized on an individual level. We elicited measures of the following types of preferences: (i) time discounting, (ii) risk preference, (iii) ambiguity aversion, (iv) altruism (generalized, in-group, and out-group), (v) anti-social behavior (generalized, in-group, and out-group), and (vi) positive reciprocity.

The experimental choices involved high stakes, in order to capture decision situations with substantial financial consequences for the subjects. Specifically, each subject received a show-up fee (KSh 450 for the survey part and KSh 250 for the experimental part, where 100 KSh was roughly equal to 1 USD during the study period) and a payoff determined by one randomly selected choice made in the experimental part. The average payoff from experiments was KSh 820, i.e., the equivalent of approximately 2.5 days' typical earnings. Each type of preference was elicited using one experimental task. The full experimental protocol is available in the Online Appendix.

For time discounting, subjects made 25 binary choices between an immediate payment or a larger payment with a three-month delay, which was increased by a fixed amount in each subsequent binary choice, using a multiple price list. Similarly, when eliciting risk preference, subjects made 21 binary choices between a lottery that yielded a positive amount or zero with equal probability, and a safe payment option that increased in each subsequent binary choice. Ambiguity aversion was measured by a binary choice between two bags – one with a known and one with an unknown composition of differently colored balls – with the payoff determined by drawing a ball of a specific color.

In the experiments focusing on the social domain, altruism was measured by the choice of how much of an endowment the participant decided to donate to a charity. One choice measured donations to a charity which helps people in Kenya (generalized altruism), the second choice elicited donations to a charity which helps people from the participant's ancestral home area (in-group altruism), and the third elicited donations to a charity which helps people in Kenya outside of the subject's own ancestral home area (out-group altruism). Anti-social behavior was measured using a binary choice in which subjects could decide to reduce the payoff of another person by sacrificing a part of their own payoff. Again, we implemented three versions, using the same wording as above to indicate generalized, in-group, and out-group versions of the task. Finally, positive reciprocity was measured by the amount of money given to a person who had been kind to the participant. This person was an anonymous participant in a different, earlier experiment in the lab who decided to leave a gift (a bag of sugar, which is a popular gift item in the setting we study) for a future visitor of the lab (i.e., decision-maker of our study), instead of keeping all the sugar for him or herself. As an alternative measure of reciprocity, we used the difference in the amount donated to this (kind) person and to another (unkind) person who had decided not to give any sugar.

In the survey part, we elicited one objective quantitative measure and one or two subjective qualitative measures for each type of preference.[8] The quantitative questions presented a hypothetical scenario that mimicked the experimental task. For time and risk preferences, instead of asking the full set of questions as in the experiment, we used the "staircase" or "unfolding brackets" procedure, in which each participant answers a sub-set of five binary choices chosen based on their answer to the previous question. The qualitative questions measure self-reported willingness to act in a certain way on a 0-10 scale. Specifically, respondents rate their own willingness to give up something that is beneficial today in order to benefit more in the future (time discounting), to take risks (risk preference), to give to a charity and to share with others (altruism), to cause trouble for other people and to do harm to other people (anti-social preferences), and to return a favor (reciprocity).

Note that the experimental validation in Nairobi is comparable to, but not strictly identical to, the preference measure validation conducted in Bonn (Becker et al. 2016). Some of the experimental tasks had to be simplified, reflecting the differences in average schooling between the Kenyan and German subject pools. We also slightly adjusted the wording in some of

[7] Anti-social preferences – malevolent willingness to harm others at a cost to self – have been shown to be relatively widespread in numerous settings in both high and low income settings (Abbink and Sadrieh 2009; Fehr, Hoff, and Kshetramade 2008; Herrmann, Thoni, and Gachter 2008; Prediger, Vollan, and Herrmann 2014; Bauer, Cahlíková, Chytilová, et al. 2018a; Bauer, Cahlíková, Celik-Katreniak, et al. 2018b).

[8] The only exception is ambiguity aversion, for which there is only one quantitative survey measure.

the GPS survey questions, based on feedback from piloting and focus-group discussions, in order to improve comprehension in the Kenyan context. In terms of procedure and data analysis, we use a similar approach as Becker et al. (2016). Please see the Online Appendix for details of each experimental task, questions used and the comparison of the Kenyan and German validation exercises.

To start, we observe that the elicited preference measures have several desirable properties (see Online Appendix Table A1 for summary statistics). First, there is substantial variation in all our measures of preferences, both survey and experimental, alleviating concerns that a failure to identify relationships between variables of interest could be mechanically driven by a lack of variation. Second, behavior in the experiments is largely comparable to previous studies. For example, in the generalized version of the dictator game (altruism measure), we observe that subjects allocate around 20% of their endowment to charity. We also find that subjects are significantly more willing to give to a charity that helps their own ethnic group, as compared to a charity that helps out-group members. Similarly, slightly fewer than 20% reduce another person's income at a cost to themselves, which is comparable to the prevalence of anti-social behavior in other settings (Abbink and Sadrieh 2009; Prediger, Vollan, and Herrman 2014), and subjects are significantly more destructive towards out-group members.

3. Results

We begin by describing the predictive power of objective quantitative survey measures. For each survey item, Table 1 displays an OLS coefficient from a regression of the standardized experimental measure on the standardized survey item (column 1) and the Spearman correlation between the survey item and a respective experimental incentivized preference measure (column 2). Below each coefficient and correlation, we report the 95% confidence interval.

We find that the quantitative survey measures of time preference, attitude to risk, generalized altruism, altruism towards one's own ethnic group, and altruism towards out-group members are strongly positively correlated with experimental measures, and the observed relationships are statistically significant. The quantitative survey measure of ambiguity aversion and all three measures of anti-social behavior correlate weakly with the experimental measure: the correlations for all are relatively small in magnitude and none is significant at traditional levels.

Specifically, in terms of magnitudes, the correlations are 0.40 for time discounting, 0.25 for risk preference, 0.29 for positive reciprocity, 0.41 for generalized altruism, 0.36 for in-group altruism and 0.38 for out-group altruism. These correlations are slightly lower than, though comparable to the correlations generated in the validation of the same set of survey preference measures in Germany (Becker et al. 2016), reported for comparison in column 3, in which the corresponding correlations were found to be 0.55 (time discounting), 0.34 (risk taking), 0.35 (positive reciprocity) and 0.39 (generalized altruism). Each of the correlations from the German study reported in Table 1 falls within the respective 95% confidence interval of our estimate in Kenya, except of the measure of time discounting for which the correlation in Germany is 0.55 and the upper bound of our estimate is 0.54. We speculate that the somewhat smaller correlations in Kenya may potentially reflect greater measurement error in the elicitation of preferences among a subject pool with lower average schooling levels.

The observed patterns are robust to controlling for the level of understanding, based on direct cross-check questions, and violations of monotonicity (in tasks eliciting time and risk preferences, which use multiple price lists), an indirect proxy of understanding. The correlations are also similar for different orderings of the survey and experimental tasks (namely, whether they were elicited during the first or second week), and robust to controlling for a set of basic individual characteristics (i.e., age, gender, being unemployed, and the number of children); the results of these robustness checks are presented in Online Appendix Table A2.

Further, we consider a concern that is inherent in this type of experimental validation, namely, that subjects may remember their choices from the previous week and choose the same options in the second week in order to appear consistent over time. To address this, we included an independent task to measure a subject's memory. Specifically, in the first week, the participants were shown a set of ten letters on a screen for twenty seconds and were incentivized to remember those letters for a short period. In the second week, they were asked to recall these ten letters, again in an incentive-compatible way. We show that the correlations observed between experimental and survey measures of preferences are not driven by subjects with more accurate recall (those remembering above the median number of letters), with the exception of the time preference measure (Table A3).

Next, we explore the predictive power of the subjective survey self-assessments. In contrast to the objective survey measures, qualitative survey measures are rather poor predictors of the experimental measures of preferences (Table 2). None of the correlations reaches statistical significance at conventional levels when we use the Spearman correlation (column 2). The magnitudes are also small. The estimated coefficients are close to zero and in many cases do not have the expected sign: nine estimated correlations have expected signs, while seven have an opposite sign to that predicted. None of the estimated 16 correlations is larger than 0.15. Specifically, the correlations are 0.06 for time discounting, -0.02 for risk preference, 0.06 and 0.14 for two measures of positive reciprocity, 0.07 for generalized altruism, -0.02 for in-group altruism and -0.09 for out-group altruism. For comparison, the German validation (Becker et al. 2016) found the correlations to be -0.41 (time discounting)[9], 0.35 (risk taking), 0.30 (positive reciprocity), and 0.23 and 0.38 (two measures of generalized altruism). We

[9] Note that the negative sign is in line with the intuition since for the experimental measure and the quantitative survey measure of time discounting higher values indicate less patience, while higher values for the qualitative survey measure indicate more patience.

Table 1
Correlations between quantitative survey measures and experimental measures.

Preference		Quantitative survey item	Kenya: Kibera residents		Germany: Bonn students	
			OLS Coefficient	Correlation	Correlation	Measures
			(1)	(2)	(3)	(4)
Time		Staircase measure: 5 interdependent choices between an early and delayed amount of money	0.33***[0.16; 0.50]	0.40***[0.24; 0.54]	0.55***	comparable
Risk		Staircase measure: 5 interdependent choices between a lottery and varying safe options	0.21**[0.03; 0.38]	0.25***[0.07; 0.41]	0.34***	comparable
Ambiguity aversion		Hypothetical choice between a bag with known and unknown number of balls of different color	0.13[-0.05; 0.31]	0.13[-0.05; 0.30]	n.a.	
Reciprocity		Hypothetical choice of the amount of a gift given to a stranger who provided help	0.12[-0.06; 0.30]	0.29***[0.12; 0.45]	0.35***	exp. different; survey comparable
Reciprocity (diff)		Hypothetical choice of the amount of a gift given to a stranger who provided help	0.06[-0.12; 0.24]	0.19**[0.02; 0.36]	n.a.	
Altruism	generalized	Hypothetical choice of the amount donated to a charity (out of Ksh3200)	0.41***[0.25; 0.58]	0.41***[0.26; 0.55]	0.39***	comparable
	in-group	Hypothetical choice of the amount donated to a charity that helps people in ancestral home area (out of Ksh3200)	0.33***[0.16; 0.50]	0.36***[0.20; 0.51]	n.a.	
	out-group	Hypothetical choice of the amount donated to a charity that helps people in other parts of Kenya than ancestral home area (out of Ksh3200)	0.40***[0.23; 0.56]	0.38***[0.22; 0.52]	n.a.	
Anti-social behavior	generalized	Hypothetical decision between (3200, 3200) or (3150, 1600) for self and for another person	0.05[-0.13; 0.23]	0.05[-0.13; 0.22]	n.a.	
	in-group	Hypothetical decision between (3200, 3200) or (3150, 1600) for self and for a person from ancestral home area	0.07[-0.12; 0.26]	0.07[-0.11; 0.25]	n.a.	
	out-group	Hypothetical decision between (3200, 3200) or (3150, 1600) for self and for a person from other parts of Kenya than ancestral home area	0.14[-0.04; 0.32]	0.14[-0.04; 0.31]	n.a.	

Notes: Column 1 is an OLS coefficient from a regression of the standardized experimental measure on the standardized survey item. Column 2 displays Spearman correlations between the survey item and the respective experimental measure (one for each preference type, except for reciprocity, where we use two experimental measures). ***, **, and *denote significance at the 1-, 5-, and 10-percent level, respectively. Below each OLS coefficient and Spearman correlation, the table reports 95% confidence interval in the square brackets. Column 3 displays the correlation between experimental and quantitative survey measures from the validation study of Becker et al. (2016) among university students in Germany. Column 4 indicates to what extent measures from our study in Kenya and measures from the German study are comparable.

also find low predictive power of qualitative survey measures when using OLS regressions (column 1), with the exception of measures of positive reciprocity and out-group altruism, for which we find positive coefficients (0.21 and 0.18, respectively), significant at the 5% level.

Thus, the overall pattern differs from the patterns observed in the German validation exercise, where quantitative survey measures as well as subjective self-assessments reliably predict behavior in experimental tasks (column 3 of Tables 1 and 2): all estimated correlations in that study are statistical significant, have the expected sign, and the magnitude is on average 0.41 for quantitative and 0.33 for qualitative survey measures, ranging between 0.23 and 0.55. While we find comparable and statistically significant correlations for the quantitative measures, for the qualitative self-assessments the correlations in Kenya are on average approximately one fifth the magnitude of those reported in the German study.

Since our sample size is smaller than the German validation (123 vs. 409), we next address a concern that the difference in findings about the lower predictive power of qualitative items is due to a lack of statistical power. First, we performed power analysis for OLS coefficients. Note that the minimum detectable effect is the same for all our measures because they are standardized. With our sample size we are powered to detect coefficient of the magnitude 0.25 and larger, for alpha=0.05 and beta=0.80. Thus, we are powered to detect medium-sized (but not small) correlations. We perceive this as a meaningful size, given that we are interested in correlations between different (experimental and survey) measures designed to uncover the same underlying preferences. Also note that the lack of statistical significance of the relationship

Table 2
Correlations between qualitative survey measures and experimental measures.

Preference		Qualitative survey item	Kenya: Kibera residents OLS Coefficient	Kenya: Kibera residents Correlation	Germany: Bonn students Correlation	Measures
			(2)	(1)	(3)	(4)
Time		Willingness to give up something that is beneficial today in order to benefit more in the future	0.04[-0.14; 0.22]	0.06[-0.12; 0.23]	-0.41***	comparable
Risk		Willingness to take risks	0.01[-0.17; 0.19]	-0.02[-0.20; 0.16]	0.35***	comparable
Reciprocity		Willingness to return a favor	0.11[-0.07; 0.29]	0.06[-0.12; 0.23]	0.30***	exp different; survey comparable
Reciprocity (diff)		Willingness to return a favor	0.21**[0.03; 0.38]	0.14[-0.04; 0.31]		
Altruism	generalized, measure 1	Willingness to give to a charity	0.03[-0.15; 0.21]	0.07[-0.11; 0.24]	0.38***	comparable
	generalized, measure 2	Willingness to share with others	-0.06[-0.23; 0.12]	-0.02[-0.20; 0.16]	0.23***	comparable
	in-group, measure 1	Willingness to give to a charity that helps people in ancestral home area	-0.03[-0.21; 0.15]	-0.09[-0.26; 0.09]	n.a.	
	in-group, measure 2	Willingness to share with others from ancestral home area	-0.05[-0.23; 0.13]	-0.05[-0.22; 0.13]	n.a.	
	out-group, measure 1	Willingness to give to a charity that helps people in other parts of Kenya than ancestral home area	0.18**[0.01; 0.36]	0.12[-0.06; 0.29]	n.a.	
	out-group, measure 2	Willingness to share with people from other parts of Kenya than ancestral home area	0.12[-0.06; 0.30]	0.13[-0.04; 0.30]	n.a.	
Anti-social behavior	generalized, measure 1	Willingness to cause troubles to other people	-0.1[-0.28; 0.08]	-0.05[-0.22; 0.13]	n.a.	
	generalized, measure 2	Willingness to make harm to other people	0.01[-0.17; 0.19]	0.05[-0.13; 0.22]	n.a.	
	in-group, measure 1	Willingness to cause troubles to people in ancestral home area	-0.02[-0.20; 0.17]	-0.003[-0.19; 0.18]	n.a.	
	in-group, measure 2	Willingness to make harm to people in ancestral home area	0.11[-0.08; 0.30]	0.15[-0.04; 0.32]	n.a.	
	out-group, measure 1	Willingness to cause troubles to people from other parts of Kenya than ancestral home area	-0.01[-0.19; 0.17]	0.02[-0.16; 0.19]	n.a.	
	out-group, measure 2	Willingness to make harm to people from other parts of Kenya than ancestral home area	0.01[-0.17; 0.19]	0.03[-0.15; 0.21]	n.a.	

Notes: Column 1 displays OLS coefficients in a regression of the standardized experimental measure on the standardized module items. Column 2 displays Spearman correlations between the survey item and the respective experimental measure (one for each preference type, except for reciprocity, where we use two experimental measures).***, **, and *denote significance at the 1-, 5-, and 10-percent level, respectively. Below each OLS coefficient and Spearman correlation, the table reports 95% confidence interval in the square brackets. Column 3 displays the correlation between experimental and qualitative survey measure from the validation study of Becker et al. (2016) among university students in Germany. Column 4 indicates to what extent measures from our study in Kenya and measures from the German study are comparable.

between the qualitative measures with experimental measures is primarily due to small magnitude point estimates, rather than due to large standard errors, as discussed above. Second, the confidence intervals are sufficiently narrow to document that estimated correlations in Kenya are smaller than those in the German study. Specifically, in Table 2 we show that the point estimate of each of the correlations for qualitative survey items from the German study is outside of the respective 95% confidence interval of our estimate in Kenya, except of one of our two measures of positive reciprocity, for which the correlation in Germany is 0.30 and the upper bound of our estimate is 0.31. Finally, it is also noteworthy that adding the qualitative survey measures into the regression that correlates a quantitative survey measure with the corresponding experimental choice adds little explanatory power, as indicated by a comparison of R-squared values in Table A5. Based on these patterns, we believe it is unlikely that the lack of statistically significant relationships between qualitative survey items and behavior in experiments in Kenya is due to the somewhat smaller sample size.

Below, we discuss potential explanations for why our findings about the low predictive power of qualitative questions are different from Becker et al. (2016). First, we consider differences in experimental design. Stakes are different in the Kenyan validation as compared to the German validation and a natural concern might be that survey responses only predict decisions with relatively low stakes.[10] The observation that it is not the case that survey questions *per se* would fail to

[10] In the validation in Nairobi, the average payoff from the experiments was 820 Kenyan shillings (approximately USD 8.2 at the time of the experiment), which is an equivalent of approximately 2.5 day's typical earnings in the area of the study. For comparison, in Bonn, the average experimental payoff was 54 Euro (approximately USD 83 at the time of the experiment), but it is not straightforward to assess how this amount compares to typical earnings since the subject pool are university students.

predict incentivized behavior in the Kenyan setting, but rather that a particular type of survey questions (self-assessments) do not predict behavior in that setting, does not favor this interpretation.

Although some of our experimental measures, in particular those for altruism and risk aversion, are closely comparable to Becker et al. (2016), elicitation of preferences in other domains differs in non-negligible ways. This was motivated by our effort to increase simplicity and variation in experimental choices. For example, we elicit reciprocity by measuring the allocation to a person who was kind to the participant, by giving a gift, instead of eliciting second-mover behavior in the Trust game that requires participant knowledge of multiplication. Also, in the time discounting task we elicit three-month discount rates, while Becker et al. (2016) elicit annual discount rates. Our design decision was informed by a small pilot, in which virtually all subjects opted for the most impatient option. Since a lack of variation in the experimental measure would mechanically lead to a low correlation with the survey measures, we decided to increase variation in choices by shortening the length of the delay of the future payment from one year to three months. We find it reassuring that the main pattern (a strong correlation with quantitative survey measures and a lack of correlation with the qualitative survey measure) holds both for the preference domains for which we use closely comparable measures, as well as for measures in preference domains that differ more from the original validation.

Next, the subject pool and setting is very different, including their education levels. While in Bonn study, all subjects are university students, around 30% of subjects in the Kenyan sample never attended secondary school. This is potentially especially important for self-assessments because the questions are relatively abstract, as compared to hypothetical versions of the experimental decisions. Since the sub-sample of subjects who have attended a university or a college is relatively small, we opt to separately estimate the correlations for subjects with below- vs. above-median years of schooling. We find no improvement in the predictive power of the qualitative survey items among subjects who have above-median schooling levels (Table A4).

Also, a long standing concern about using self-assessments to measure personality characteristics across cultures is that they might be understood and interpreted in different ways across settings with different languages and populations with different real life experiences. Finally, low experimental validity of personal self-assessments could, in principle, originate in social desirability biases. Certain personality traits, such as ability to delay consumptions, accept risk or willingness to share might be perceived as socially desirable and may introduce systematic biases in responses. Arguably, general personal self-assessments are more prone to such biases, as compared to specific choices with well-defined parameters and it might be the case that a tendency to misreport according to what subjects perceive as socially desirable differs across setting.

Overall, with just two validation studies at hand it is difficult to pin down one single factor that explains the differences in experimental validity of self-assessments in the German and Kenyan contexts. Thus, our paper highlights the need for more validation studies in additional setting to make progress on these open questions.

4. Concluding remarks

An experimental validation of survey preference measures among residents of a working class Nairobi neighborhood reveals several noteworthy patterns. First, we show that *quantitative survey measures* of time preference, attitude to risk and altruism are good predictors of choices in incentivized experiments. This finding reinforces the findings from a similar validation exercise performed among university students in Bonn (Becker et al. 2016), and thus, together, the two studies document the experimental validity of these measures at opposite ends of the global income and education distribution. Second, this study tested the experimental validity of survey preference measures in a new domain, *anti-social preferences*, which is arguably most prone to social desirability biases. We document that survey measures of anti-social preferences only weakly predict incentivized behavior, which strengthens the case for investing resources into gathering incentivized measures in this domain. Third, we find that the *subjective qualitative questions* on preferences do not meaningfully correlate with the experimental measures in the Kenyan sample, in contrast to the German sample. Thus, caution is needed before interpreting these measures as proxies of preferences in all contexts.

What lessons about measuring preferences using survey questions can we draw from the available evidence? First, our results should boost confidence in the ability of *objective quantitative* GPS survey measures of preferences, based on hypothetical tasks, to predict high-stakes incentivized behavior in experiments designed to measure a range of preferences across economically and culturally diverse settings. Second, qualitative survey questions have been shown to do a good job of predicting behavior in incentivized experiments in rich (mostly German) settings (Dohmen et al. 2011; Becker et al. 2016) and a range of real-life behaviors (Barasinska, Schaefer, and Stephan 2012; Bauernschuster et al. 2014; Bonin et al. 2007; Fouarge, Kriechel, and Dohmen 2014; Jaeger et al. 2010; Dohmen et al. 2011).[11] In light of our findings it might be tempting to conjecture that self-assessments are generally unreliable in low-income settings, in contrast to high-income settings.

[11] In contrast to rich country settings, validation studies conducted in low income countries are still rare and typically focus on measures of a single preference type, specifically on risk preference. Following up on Dohmen et al. (2011), who conducted a validation experiment in Germany, positive correlations between survey and experimental measures of risk preference were documented in rural Thailand (Hardeweg, Menkhoff, & Waibel 2013) and among Chinese students (Ding, Hartog, and Sun 2010). A recent cross-cultural study on risk-taking from 30 countries (Vieider et al. 2015) documents that qualitative survey measures are positively correlated with choices in incentivized experiments in a majority of cases, but the magnitude and statistical significance of the correlation varies substantively across countries, which also suggests that the experimental validity of qualitative survey measures may be context specific.

However, since we do not know which factor (participant education levels, exposure to abstract concepts, social desirability biases, culturally-specific ways of interpreting self-assessments, etc.), or which combination of factors, explains the lower experimental validity of self-assessments in the Kenyan context, generalizing from a single study to all low-income environments would still be premature. Rather, our paper highlights the need for more validation studies, ideally a series of comparable validation exercises in a diverse set of contexts across the globe, in order to better understand the characteristics of individuals or societies for which the qualitative self-assessments are informative. Future research may also need to determine whether alternative formulations of qualitative questions can be more robust predictors of preferences than current self-assessments.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2020.103493.

References

Abbink, Klaus, Sadrieh, Abdolkarim, 2009. The Pleasure of Being Nasty. Econ. Lett.105 (3) 306–308.

Ashraf, N., Karlan, D., Yin, W., 2006. Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines. Quart. J. Econ.121 (2) 635–672.

Baird, Sarah, Hicks, Joan Hamory, Kremer, Michael, Miguel, Edward, 2016. Worms at Work: Long-Run Impacts of a Child Health Investment. Quart. J. Econ.131 (4) 1637–1680.

Banerjee, Abhijit V., Duflo, Esther, 2012. Poor Economics. Public Affairs.

Barasinska, N., Schaefer, D., Stephan, A., 2012. Individual Risk Attitudes and the Composition of Nancial Portfolios: Evidence from German Household Portfolios. Quart. Rev. Econ. Finance 52 (1), 1–14.

Bauer, Michal, Cahlíková, Jana, Celik-Katreniak, Dagmara, Chytilová, Julie, Cingl, Lubomír, Želinský, Tomáš, 2018b. "Antisocial Behavior in Groups. CEPR Discussion Paper 13315.

Bauer, Michal, Cahlíková, Jana, Chytilová, Julie, Želinský, Tomáš, 2018a. Social Contagion of Ethnic Hostility. In: Proceedings of the National Academy of Sciences, 115, pp. 4881–4886.

Bauernschuster, S., Falck, O., Heblich, S., Suedekum, J., Lameli, A., 2014. Why Are Educated and Risk-Loving Persons More Mobile across Regions? J. Econ. Behav. Org.98 56–69.

Becker, Anke, Dohmen, Thomas, Huffman, David, Falk, Armin, Sunde, Uwe, 2016. "The Preference Survey Module: A Validated Instrument for Measuring Time, Risk, and Social Preferences. IZA Discussion Paper 9674.

Becker, Anke, Enke, Benjamin, Falk, Armin, 2018. "Ancient Origins of the Global Variation in Economic Preferences. Working Paper.

Blattman, C., Miguel, E., 2010. Civil War. J. Econ. Lit.48 (1) 3–57.

Bonin, H., Dohmen, Thomas, Falk, Armin, Huffman, David, Sunde, Uwe, 2007. Cross-Sectional Earnings Risk and Occupational Sorting: The Role of Risk Attitudes. Labour Econ.14 (6) 926–937.

Bowles, Samuel, 2006. Microeconomics: Behavior, Institutions, and Evolution. Princeton University Press Princeton, PrincetonNJ.

Burks, S.V, Carpenter, J.P, Goette, L., Rustichini, A., 2009. Cognitive Skills Affect Economic Preferences, Strategic Behavior, and Job Attachment. In: Proceedings of the National Academy of Sciences, 106, p. 7745.

Christensen, Garret, Miguel, Edward, 2018. Transparency, Reproducibility, and the Credibility of Economics Research. J. Econ. Lit.56 (3) 920–980.

Ding, X., Hartog, J., Sun, Y., 2010. Can We Measure Risk Attitudes in a Survey? IZA Discussion Paper 4807.

Dohmen, Thomas, Falk, Armin, Huffman, David, Sunde, Uwe, Schupp, J., Wagner, G., 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences.". J. Eur. Econ. Assoc.9 (3) 522–550.

Falk, A., Fehr, E., Fischbacher, U., 2005. Driving Forces behind Informal Sanctions. Econometrica 73 (6), 2017–2030.

Falk, Armin, Becker, Anke, Dohmen, Thomas, Enke, Benjamin, Huffman, David, Sunde, Uwe, 2018. Global Evidence on Economic Preferences. Quart. J. Econ.133 (4) 1645–1692.

Falk, Armin, Hermle, Johannes, 2018. Relationship of Gender Differences in Preferences to Economic Development and Gender Equality. Science 362 (6412).

Fehr, E., Fischbacher, U., 2003. The Nature of Human Altruism: Proximate Patterns and Evolutionary Origins. Nature 425, 785–791.

Fehr, E., Hoff, K., Kshetramade, M., 2008. Spite and Development. Am. Econ. Rev.98 (2) 494–499.

Fouarge, D., Kriechel, B., Dohmen, Thomas, 2014. Occupational Sorting of School Graduates: The Role of Economic Preferences. J. Econ. Behav. Org.106 335–351.

Gurven, Michael, Christopher von, Rueden, Maxim, Massenkoff, Hillard, Kaplan, Marino, Lero Vie, 2013. How Universal Is the Big Five? Testing the Five-Factor Model of Personality Variation Among Forager–Farmers in the Bolivian Amazo. J. Person. Soc. Psychol.104 (2) 354–370.

Hardeweg, B., Menkhoff, L., Waibel, H., 2013. Experimentally- Validated Survey Evidence on Individual Risk Attitudes in Rural Thailand. Econ. Dev. Cult. Change 61 (4), 859–888.

Haushofer, Johannes, MarieCollins, Giovanna DeGiusti, Joseph MuiruriNjoroge, AmosOdero, CynthiaOnyago, JamesVancel, ChaningJang, KuruvillaManeeshV, and ConorHughes. 2014. "A Methodology for Laboratory Experiments in Developing Countries : Examples from the Busara Center."Working Paper.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The Weirdest People in the World. Behav. Brain Sci.33 (2–3) 61–83.

Herrmann, B., Thoni, C., Gachter, S., 2008. Antisocial Punishment across Societies. Science 319 (5868), 1362.

Jaeger, D., Dohmen, Thomas, Falk, Armin, Huffman, David, Sunde, Uwe, Bonin, H., 2010. Direct Evidence on Risk Attitudes and Migration. Rev. Econ. Stat.92 (3) 684–689.

Laajaj, Rachid, Macours, Karen, Hernandez, Daniel Alejandro Pinzon, Arias, Omar, Gosling, Samuel, Potter, Jeff, Rubio-Codina, Marta, Vakis, Renos, 2019. Challenges to Capture the Big Five Personality Traits in Non-WEIRD Population. Sci. Adv.5 (7).

Maniadis, Zacharias, Tufano, Fabio, List, John A, 2014. One Swallow Doesn' t Make a Summer : New Evidence on Anchoring Effects. Am. Econ. Rev.104 (1) 277–290.

Meier, S., Sprenger, C., 2010. Present-Biased Preferences and Credit Card Borrowing. Am. Econ. J. Appl. Econ.2 (1) 193–210.

Miguel, Edward, Kremer, Michael, 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Econometrica 72 (1), 159–217.

Prediger, Sebastian, Vollan, Bjorn, Herrman, Benedikt, 2014. Resource Scarcity and Antisocial Behavior. J. Public Econ.119 1–9.

Rustagi, Devesh, Engel, Stefanie, Kosfeld, Michael, 2010. Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management. Science 330 (6006), 961–965.

Sutter, Matthias, Kocher, Martin G., Glätzle-Rützler, Daniela, Trautmann, Stefan T., 2013. Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior. Am. Econ. Rev.103 (1) 510–531.

Vieider, F., Lefebre, M., Bouchouicha, R., Chmura, T., Hakimov, R., Krawczyk, M., Martinsson, P., 2015. Common Components of Risk and Uncertainty Attitudes across Contexts and Domains: Evidence from 30 Countries. J. Eur. Econ. Assoc.13 (1) 421–452.