Discussion Paper Series – CRC TR 224

Discussion Paper No. 397
Project A 05

Accounting for Individual-Specific Reliability of Self-Assessed Measures of Economic Preferences and Personality Traits

Thomas Dohmen [1]
Tomáš Jagelka [2]

March 2023

[1] University of Bonn, IZA and University of Maastricht, Email: tdohmen@uni-bonn.de
[2] University of Bonn, ECONtribute, IZA and CREST, Email: tjagelka@uni-bonn.de

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Accounting for Individual-Specific Reliability of Self-Assessed Measures of Economic Preferences and Personality Traits

Authors:

Thomas DOHMEN[‡] and Tomáš JAGELKA[*]

March 7, 2023

We would like to give special thanks to Daniel Navarro-Martinez and Hannah Schildberg-Hörisch for their valu-able comments. We also thank participants at the YEP Seminar of the ECONtribute Cluster of Excellence, at the seminar of the Düsseldorf Institute for Competition Economics, at the ECONtribute Young Economist Workshop in Applied Economics, at the 2022 World Economic Science Association Conference, and at the 2022 conference of the Slovak Economic Association. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866. Dohmen also acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (Project A05).

[‡]University of Bonn, IZA Institute of Labor Economics, and University of Maastricht. Email: tdohmen@uni-bonn.de.

[*]ECONtribute Cluster of Excellence, IZA, Institute for Applied Microeconomics at the University of Bonn, and CREST. Email: tjagelka@uni-bonn.de.

Abstract

Measures based on self-assessments, which are increasingly important in empirical economic research, are plagued by measurement error. This paper presents the first attempt at measuring both *revealed* and *self-reported* reliability of individuals' answers on self-reports of latent characteristics. We show that measurement error on self-reports relevant to economists is heterogeneous across individuals and can be reasonably approximated by a distribution with two unobserved types. We propose a straightforward survey question which allows to distinguish individuals who give highly reliable answers from those who do not, using cross-sectional data. We demonstrate that it predicts revealed individual reliability over and above all measured characterises, survey conditions, and experimental treatments. We show how our simple self-reported reliability measure can be used to cost-effectively reduce attenuation bias in estimates of cognitive and non-cognitive determinants of high school GPA, college graduation, unemployment, and life satisfaction. Without requiring panel data, the achieved correction is similar to some of the most effective reduced-form theory-based approaches in the existing literature. Finally, we clarify the role of effort and self-knowledge in generating measurement error and propose a simple model which rationalizes our findings.

1 Introduction

Self-assessments are an important source of information in social science research. They are increasingly used by empirical economists to obtain measures of individual attributes, such as personality traits or preferences, which influence economic decision-making and important life outcomes.[1] It is now widely accepted across fields in social science that self-assessments, which measure constructs that are typically not directly observable, are plagued by measurement error.[2] This problem can be especially severe for qualitative self-assessments that are difficult to incentivize properly.[3] Measurement error might be individual-specific due to differences in attention and self-knowledge. Random noise reduces the precision of measurements and obscures true relationships between measured constructs and outcomes. Suitable methods for addressing measurement error often require panel data or are cumbersome to implement.

We propose a straightforward survey question that allows to distinguish individuals who give highly reliable answers from those who do not. It asks respondents to directly self-report the extent to which their responses to the survey describe them accurately. This simple question can easily be added to any survey or experiment. Importantly, it allows researchers to account for measurement error using cross-sectional data only.

We implement our novel survey instrument in a two-wave study with 651 respondents from four major English speaking countries, in order to assess the properties and quality of our reliability measure. Panel data enable us to construct a quantitative person-specific measure of revealed reliability, which provides a benchmark to which self-reported reliability can be compared. Panel data also allow us to calculate test-retest correlations for measures of preferences, skills, and well-being which are most relevant to economists.[4] The test-retest correlations are invariant to

[1] See, e.g., Epstein (1979); Barrick and Mount (1991); Salgado (1997); Nyhus and Pons (2005); Heckman, Stixrud, and Urzua (2006); Roberts et al. (2007); Mischel et al. (2011); Heckman et al. (2010); Cobb-Clark and Tan (2011); Dohmen et al. (2011); Heckman, Pinto, and Savelyev (2013); Heckman, Humphries, and Kautz (2014); Caliendo, Cobb-Clark, and Uhlendorff (2015); Falk et al. (2018); Todd and Zhang (2020); Falk, Neuber, and Strack (2021); Fiala et al. (2022); List et al. (2022); Stango and Zinman (2020). See Heckman, Jagelka, and Kautz (2021) for a recent summary of the literature.

[2] See, e.g. Lord and Novick (1968); Epstein (1979); Schmidt and Hunter (1996); Bound, Brown, and Mathiowetz (2001); Barrick, Mount, and Judge (2001); Heckman, Stixrud, and Urzua (2006); Gillen, Snowberg, and Yariv (2019); Beauchamp, Cesarini, and Johannesson (2017); Jagelka (2020); Stango and Zinman (2020).

[3] Jagelka (2020) finds that qualitative measures of personality have 3 to 10 times lower signal-to-noise ratios relative to measures of economic preferences obtained from incentivized choice tasks.

[4] Some of these are novel while others were thus far scattered across various papers in economics and psychology. We not only centralize the calculated test-retest correlations in one place but also provide measures which are directly comparable. We are not aware of existing attempts to measure test-retest correlations relevant to economists based

the time lapsed between the first and second experimental wave which ranges from 2 to 11 weeks across the surveyed individuals. This provides support for the assumption that the underlying constructs of interest are largely stable within our timeframe. Therefore the calculated test-retest correlations are informative about the size of measurement error inherent in a given measure, such that a higher test-retest correlation implies that a survey instrument measures the latent construct of interest more precisely.

We show that heterogeneity in answer precision at the individual level can reasonably be approximated by a classification of two types, a reliable one and an unreliable one. We demonstrate that test-retest correlations are higher for reliable types than for unreliable types. Crucially, self-reported reliability is highly predictive of the actual reliability of respondents' answers. In fact, self-reported reliability is the single best predictor of revealed reliability and captures measurement error due to both lack of effort and imperfect self-knowledge. Individuals who self-report to provide highly reliable answers have 40% less noise content relative to those who do not. This is comparable to a reduction in measurement error achieved by increasing the number of indicators per construct from 1 to 5.

We demonstrate the effectiveness of our proposed method in eliminating attenuation bias in estimates between outcomes (college graduation, unemployment, high school GPA, life satisfaction) and their latent cognitive and non-cognitive determinants, and compare it to that of existing approaches. Our approach largely outperforms standard screening criteria. The reduction in attenuation bias is similar to that achieved by theory-based alternatives proposed by Gillen, Snowberg, and Yariv (2019) – henceforth GSY – and Falk, Neuber, and Strack (2021) – henceforth FSN –, both of which require panel data and non-trivial computation. We encourage researchers to implement it in their surveys as a cost effective means to account for individual-specific measurement error.

Having constructed a quantitative person-specific measure of revealed reliability also allows us to explore the determinants of measurement error. Our findings suggest that the reliability of an individual's answers depends on how *willing* and *able* the individual is to answer the tasks that he is presented with. *Willingness* to provide highly reliable answers determines the amount of effort the individual decides to put into the survey. Some individuals are motivated by external incentives and choose to put in effort in order to hedge their bets and make sure they get paid, even though we explicitly informed them that their survey payoff is only contingent on survey completion and not the answers they provide. They are characterized by high risk

on a single dataset.

aversion and neuroticism and low trust. Others simply want to provide informative answers to the researchers due to their prosociality. These individuals have high agreeableness and positive reciprocity and low negative reciprocity. The *ability* to provide highly reliable answers requires good self-knowledge, a precise understanding of the tasks at hand, and the ability to accurately select the answer which the individual considers most appropriate. It is facilitated by high conscientiousness, openness to experience, and cognitive ability.

We clarify the role of effort and imperfect self-knowledge in explaining differences in individual reliability. Individuals need to spend at least a minimal amount of time (and effort) in order to correctly process the survey items. The precision of answers is initially sharply increasing in survey time and then flattens out. Responses of individuals above the effort threshold contain 70% less noise relative to those who are below.[5] This is comparable to a reduction in measurement error achieved by increasing the number of indicators per construct from 1 to 12. We estimate the threshold to be at approximately the 5th percentile in survey times, which corresponds to a little over 2 seconds per survey item.[6] Spending additional time is largely ineffectual in terms of increased answer precision.

No matter the amount of effort exerted, an important source of measurement error remains. We attribute this residual imprecision to a fundamental lack of self-knowledge, the extent of which is heterogeneous across individuals. It is the self-knowledge component of individual reliability which is likely responsible for its trait-like features. A reliability question introduced after the first set of personality questions on our survey predicts answer reliability on all types of survey items just as well as a reliability question which we ask at the end of the survey. Both questions are correlated with each other within a survey wave and also across waves, and the magnitude of the correlations is in line with test-retest correlations for single item measures of other latent characteristics such as personality and preferences. These results corroborate and complement Jagelka (2020) who finds that imperfect self-knowledge regarding one's risk and time preferences maps strongly to the conscientiousness personality trait and Enke and Graeber (2021) who find that cognitive uncertainty is correlated across domains of survey expectations, inter-temporal choice, and choice under risk.

We propose a simple model which accounts for our findings. Individuals first choose whether

[5]The effort threshold is important above and beyond standard criteria and attention checks often employed by survey providers to screen out unreliable observations. Our dataset was already "cleaned" by our survey provider Dynata to exclude individuals who obviously "streamlined" the survey (e.g., who always picked the middle option) or who spent too little time (according to Dynata). Our results suggest that their time criterion is too lenient.

[6]This includes time spent reading instructions.

or not to exert sufficient effort needed to correctly comprehend and evaluate the experimental tasks. If they are below the effort threshold, their answers will be largely uninformative. If they are above the threshold, the reliability of their answers will only be constrained by their level of self-knowledge, which is responsible for the remaining noise after accounting for effort.[7]

The rest of the paper is organized as follows: Section 2 situates our contribution within the existing literature, Section 3 describes the data, Section 4 documents the extent of the measurement error problem and its correlates in our dataset, Section 5 demonstrates the effectiveness of self-reported reliability in identifying identifying individuals who give highly reliable answers and in eliminating attenuation bias, Section 6 examines the sources of measurement error, Section 7 develops a simple model which rationalizes our empirical results, and Section 8 discusses the implications of our research for practitioners and concludes.

2 Background

Any attempt to measure a latent construct relies on observing an individual's performance on a task (see Heckman, Jagelka, and Kautz, 2021).[8] Much of the existing literature has focused on lack of effort, or inattentiveness, as the main source of error in measuring latent constructs in surveys and experiments. This idea corresponds to the "insufficient effort responding" (IER) concept of Huang et al. (2012).[9] However, the relationship between survey time (effort) and measurement error is controversial in the literature but important to psychologists and, more recently, also to economists (e.g., Ratcliff and Rouder, 1998; Smith, Krajbich, and Webb, 2019; Alós-Ferrer, Fehr, and Netzer, 2021). One side argues that answers are more precise when individuals take more time to answer the assigned tasks (e.g., Wise and Kong, 2005; Meade and Craig, 2012). Underlying this argument is the assumption that more time is associated with higher effort, which is assumed to improve response quality. The other side (e.g., Shadlen and Kiani, 2013; Alós-Ferrer et al., 2016; Fudenberg, Strack, and Strzalecki, 2018; Liu and Netzer, 2021) points out that individuals spend more time on a question when they are close to indifference between two available options. This implies that, at least on questions with only two answer categories, more time is associated with lower answer precision, provided that respondents exert sufficient effort to answer the question.[10] This argument is consistent with the idea that cognitive certainty and

[7] We assume that the individuals are benevolent towards the researcher and do not actively seek to deceive him.

[8] A task could be a choice in an experiment, a test, a real effort task, or a self-assessment.

[9] Huang et al. (2012) define IER as "a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses."

[10] It is not entirely clear whether "the other side" predicts that higher response times suggest more measurement error when choice options are not binary. For example, if an answer has to be given on a 11-point Likert scale,

self-knowledge are inversely related to response time.

Our results support and clarify the first hypothesis: in the context of commonly used qualitative self-assessments of personal characteristics such as those which we employ in our experiment, individuals need to spend at least a minimal amount of time (and effort) in order to correctly process the survey items. This implies that the reliability of answers is increasing in survey time but only up to a certain threshold. The idea of a non-linearity in the effect of effort on response accuracy is explored also in the psychology literature, see Meade and Craig (2012).

More recently, the idea of imperfect self-knowledge has been explored as a separate and additional source of answer imprecision. It has been incorporated into theoretical models of decision-making. In the words of Loomes and Sugden (1995): "the stochastic element derives from the inherent variability or imprecision of the individual's preferences, whereby the individual does not always know exactly what he or she prefers." The concept of imperfect self-knowledge - or cognitive uncertainty - is supported by recent empirical evidence and found to impact individuals' choices both on qualitative survey questions and on incentivized choice tasks (e.g., Jagelka, 2020; Enke and Graeber, 2021; Falk, Neuber, and Strack, 2021).

Finally, even if individuals put in sufficient effort and know themselves perfectly, their responses to latent trait elicitation tasks might still be noisy because their perception of the tasks and their attributes might be changing.[11] Evidence in neuroeconomics finds that decision values are formed from neural activity in the ventromedial prefrontal cortex of the human brain. The neural activity itself is stochastic (for a summary of the evidence, see Fehr and Rangel, 2011).

Once a researcher has identified the presence of measurement error in his dataset, there are three ways to proceed. The researcher can either exclude the unreliable individuals, take steps to improve the precision of available responses in a reduced-form framework, or apply a structural model to take into account the randomness in observed choices.

The first strand of the literature relies on the identification of the problematic individuals. Existing methods focus almost exclusively on screening out inattentive respondents. Techniques include asking attention check questions, tracking time spent answering the survey, and constructing various inattentiveness indices based on response patterns (e.g., streamlining a survey,

respondents might not be able to decide between 7 and 8, as they are (almost) indifferent between these two categories, and therefore deliberate, but either answer would be more precise than quickly but thoughtlessly choosing an arbitrary answer category.

[11] Responses of individuals might also be unreliable due to willful misreporting. This would be a source of bias rather than noise.

characterized by the respondent systematically selecting the same response across survey tasks), answer consistency across related tasks (equivalent questions should receive the same response), or outliers. Many of these approaches are well summarized in Meade and Craig (2012) and more recently in Stantcheva (2022). In addition, see Read, Wolters, and Berinsky (2022) for a nice discussion of the nuances of detecting low quality responses based on survey time patterns, whether it be too fast or too slow. Finally, see FNS for a clever, theory-based attempt to identify individuals with low self-knowledge.

An intriguing possibility involves directly asking individuals about the reliability of their responses, instead of trying to infer it from response patterns. Our review of the literature revealed very few such attempts: Wise and Kong (2005), Meade and Craig (2012), and Alesina, Miano, and Stantcheva (2023). All of these focused on interrogating individuals on their effort or attention. The instruments used do not elicit the overall reliability of responses, which we show requires much more than effort provision (see Section 2).

The second strand of the literature takes measurement error as given and tries to account for it using various methodological procedures. The simplest consists of asking multiple tasks related to a given construct and averaging over the responses (e.g., Soto and John, 2017; Falk et al., 2022). Somewhat more complex are approaches using the instrumental variables (IV) framework. The recent application of the IV approach by GSY inspired an enthusiastic following (e.g., Chapman et al., 2023; Stango and Zinman, 2020). Their obviously related instrumental variables (ORIV) approach uses panel data on a given measure as instruments for one another. Finally, researchers may use structural models to model and account for various mental processes driving inconsistent choice (e.g., Cunha, Heckman, and Schennach, 2010; Jagelka, 2020; Belzil and Jagelka, 2020).

3 Data

The experiment was conducted online through the Dynata platform between December 2020 and February 2021. It includes two waves of data-collection with an average delay of 5 weeks between the initial survey and the recontact.[12] The target population was English-speaking individuals between 18 and 25 years old from Australia, Canada, the United Kingdom, and the United States. Our main analysis includes 1,400 individuals, 651 of whom completed both survey waves.[13] Descriptive statistics for the sample are provided in Table A.1 of the Appendix.

[12]Subjects had a window of time to participate in both the first and the second wave of the experiment. The time lapse between an individual's responses to the first and second wave ranges between 2 and 11 weeks.

[13]The survey provider replaced at no cost individuals who they considered as providing very low quality responses. We were able to gain access to the excluded observations (119 in the first wave and 49 in the second wave) and use

Invitations for participation in the second survey wave were sent to all participants who successfully completed the first wave. Selection into our wave 2 based on observed characteristics is limited. Individuals who participated in both survey waves are slightly older, more likely to be employed, and based in the UK. An additional analysis investigating selection on ability, personality, and preferences reveals that wave 2 participation is predicted by conscientiousness including all of its facets, by the emotional stability facet of neuroticism, by higher risk aversion, and by lower present bias (see Table A.3 of the Appendix).

Individuals answered questions which are used to measure well-being, personality, and economic preferences in well-known and often-studied datasets. These include a 60-item BFI-2 questionnaire (Soto and John, 2017) designed to measure personality; 6 additional questions which are used to measure personality in the German SOEP dataset and which do not overlap with the BFI-2 questionnaire; qualitative measures of economic preferences used in the Global Preferences Survey (Falk et al., 2018); the Satisfaction with life scale (SWLS); and an overall life satisfaction question used in the World Values Survey (Bjørnskov, 2010).[14] The survey also contains questions regarding intellectual ability. At the end of the BFI-2 section and also at the end of the whole experiment, each individual was asked to evaluate the reliability of the answers which they provided. The full list of tasks is included in the online appendix.[15]

The experiment includes several treatment conditions which involve altering the magnitude of monetary incentives offered for participation in the experiment and the order in which questions were asked. Dynata standardly offers a small compensation to its survey takers. On top of this, we randomized half of the sample into an *extra incentive treatment* condition. At the end of the welcome screen, all individuals received the text: "We thank you for your participation and careful attention". Survey participants who were randomized into the *extra incentive treatment* in addition had the text: "As a token of our gratitude, on top of the tokens which you usually receive from Dynata, we offer you an additional compensation worth ..." These individuals received the equivalent of an extra 3 Euros expressed in their local currency (British Pounds, Australian Dollars, Canadian Dollars, or American Dollars). The individuals who were randomized into the *extra incentive condition* in the first wave, were further randomized into either receiving the same extra incentives for answering the second survey wave, or into receiving the double of that (i.e. the equivalent of 6 Euros in local currency).

them to compare our method to standard exclusion criteria.

[14] We do not have data on life satisfaction for all individuals as the questions was only added midway through the survey.

[15] Several questions were added in the second survey wave only.

The second treatment randomizes the order in which the survey questions appeared on the screen. Specifically, the modifications involved (1) switching whether the BFI-2 Questionnaire the section on ability appeared before or after the ability section in the survey; and (2) whether in the ability section participants were first asked to provide their high school GPA or to qualitatively evaluate their intellectual ability. This generated a total of four treatments regarding question order. The randomization generated in the first experimental wave was maintained in the second wave also. Therefore while question order was different among individuals, each individual faced the same question order in both survey waves.

The extra incentive treatment has two noteworthy effects significant at 5%. First, it improves individuals' mood at the beginning of the survey in both survey waves. The boost to mood fades over time and becomes insignificant by the end of the survey. Doubling the extra incentive has no additional impact. Second, extra incentives increase the amount of time that a person spends on the survey by almost 2 minutes on average. This corresponds to an approximately 15% and 10% increase respectively given the average survey time on the first and second survey wave.[16] The increase in time taken to complete the survey is confined to the higher end of the time distribution.

Treatments which alter the order in which questions were asked have no statistically significant impact on any of the analyzed traits or measures of well-being. Time spent on the overall survey as well as on its subsections is also unaffected. This suggests that the order in which sections of the survey appear does not impact time spent on them. We thus see no evidence of learning or fatigue within a survey wave.[17]

4 Test-retest Correlations of Commonly Used Measures of Personality and Preferences

In this section, we provide test-retest correlations for measures of preferences, skills, and well-being which are most relevant to economists.[18] We show that test-retest correlations are invari-

[16] While the absolute magnitude of the effect is similar in both survey waves, it looses statistical significance in the second one.

[17] We see no evidence of learning between survey waves either as there are no statistically significant differences between time spent on equivalent sections of wave 1 and wave 2.

[18] Psychologists in the field of psychometrics pioneered methods to quantify the reliability (precision) of elicited measures and hence the extent of the measurement error problem. The test-retest method is a prominent example. It entails eliciting repeated measures of the same construct for a given group of individuals within a short enough time period (generally a couple of weeks) such that the underlying construct of interest can be reasonably assumed

ant to the time lapsed between the first and second experimental wave which varies from 2 to 11 weeks across the surveyed individuals. This provides support for the assumption that the underlying constructs of interest are largely stable within our timeframe. Furthermore, test-retest correlations are invariant across four major English-speaking countries and unaffected by our experimental treatments, which randomly increase the survey payment and modify question order. We document a monotonic pattern between the number of indicators used to measure a given construct and the construct's test-retest correlation. We quantify the observed empirical relationship and use it to benchmark the effectiveness of our novel self-reported reliability measure. We also show how it compares to the theoretical prediction and find that it offers a simple explanation for a number of puzzling previous claims, such as economic preferences having lower test-retest stability or explanatory power than various psychological personality traits.

As can be seen in Table 1, all test-retest correlations are substantially below one. Assuming that the underlying constructs of interest are stable within the studied time period, this confirms the intuition that measurement error is present and important. The assumption of construct stability is supported by the lack of a time trend in measured individual answer reliability with time lapsed between the survey waves, see Table A.8 of the Appendix.[19]

BFI-2 personality traits exhibit the highest test-retest correlations with an average of 0.82 across the 5 traits. The 15 facets exhibit an average correlation of 0.72 and the same is found for personality traits from the SOEP inventory. It is worth noting that while the BFI-2 inventory includes 12 measures per personality trait and 4 per facet, the SOEP only includes 3 measures per trait (4 for openness to experience).

Qualitative economic preference test-retest correlations average 0.56 across the 9 studied preference measures which capture risk preference, time preference (including present bias), and social preferences. The preferences have only one dedicated measure each with the exception of negative reciprocity with respect to oneself which has two measures. Notably, single-item preference measures have similarly sized test-retest correlations as single items of the BFI-2 inventory (see Table A in the Appendix). Test-retest correlations average 0.75 across the two included measures

stable (see, e.g., Cozby and Bates, 2012). A "test-retest" correlation is then taken for each measure of interest across survey waves. Psychologists often rely on test-retest correlations as an indicator of a measure's reliability (e.g., Soto and John, 2017).

[19] As a reminder, we have substantial heterogeneity in recontact times which range between 2 and 11 weeks. Combined with a large number of observations, we should have ample power to detect a time trend, if any were present. Yet recontact time is far from being statistically significant. The p-value on a daily time-trend is almost 0.3 while that on a dummy which splits the sample in half by recontact time is almost 0.4.

Table 1: Test-retest Correlations of Standardly Used Qualitative Behavioral Measures: Country Variation

Group	Instrument	Construct	Test-Retest Correlation	Australia	Canada	UK	USA
Personality	BFI-2 60-item	Extraversion	**0.85**	0.83	0.88	0.82	0.91
		- Sociability	**0.82**	0.80	0.82	0.83	0.83
		- Assertiveness	**0.74**	0.71	0.80	0.73	0.71
		- Energy	**0.68**	0.60	0.70	0.64	0.84
		Conscientiousness	**0.83**	0.81	0.77	0.86	0.85
		- Organization	**0.77**	0.74	0.78	0.78	0.80
		- Productiveness	**0.75**	0.71	0.73	0.77	0.76
		- Responsibility	**0.71**	0.74	0.60	0.74	0.70
		Neuroticism	**0.85**	0.88	0.85	0.85	0.81
		- Anxiety	**0.77**	0.81	0.74	0.76	0.71
		- Depression	**0.79**	0.78	0.79	0.80	0.76
		- Emotional Volatility	**0.75**	0.77	0.75	0.74	0.73
		Agreeableness	**0.77**	0.75	0.73	0.79	0.77
		- Compassion	**0.67**	0.62	0.61	0.72	0.66
		- Respectfulness	**0.69**	0.72	0.63	0.72	0.65
		- Trust	**0.64**	0.60	0.68	0.64	0.66
		Openness to Experience	**0.78**	0.74	0.80	0.77	0.80
		- Curiosity	**0.67**	0.64	0.73	0.67	0.65
		- Aesthetic_Sense	**0.72**	0.72	0.75	0.69	0.68
		- Imagination	**0.69**	0.61	0.73	0.67	0.79
	SOEP	Extraversion	**0.79**	0.74	0.81	0.80	0.82
		Conscientiousness	**0.68**	0.73	0.56	0.69	0.67
		Neuroticism	**0.78**	0.78	0.82	0.76	0.75
		Agreeableness	**0.65**	0.73	0.60	0.64	0.62
		Openness to Experience	**0.68**	0.69	0.57	0.68	0.77
Economic Preference	Global Preference Survey	Risk Preference	**0.71**	0.77	0.74	0.65	0.73
		Patience	**0.42**	0.58	0.26	0.40	0.35
		Present Bias	**0.58**	0.56	0.61	0.59	0.49
		Altruism	**0.57**	0.61	0.50	0.54	0.69
		Trust	**0.60**	0.44	0.65	0.62	0.77
		Positive Reciprocity	**0.53**	0.54	0.45	0.53	0.58
		Neg Reciprocity Self	**0.56**	0.59	0.61	0.51	0.55
		Negative Reciprocity Self2	**0.61**	0.53	0.67	0.61	0.66
		Neg Reciprocity Other	**0.48**	0.48	0.49	0.45	0.55
Well-being	Gallup 1-Item	Life Satisfaction	**0.77**	0.86	0.84	0.79	0.50
	SWLS 5-item	Life Satisfaction	**0.72**	0.67	0.76	0.75	0.66
	Current Mood	Mood at Beginning of Survey	**0.61**	0.58	0.51	0.68	0.52
		Mood at End of Survey	**0.65**	0.58	0.58	0.74	0.55
Cognitive Ability	Qualitative Assessment	Ability Computer	**0.62**	0.58	0.58	0.64	0.64
		Ability Writing	**0.68**	0.77	0.74	0.60	0.65
		Ability Reading	**0.60**	0.59	0.70	0.58	0.54
		Ability Communication	**0.64**	0.61	0.72	0.60	0.70
		Ability Problem-Solving	**0.58**	0.56	0.61	0.56	0.67
		Ability Math	**0.72**	0.69	0.77	0.69	0.76

of life satisfaction. Cognitive ability test-retest correlations average 0.64 across the 6 studied items (computer skills, writing, reading, communication, math, problem-solving).

The obtained test-retest correlations are broadly in line with corresponding test-retest correlations reported in various papers in the economics and psychology (see Figure A.6 of the Appendix). Previous results are based on disparate datasets and do not cover all constructs relevant to economists.

4.a Impact of the Number of Measures on Test-retest correlations

Table 1 reveals a pattern between test-retest correlations and the number of measures included in a construct. The 12-item BFI-2 personality traits are more stable than 3-item facets and SOEP traits which are in turn more stable than 1-item preference and ability questions.[20]

The pattern is intuitive. Each measure of a particular construct can be seen as reflecting the underlying trait of interest plus random noise. Increasing the number of measures helps average out the noise and yields a more accurate indicator.[21] Test-retest correlations increase in line with reduced measurement error in the indicators.

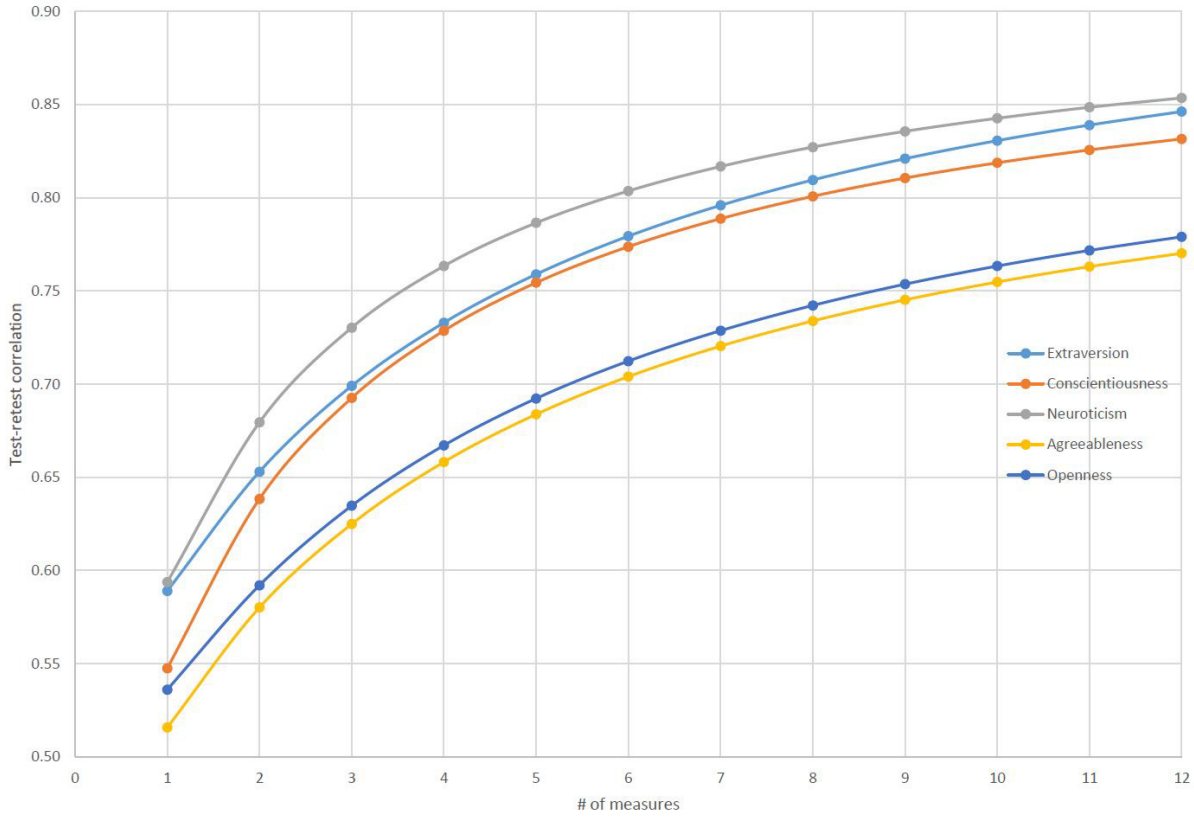
We quantify this relationship using measures from the BFI-2 questionnaire which includes 12 items per personality trait. Figure 1 confirms that we can improve our measurement system simply by using more measures and quantifies the empirical rate of the improvement. It plots test-retest correlations for each of the Big 5 personality traits against the number of items used to measure them. Specifically, it shows the average test-retest correlation of all the possible combinations of the 12 dedicated items which produce 1, 2, ..., 12 measures for each trait.[22] The results are very clear: increasing the number of items increases test-retest correlations monotonically all the way up to 12 which suggests that measurement error is present but decreasing in the number of items used. Across the five personality traits, the average test-retest correlation increases by approximately 50% from 0.56 to 0.82 when using all 12 items instead of only 1 item. Test-retest stability is thus high and likely underestimated in low-dimensional personality tests. The largest gains in precision from adding an extra item occur when few measurement items are used. However, the gains are still positive at 12 items which leaves open the possibility that

[20] In fact, on average test-retest correlations for individual measures from the BFI-2 questionnaire are similar to test-retest correlations for individual measures from the SOEP questionnaire or economic preferences. See Table A of the Appendix.

[21] This insight was already provided by Epstein (1979) who showed that behavior averaged over a number of days demonstrates much higher temporal stability than when it is evaluated as a single instance.

[22] Medians are virtually identical.

Figure 1: Test-retest Correlations of Big 5 Personality Traits as a Function of the Number of Measures Used



test-retest correlations approach 1, indicating stability of the underlying construct.

The decrease in measurement error as the number of measures increases offers a natural point of comparison against which we can benchmark the effectiveness of our proposed method of identifying reliable individuals based on self-reports. I also offers a simple explanation for a number of puzzling previous claims based on comparing latent constructs with unequal dedicated measures. For example qualitative measure of an economic preference often has only one dedicated survey question whereas a given personality trait is typically measured using 3 or more indicators. This may confound results and lead to potentially misleading conclusions such as economic preferences having lower test-retest stability (see, e.g., Schildberg-Hörisch, 2018)[23] or explanatory power (see, e.g., Cobb-Clark et al., 2019) than various psychological personality traits.

Figure A.1 of the Appendix shows the theoretical relationship between test-retest correlations

[23]In the conclusion of her article, Schildberg-Hörisch (2018) indeed calls for economists to "measure a single construct like risk preferences with multiple items ... and to average over those items in order to reduce measurement error."

an the number of measures for an underlying construct. It is calibrated to match the observed single-item noise content of *Big Five* personality traits in our dataset, averaged over the 5 traits. The documented empirical relationship is broadly consistent with the theoretical prediction. The actual increase in test-retest correlations is more gradual, as one would expect when not all available indicators measure exactly the same underlying construct. This is the case of the 12 indicators which we have for each *Big Five* personality trait, each of which in turn has three facets as seen in Table 1.

While we have fewer items per construct to measure economic preferences, life satisfaction, and ability, a similar monotonic increase in test-retest correlations is obtained when we combined related from each of these categories.[24]

4.b Impact of Survey Time on Test-retest correlations

Time taken to answer a survey is a proxy for the effort exerted by a particular respondent. It is thus plausible that it affects measurement error. Despite this variable's intuitive importance, it's impact is at present not well understood. Even the direction of its effect is controversial: As described in more detail in Section 2, some argue (e.g., Wise and Kong, 2005; Meade and Craig, 2012) that more time spent implies a more careful answer and thus less noise while others (e.g., Alós-Ferrer et al., 2016; Fudenberg, Strack, and Strzalecki, 2018; Liu and Netzer, 2021) claim that subjects spend more time when they are uncertain of an answer which would imply more noise in their responses. We shed light on this debate.

Figure 2 plots test-retest correlations for the BFI-2 Big Five personality traits by decile of the distribution of observed time taken to complete the survey.[25] The picture is quite striking. Test-retest correlations initially sharply increase with time taken and then stabilize.

The initial increase in precision with time spent supports the effort hypothesis; the plateauing-off suggests that there is a threshold beyond which additional effort is no longer valuable. In our survey this threshold appears to lie in the vicinity of the 5th percentile of the distribution of survey times.[26,27] Accordingly we say that individuals "rushed" the survey if their response

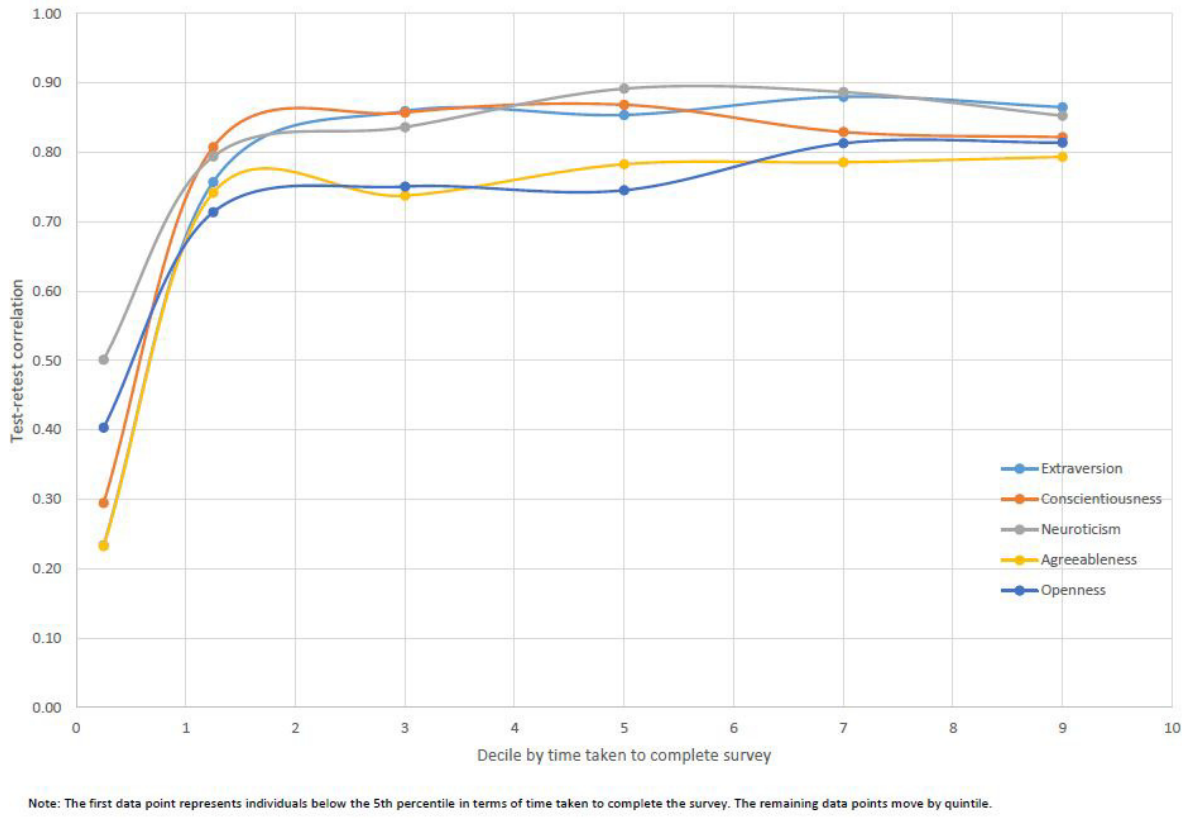
[24] Results are available from the authors upon request.

[25] Figure 2 averages across the 2 survey waves. Results are analogous when looking at time taken for Wave 1 and Wave 2 separately.

[26] This is corroborated by a statistically insignificant effect of minutes spent on answer accuracy once a dummy for being in the fastest 5% is included in the regression (see Table A.10 of the Appendix). The threshold behavior is robust to alternative cutoff points at the 10th, 15th, or 20th percentile of the survey time distribution.

[27] This is also an explanation for why we observe no impact of extra incentives on answer accuracy: while the incentive treatment increases average response times, it does not affect the probability of "rushing" the survey.

Figure 2: BFI-2 Test-Retest Correlations by Time Decile



Note: The first data point represents individuals below the 5th percentile in terms of time taken to complete the survey. The remaining data points move by quintile.

times fall within the fastest 5% of the distribution of survey times on either survey wave.

Rushing the survey has a large impact on response accuracy. It reduces test-retest correlations by 0.43 on average across the Big 5 personality traits. This is larger than the effect of increasing the number of measures per trait from 1 to 12.

Even though some research suggests that individuals who take very long to complete a survey might also give unreliable responses (e.g., Read, Wolters, and Berinsky, 2022), we find no detrimental impact on reliability from individuals with very long response times. In this case standard exclusion criteria applied by survey providers such as Dynata appear sufficient.

4.c Impact of Other Variables on Test-Retest Correlations

Table 1 shows that test-retest correlations are stable across the four examined major countries of the English-speaking world. In Table A.5 of the Appendix we further document that test-retest correlations are stable across the studied period of 2-11 weeks between survey waves and largely invariant to basic demographic variables such as sex, to survey conditions such as participating in the experiment on a PC or on a hand-held device, to providing extra incentives equivalent to

3 Euros (and to doubling them), and to question order. Finally, Table A of the Appendix displays test-retest correlations for the 87 individual items underlying the measures described above.

5 Empirical Results

As seen in the previous section, self-assessments can be noisy. We ask individuals to evaluate the reliability of their answers directly. We do this twice, once after the end of the BFI-2 questionnaire (henceforth "BFI reliability") and the second time at the very end of the survey (henceforth "overall reliability").[28]

The overall self-reported reliability of answers in our survey is high. Out of the 651 individuals who completed both survey waves, 70% reported a reliability of 9 or above on our 11-point Likert scale from 0 to 10 averaged across the survey waves. On the personality section, over half of the individuals indicated that they "strongly agree" that the answers which they provided describe them accurately. This suggests that a classification of respondents into two types based on the reliability of their responses might be fruitful. Accordingly, we consider an individual to be the "reliable type" if he self-reports, on each survey wave, an overall reliability of 9 or above on an 11-point scale from 0 to 10 or, equivalently, a personality section reliability of 5 on a 5-point scale from 1 to 5. The distribution of overall and BFI reliability can be seen in Figures A.2 and A.3 of the Appendix respectively.

For both the reliable and the unreliable types, there is no time trend in measured individual answer reliability with time lapsed between the survey waves.[29] This supports the assumption that there are no differences in actual construct stability of latent characteristics between the two types. Measured constructs are stable for both types between the two survey waves. This means that any difference in our objective measure of measurement error – test-retest correlations – between the two types can be attributed to differences in the amount of measurement error in their responses.

5.a Construct Validity

We now demonstrate that our self-reported reliability question allows us to separate individuals who give reliable responses on our survey from those who do not. In Table 2 we show test-retest

[28] The wording for the BFI reliability question is: "I am someone who is sure that my answers to these questions describe me accurately". The wording for the overall reliability question is: "Please indicate on the scale below how reliable are your answers to this survey."

[29] See Table A.8 of the Appendix.

Table 2: Test-retest Correlations by Self-Reported Reliability

Trait	Sure about BFI	Unsure about BFI	Reliable Survey Answers	Unreliable Survey Answers
BFI-2 Extraversion	0.89	0.76	0.89	0.72
BFI-2 Conscientiousness	0.88	0.73	0.86	0.74
BFI-2 Neuroticism	0.88	0.81	0.87	0.80
BFI-2 Agreeableness	0.80	0.68	0.80	0.65
BFI-2 Openness to Experience	0.82	0.70	0.80	0.70
SOEP Extraversion	0.86	0.67	0.83	0.68
SOEP Conscientiousness	0.79	0.50	0.73	0.52
SOEP Neuroticism	0.83	0.70	0.81	0.69
SOEP Agreeableness	0.74	0.49	0.65	0.56
SOEP Openness to Experience	0.77	0.53	0.74	0.53
GPS Risk	0.77	0.63	0.73	0.65
GPS Time	0.45	0.39	0.43	0.41
GPS Present Bias	0.64	0.47	0.61	0.48
GPS Altruism	0.62	0.50	0.59	0.48
GPS Trust	0.65	0.55	0.61	0.58
GPS Pos Reciprocity	0.56	0.43	0.52	0.44
GPS Neg Reciprocity Self	0.63	0.47	0.59	0.47
GPS Neg Reciprocity Self2	0.66	0.53	0.65	0.49
GPS Neg Reciprocity Other	0.51	0.44	0.51	0.41
Gallup General Life Satisfaction	0.83	0.72	0.79	0.72
SWLS	0.73	0.70	0.73	0.71
Ability Computer	0.63	0.59	0.59	0.62
Ability Writing	0.73	0.60	0.70	0.56
Ability Reading	0.70	0.50	0.63	0.48
Ability Communication	0.72	0.54	0.68	0.54
Ability Problem-Solving	0.66	0.51	0.61	0.50
Ability Math	0.82	0.59	0.78	0.55
Observations	297	354	407	244

Notes: Test-retest correlations highlighted in red are higher for individuals who reported a high reliability of answers (i.e. overall self-reported reliability ≥ 10/11 on both survey waves; self-reported BFI reliability = 5/5 on both survey waves).

correlations for the reliable type and the unreliable type based on their self-reports on each of our survey reliability questions: BFI reliability and overall reliability.

There are several takeaways from this table. First, individuals' *self-reported* answer reliability is related to *measured* answer reliability. When splitting the sample into individuals who report giving highly accurate answers and those who do not, test-retest correlations are universally higher for the individuals who claim to be more accurate. Second, both self-report reliability measures apply across the examined indicators and produce a similar pattern in the data. This

may seem surprising given that the first measure was asked at the end of the personality section and was adapted to the format of the section. The fact that it is predictive of measurement error in all survey sections suggests that reliability may be a personal characteristic, perhaps akin to a trait. The latter hypothesis is supported by the fact that test-retest correlations between the two indicators *within* a survey wave are very similar to test-retest correlations of either indicator *across* survey waves.[30] They are also broadly in line with single item test-retest correlations for other latent constructs measured in our dataset, presented in Table A of the Appendix. These findings corroborate and complement recent research of Jagelka (2020) who finds that imperfect self-knowledge regarding one's risk and time preferences maps strongly to the conscientiousness personality trait and of Enke and Graeber (2021) who find that cognitive uncertainty is correlated across domains of survey expectations, intertemporal choice, and choice under risk. Third, the "BFI question" performs better in discriminating between individuals who give highly reliable answers and those who do not. The average difference in test-retest correlations between the reliable and unreliable group is 0.13 using the BFI measure and 0.11 using the overall reliability measure. This finding is corroborated by additional analyses which we present later on.

As a point of comparison, the gain in test-retest correlations between highly reliable and not highly reliable individuals based on either measure is comparable to the gain in test-retest correlations from increasing the number of items used to measure a personality trait from 1 to 5 or from 3 to 12. As a concrete illustration, take the 3-item SOEP system for measuring personality traits. If a researcher administers the SOEP questionnaire and identifies highly reliable individuals by asking our proposed self-reported reliability question, he will be able to obtain measures for personality traits of a similar accuracy as he would have gotten using the 12-item BFI-2 survey (test-retest correlation around 0.8).

One wave of data collection featuring a single reliability question is sufficient to identify and separate the reliable and the unreliable types. Furthermore, our self-reported reliability questions are able to separate individuals who give reliable responses from those who do not even if we exclude individuals who "rushed" the survey. This confirms that the self-report measure contains complementary important information even once survey time is taken into account. See Table A.9 of the Appendix which replicates Table 2 using alternatively only reliability data from our first survey wave or only individuals who did not rush the survey.

[30]The correlations between the overall precision indicator and the personality section one are 0.44 and 0.43 in wave 1 and wave 2 respectively. Test-retest correlations are 0.44 and 0.46 for the overall and personality precision indicators respectively. The test-retest correlation for the combined measure is 0.53.

Table 3: Revealed vs. Self-Reported Individual Survey Reliability

	(1)	(2)	(3)
VARIABLES	Revealed Individual Reliability		
Self-Reported BFI Reliability	0.11***		0.10***
	(0.01)		(0.01)
Self-Reported Overall Reliability		0.07***	0.04***
		(0.01)	(0.01)
Self-Reported BFI Reliability#Self-Reported Overall Reliability			0.05***
			(0.01)
Constant	0.52***	0.53***	0.5***
	(0.01)	(0.01)	(0.01)
Observations	651	651	651
R-squared	0.16	0.07	0.19

Standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes: Revealed individual reliability represents the test-retest correlation of an individual's responses on all 87 measures included in the survey.

5.a.i Revealed Reliability at the Individual Level

Thus far we considered general test-retest correlations which are used in the literature to determine the reliability of a given measure. We demonstrated that test-retest reliability is correlated with self-reported reliability. In spite of their widespread use, general test-retest correlations fail to capture heterogeneity in answer reliability. We propose a twist to this commonly used concept and construct test-retest correlations for a given person across measures. We thus obtain a quantitative individual-specific measure of *individual revealed reliability* which allows us to explore heterogeneity in reliability at a deeper level.

For each individual, we construct a vector of normalized scores for 87 items contained in our survey which measure personality, preferences, cognitive ability, and well-being. We obtain two such vectors for each individual, one per survey wave. Our measure of *individual revealed reliability* is the correlation between these two vectors across the two survey waves for each individual. Table 3 shows that self-reported individual reliability is strongly related to revealed self-reported individual reliability. Both self-reported measures are relevant and predictive of revealed reliability.

In line with results from the previous sections, the self-reported measure elicited at the end of the BFI section outperforms the one elicited at the end of the survey. This time, evidence in support of the first measure is overwhelming: its share of explained variation is more than twice as large

Table 4: Revealed vs. Self-Reported Individual Survey Reliability: Controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Revealed Individual Reliability							
Self-Reported BFI Reliability	0.11***	0.11***	0.10***	0.11***	0.11***	0.09***	0.08***	0.07***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Experimental Treatments		x						x
Rushed Survey			x					x
Demographics				x				x
Cognitive Ability					x			x
Personality						x		x
Economic Preferences							x	x
Constant	0.52***	0.53***	0.53***	0.56***	0.52***	0.52***	0.52***	0.56***
	(0.01)	(0.02)	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)	(0.03)
Observations	651	651	651	613	651	651	651	613
R-squared	0.16	0.17	0.18	0.19	0.16	0.21	0.24	0.29

Standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes: Experimental Treatments include: extra incentives, order of personality section, order of ability section.
Demographics include: sex, age, country, current profession, highest achieved education, and using a PC/handheld device for this survey.
Rushed Survey includes a dummy variable indicating whether an individual was below the 5th percentile in survey times on either survey wave.
Cognitive Ability includes qualitative questions regarding ability.
Personality includes 60 BFI2 questions.
Economic Preferences include qualitative questions regarding preferences for risk, time (including present bias), reciprocity, altruism, and trust.
Regressions which include demographics exclude marginal categories which have near 0 mass. This excludes 38 individuals.

and the impact of a 1 standard deviation increase in self-reported "BFI" reliability on revealed individual reliability is 60% larger than for self-reported "overall" reliability. When both are included, the coefficient on the BFI measure maintains its magnitude while the coefficient on the overall measure is halved.[31] These results are robust to constructing the individual reliability measure from all 87 available measures, or separately for the 66 personality measures and the 21 non-personality measures contained in our survey (see Tables A.11 and A.12 of the Appendix).

Table 4 shows that self-reported BFI reliability is by far the single best predictor of revealed individual reliability and explains 16% of its cross-sectional variation. Self-reported reliability provides information above and beyond that which is contained in experimental treatments, demographic variables, time spent on the survey, cognitive ability, personality, and economic preferences combined. It is followed at a distant second place by whether or not someone "rushed" the survey. The share of variation in revealed reliability explained by our one self-reported reliability question dwarfs the share explained by demographics and the *Big 5* personality traits

[31] The interaction term is positive suggesting that the two measures are complementary.

and cognitive ability combined. The share of explained variation is equal to that of all economic preferences[32] together with the present bias indicator.

5.b Attenuation Bias with Unobserved Types

Measurement error in explanatory variables leads to attenuation bias in estimates of their impact on outcomes. Estimated coefficients for noisy variables are biased towards zero. Self-reports of latent characteristics tend to be particularly noisy. This can lead to a Type 2 error: falsely failing to reject the null hypothesis of no effect of an explanatory variable on an outcome.

Take a latent explanatory variable X. When it is measured with error, the researcher will not observe X but rather $\widehat{X} = X + \mu$, where μ is a random variable representing measurement error. Let Y be an outcome that depends on X, such that $Y = \alpha + \beta * X + \xi$. The term ξ is the equation error term and represents the influence of unobserved variables. For simplicity we assume that X is uncorrelated with ξ, and μ is mean 0 and uncorrelated with X, Y, and ξ. We thus obtain the classical errors-in-variables model. The measurement error in the explanatory variable enters the equation error term and creates an endogeneity bias:

$$Y = \alpha + \beta * \widehat{X} + (\xi - \beta * \mu) \tag{1}$$

When the researcher regresses outcome Y on the measured noisy explanatory variable $\widehat{X}$, instead of obtaining the true coefficient β, he will obtain a biased $\widehat{\beta}$:

$$\widehat{\beta} = \frac{cov(Y, \widehat{X})}{var(\widehat{X})} = \frac{cov(\alpha + \beta * X + \xi, X + \mu)}{var(X + \mu)} \tag{2}$$

which converges to

$$plim\widehat{\beta} = \beta * \frac{var_x}{var_x + var_\mu} \tag{3}$$

We can see that for $var_\mu > 0$, the absolute value of $\widehat{\beta}$ is lower than the absolute value of β. In the presence of classical measurement error, the impact of attenuation bias on estimated coefficients is unambiguous: it artificially pushes them towards zero.[33]

In line with our empirical results, we outline a model with two unobserved types: a reliable one and an unreliable one. Denote the prevalence of the unreliable type in the sample p_u, with the

[32] As a reminder, these are: risk preference, time preference (including present bias), positive and negative reciprocity, altruism, and trust

[33] While standard errors may increase or decrease, the impact on calculated t-statistics is also unambiguous. Attenuation bias in the presence of classical measurement error biases t-statistics downward and thus lowers the statistical significance of estimates. Furthermore, the OLS estimator of R^2 converges to a limit which is always lower than the true R^2 due to attenuation bias.

prevalence of the unreliable type being $1 - p_u$. Let $var_r(\mu)$ be the response error variance of the reliable type and $var_u(\mu)$ be the response error variance of the unreliable type. Our findings suggest that $var_r(\mu) << var_u(\mu)$. For simplicity, let us assume $var_r(\mu) = 0$. The presence of the unreliable type in a sample introduces attenuation bias into estimates of all relationships between outcomes Y and noisily measured latent explanatory variables X. Estimated coefficients, their associated t-statistics, and the equation R^2 will all be biased towards zero. The bias is increasing in the response error variance $var_u(\mu)$ and in the sample prevalence of the unreliable type p_u.

We showed that our self-reported reliability measure allows us to separate individuals who provide reliable answers from those who do not. It allows us to identify the unreliable type. Under the conditions outlined above, the exclusion of the unreliable type eliminates attenuation bias due to measurement error on latent explanatory variables. The benefit is unbiased estimates of relationships between latent variables and outcomes of interest.[34] While excluding unreliable individuals eliminates measurement error μ on the explanatory variable X, we are still left with the equation error ξ. The unbiased estimates may thus come at the cost of increased standard errors due to reduced sample size.[35] The application of our method will be most advantageous to the applied researcher when the measurement error problem is substantial and sample size is sufficiently large, so that the exclusion of the unreliable type will leave sufficient statistical power. We proceed to demonstrate its effectiveness in an empirical application.

5.c De-Biasing Estimates Using Self-Reported Reliability

We analyze the effectiveness of our proposed method in reducing attenuation bias in estimates of the cognitive and non-cognitive determinants of four important life outcomes: high school GPA, graduation from college, unemployment, and life satisfaction. We follow the implications of the simple model presented in Section 5.b which suggests that we need to identify and exclude the unreliable type in order to de-bias estimates. *We only use self-reported reliability information from the first wave BFI reliability question in order to illustrate the power of our method with minimal data requirements.*

We regress each life outcome on latent traits (cognitive ability, economic preferences, personality) and demographics. We do so first for the full sample, then only for the unreliable types

[34] In practice, survey responses of reliable types also suffer from measurement error, but to a much lesser degree, as can be seen in Table 2. This means that while attenuation bias is alleviated, it will not be entirely eliminated.

[35] This holds for all methods which attempt to exclude individuals who give low quality responses outlined in Section 2 (e.g., FNS). Standard errors also increase when the instrumental variables approach is used (e.g., GSY.

Table 5: Explanatory Power of Cognitive and Non-Cognitive Characteristics for Individuals who Self-report to be Sure/Unsure of their Responses to the BFI Questionnaire

	Graduated from College			**High School GPA**			**Unemployed**			**Life Satisfaction**		
VARIABLES	Pooled	Unsure	Sure	Pooled	Unsure	Sure	Pooled	Unsure	Sure	Pooled	Unsure	Sure
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Demographics	x	x	x	x	x	x	x	x	x	x	x	x
Cognitive Ability	x	x	x	x	x	x	x	x	x	x	x	x
Personality	x	x	x	x	x	x	x	x	x	x	x	x
Economic Preferences	x	x	x	x	x	x	x	x	x	x	x	x
Constant	-1.25***	-1.10***	-1.45***	0.82***	0.71***	0.91***	0.06	0.14	-0.04	0.13	0.02	0.28
	-0.12	-0.18	-0.17	-0.11	-0.2	-0.1	-0.1	-0.14	-0.14	-0.33	-0.51	-0.46
Observations	1423	696	727	1471	698	773	1437	668	769	644	284	360
R-squared	0.19	0.16	0.26	0.02	0.03	0.05	0.05	0.04	0.09	0.33	0.29	0.37

Standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes: Includes data from the full 1st wave of data collection only. Observations originally excluded by Dynata as low quality are added back for the purposes of this analysis.
Reliability is determined solely based on self-reported "BFI Reliability" in the first survey wave.
Individuals are defined as having graduated from college if they report having finished a bachelor's or a master's degree (as opposed to high school or less).
High School GPA is calculated as a percentage.
Individuals are defined as unemployed if they report their current professional status as unemployed (as opposed to being employed or in college).
Life satisfaction is measured using the general question on life satisfaction.
Demographics include: sex, age, and country.
Cognitive Ability includes responses to qualitative questions regarding ability.
Personality includes responses to 60 BFI2 questions.
Economic Preferences include responses to qualitative questions regarding preferences for risk, time (including present bias), reciprocity, altruism, and trust.

(individuals who self-reported to be *unsure of their responses*), and finally for the reliable types (individuals who self-reported to be *sure of their responses*). Following FNS, we focus on R^2 as a measure of the improvement offered by our technique. Table 5 shows that explanatory power increases for each of the 4 studied outcomes, as one would expect when eliminating attenuation bias due to the the presence of unreliable types in the sample. The difference in explanatory power between the reliable types and the unreliable types is large in all cases. The increase in R^2 ranges from 28% for life satisfaction to 125% for unemployment, with an average of 70% across the studied outcomes. A similar improvement in explanatory power is achieved when separately studying the relationship between each outcome and alternatively cognitive ability, personality, or preferences.[36]

We next compare the effectiveness of our self-reported reliability measure to alternative methods for reducing measurement error in responses: averaging over measures elicited at different points in time, applying standard exclusion criteria used by survey providers, excluding individuals who rushed the survey, being above median in the self-knowledge criterion proposed by FNS, and using the ORIV approach employed by GSY. Both the self-knowledge criterion of FNS and the use of instrumental variables to counter measurement error problems, are grounded in theory. The disadvantage of these methods is that they require panel data and can be cumber-

[36]Results are available from the authors upon request.

Table 6: Explanatory Power of Cognitive and Non-Cognitive Characteristics: A Comparison of Techniques Used for Eliminating Unreliable Observations

Outcome		Base (1)	Base with Xs Averaged Over both Survey Waves (2)	Dynata Streamliner Exclusion (3)	Dynata Fast/Slow Exclusion (4)	Dynata Low Quality Observation Exclusion (5)	No Rush (6)	Self-Reported Reliable (7)	Self-Reported Reliable + No Rush (8)	Above Median in Falk et al. (2021) Self-knowledge (9)	Above Median in Falk et al. (2021) Self-knowledge + Self-Reported Reliable (10)
Graduated from College	Observations	657	657	651	616	603	607	353	345	324	216
	R-squared	0.19	0.20	0.19	0.19	0.19	0.19	0.26	0.27	0.25	0.32
High School GPA	Observations	683	683	677	639	625	633	377	369	338	230
	R-squared	0.04	0.03	0.04	0.04	0.04	0.04	0.04	0.05	0.26	0.28
Unemployed	Observations	664	664	659	620	607	616	371	364	334	230
	R-squared	0.07	0.08	0.07	0.07	0.07	0.08	0.12	0.11	0.09	0.14
Life Satisfaction	Observations	325	325	322	306	299	312	190	190	163	116
	R-squared	0.41	0.41	0.41	0.40	0.41	0.43	0.51	0.51	0.48	0.53

Notes: Includes data from individuals who answered both survey waves. Observations originally excluded by Dynata as low quality are added back for the purposes of this analysis.

Reliability is determined solely based on self-reported "BFI Reliability" in the first survey wave in order to illustrate the effectiveness of our method of using one simple question without needing panel data.

High School GPA is calculated as a percentage.

Individuals are defined as unemployed if they report their current professional status as unemployed (as opposed to being employed or in college).

Life satisfaction is measured using the general question on life satisfaction.

All regressions include demographics, cognitive ability, psychological personality traits, and economic preferences.

Demographics include: sex, age, and country.

Cognitive Ability includes responses to qualitative questions regarding ability.

Personality includes responses to 60 BFI2 questions.

Economic Preferences include responses to qualitative questions regarding preferences for risk, time (including present bias), reciprocity, altruism, and trust.

some to implement. For the remainder of this Section we focus on individuals who answered both survey waves in order to be able to implement existing methods which require panel data, and to compare results.

Table 6 shows the R^2 of regressions of each outcome on latent traits (cognitive ability, personality, preferences) and demographics. The first column includes data from all individuals who answered both survey waves and serves as a point of comparison. The method of FNS shown in Column 9 produces a comparable increase in R^2 as our method (Column 7), which requires neither complex computation nor panel data. Furthermore, our self-reported reliability indicator complements the self-knowledge criterion of FNS. Applied jointly (Column 10), the two methods increase the explanatory power of latent traits above and beyond that which is achieved by either method alone. Methods for reducing measurement error shown in Column 1-6 produce only negligible increases in explanatory power, likely because they leave too many unreliable observations in the sample. We omit the ORIV method from Table 6 because R^2 obtained from instrumental variables regressions has no natural interpretation. It is not comparable to R^2 from regular regressions and thus not useful for the purposes of Table 6.

We next examine the impact of excluding the unreliable type identified by our self-reported reli-

ability measure, on estimated coefficients and statistical significance of estimates, and compare it to the effect of applying the methods used by GSY and by FNS. To this end, we run individual regressions of each outcome (college graduation, unemployment, high school achievement, life satisfaction) on each measured latent construct (cognitivie ability, personality, and preferences), controlling for age, sex, and country. This means that we run separate regressions for each outcome-latent characteristic combination, which we have in our survey. We count the number of instances in which the application of our method increases raw coefficient sizes and the significance of estimates. We divide it by the total number of outcome-latent characteristic pairs and compare the percentage effectiveness of our method to corresponding percentages in which the application of alternatively GSY and of FNS increases the raw coefficient sizes and significance of estimates.

The application of our method increases the absolute magnitude of the estimated relationships in a majority of the cases (84%), as one would expect when correcting for attenuation bias. This is very similar to what is achieved by applying the IV approach used in GSY and the technique of FNS. These increase coefficients of interest in 87% and 70% of the studied cases respectively. In addition, despite excluding a part of the sample, the statistical significance of estimates increases in 48% of cases using our technique compared to 70% of cases using that of GSY and 42% using FNS. As a reminder, while for the implementation of the GSY and FNS corrections we need information also from the second survey wave, for ours we do not (and we did not use it to obtain the above-mentioned estimates).

The IV approach is the textbook cure for the classical measurement error problem, when panel data with repeatedly elicited measures are available.[37] It is thus unsurprising that we find that it generally yields the strongest results. The correction resulting from the application of our proposed method is somewhat weaker than the IV but in general slightly stronger than the one we obtain when applying the FNS method. We interpret this as being due to the fact that our self-reported reliability measure captures both differences in effort and differences in self-knowledge, whereas the FNS measure by design only captures differences in self-knowledge.

5.d Representativeness of the Reliable Type

Applied researchers might be wondering whether relying solely on the reliable type for their empirical analysis might compromise the external validity of their estimates. This would be the case if (1) the reliable sample was substantially different from the unreliable sample, and (2)

[37] In the classical errors-in-variables model, the reliability of an explanatory variable, which can be measured by the test-retest correlation, can also be used to correct for attenuation bias.

the relationship of interest was highly non-linear in the dimensions on which the two samples differed. We do not find evidence for either in our dataset.

As described in Section 5.c, estimated relationships between outcomes and latent traits using the "reliable" sample appear reasonable. They are in general simply stronger (the absolute magnitude of estimated coefficients is larger) than those obtained using the full sample. This is exactly what one would expect to see when eliminating attenuation bias.

Furthermore, while we do find some statistically significant correlates of response reliability (see Section 6), the estimated relationships are too weak to compromise the representativeness of the reliable type. Indeed, there is no evidence that focusing on the "reliable" sample diminishes representativeness on observed characteristics, which researchers usually care about. Table A.2 of the Appendix shows that the self-identified reliable sample has full coverage, relative to the full sample, on all observed characteristics which we measure.[38] Differences in means are also small, rarely exceeding 5%.

Since in addition to observed characteristics we also have extensive information on latent traits and preferences, we test for the support of the "reliable" sample also in terms of personality, economic preferences, cognitive ability, and life satisfaction. Table A.4 of the Appendix shows that the self-identified reliable sample has virtually full coverage also on latent traits and preferences. Any differences in means for the reliable sample on these dimensions are again small.

This is in line with results presented in FNS who derive a theoretical measure of self-knowledge. They find that while self-knowledge is predicted by certain characteristics – and the estimated relationships are statistically significant – the share of explained variation in it is rather low and selection does not play a significant role in their findings. They find no evidence that true relationships between outcomes and latent characteristics differ between the individuals who are above the median in self-knowledge and those who are below median. Instead, they attribute stronger estimated relationships for the former to a reduction in attenuation bias. We observe similar patterns in results when analyzing our "reliable" and "unreliable" subsamples. If a researcher were nevertheless concerned about the representativeness of the "reliable sample", he could simple re-weigh estimates given the virtually full support on all observed (and unobserved) characteristics which we measure.

[38] The only exception is the "other" category in education, which is marginal. We say that the reliable sample has "full coverage" on a particular dimension if its minimum and maximum value on that dimension coincide with those of the full sample.

6 Understanding the Sources of Measurement Error

As can be seen in Table A.14 of the Appendix, individuals with higher cognitive ability report higher reliability. Non-cognitive skills are also important. Conscientiousness, neuroticism, and agreeableness, and openness to experience are associated with a higher reported reliability of answers while extraversion has the opposite effect. From economic preferences, positive reciprocity and risk aversion increases revealed reliability while negative reciprocity and trust decreases revealed reliability. If both personality and preferences are included in a single regression, the respective estimated coefficients decrease, especially for personality. This corroborates Jagelka (2020) finding that economic preferences and psychologists' personality traits are strongly related.

These relationships can be rationalized in two broad categories, each with two subdivisions. The first concerns how well individuals are *able* to answer the survey tasks. There is an external and an internal component to this. The external part governs how well the individual comprehends what is asked of him and his ability to choose the option which best fits his desired response. Cognitive ability and conscientiousness fall intuitively into this category as they plausibly govern how well an individual understands a given task and how much care he puts into answering accurately. The internal part governs an individual's awareness of what answer truly fits him best. Perceptive and curious individuals who have high openness to experience should perform better on this dimension. Furthermore, Jagelka (2020) ties the conscientiousness personality trait to self-knowledge. Thus conscientiousness may impact both the external and internal aspect.

The second broad category concerns how much a given individual *wants* to answer the survey tasks well. Once again, the internal-external distinction may prove useful. The external component will depend on the individual's assessment of how the quality of his answers will impact the monetary benefits which he expects to receive. Even though our survey participants were told that how they answer the survey will not impact their compensation, neurotic and risk averse individuals may not want to risk giving unreliable answers while individuals who are low on trust may also want to hedge their bets. Finally, certain individuals may want to give reliable answers for internal reasons. Those who are agreeable may want to simply help the researcher, those who are high on positive reciprocity may want to reward the researcher for his trust and compensation.

From demographics, only age is statistically significant: younger individuals provide less reliable answers. They plausibly have lower self-knowledge than older individuals as they had less time

to learn about themselves. Experimental treatments have no meaningful influence on reliability.

7 A Simple Model of Measurement Error

We have shown that measurement error is present in standardly used items for measuring constructs such as personality, preferences, and ability which are of eminent interest to economists. We capture it at the measure level across individuals by general test-retest correlations and at the person level across measures by our indicator for revealed individual reliability presented in the preceding section. It is represented by the μ parameter in Equation 1. When unaccounted for, it is responsible for attenuation bias in the estimated coefficient on the noisy explanatory variable.

In Section 4.b we demonstrated that measurement error is decreasing in effort spent answering the assigned tasks (up to some threshold). However, measurement error is present to varying degrees even in responses of individuals who exert effort which should be sufficient for reliably answering the assigned survey tasks. The variation in revealed individual reliability is correlated with self-reported individual reliability even after controlling for a host of individual demographics, preferences, and skills. This suggests that an additional individual characteristic determines the reliability of answers. Let us call it self-knowledge.

We now exposit a simple model of decision-making which rationalizes our findings. As mentioned in Heckman, Jagelka, and Kautz (2021), answering a qualitative survey question can be seen as performing a task. Specifically, the individual's task is to pick the option which best describes him in a menu of alternatives.

Let us now outline the basic assumptions of our model: (1) Individuals are benevolent towards the researcher i.e. they do not deliberately attempt to deceive him;[39] (2) Some individuals have imperfect self-knowledge but each individual has at least some positive level of self-knowledge; (3) Individuals need to exert effort in order to provide informative answers and time spent answering the survey is proportional to effort exerted; (4) The reliability of individuals' answers is increasing in self-knowledge and in effort. The latter only matters up to some threshold i.e. once sufficient effort is spent, exerting more effort will no longer improve reliability.

[39] In our experiment, there is no benefit to lying apart from perhaps presenting oneself in a more positive light. Yet, as the questions we ask generally concern preferences and behavioral tendencies, it is often not even clear what the desirable answer would be (e.g., whether a person should prefer to be more or less risk averse). Furthermore, there is evidence that individuals in general have a preference for telling the truth (see, e.g., Abeler, Nosenzo, and Raymond, 2019).

When individual i starts a survey, he first decides how much effort to put in. In our context, the decision to exert sufficient effort can be understood as an individual's decision to genuinely look within himself and determine which response fits him best.

The minimal required amount of effort, $\underline{E}$, is plausibly a function of survey characteristics such as the number of questions asked and their complexity. For simplicity, we assume that $\underline{E}$ is proper to a survey and common to all individuals who take it. When $E_i < \underline{E}$, we can say that the individual decided not to take the survey seriously (in our terminology, he decided to "rush it"). In this case he may use some form of randomization strategy or heuristic. This idea is similar to the two-stage model of decision-making developed by Belzil and Jagelka (2020) to explain choice behavior on incentivized experiments to elicit economic preferences. It is supported by results of Carpenter and Munro (2022) who find that at some level of effort individuals switch from the more intuitive System 1 choice behavior to more deliberative System 2 reasoning.

If individual i chose effort $E_i > \underline{E}$, the informativeness of the individual's answers will only be constrained by his level of self-knowledge ζ_i. Following Section 5.b, let us assume that there are two unobserved types, one of which has imperfect self knowledge.[40] For the type with imperfect self-knowledge, $\zeta_i \in (0,1)$. We will focus on this type for the remained of this Section.

When choosing from a menu of ordered options on task t, such an individual, even if he decided to seriously look within himself, will only receive a noisy impression s_i of the characteristic of interest, centered around its true value and with a standard deviation $\sigma_i = \frac{1}{\zeta_i} - 1$.[41] He will then choose the option which corresponds to the s_i he received.

Take a characteristic of interest c. The noisy impression s_i drives choice behavior and is unobserved by the econometrician. We can decompose it into its true value c_i and the error with which the individual perceives it, ϵ_i, with a standard deviation σ_i:

$$s_i = c_i + \epsilon_i \tag{4}$$

On the one hand, if a given survey task t results in a continuous measure of c, we will record s_i for each individual. On the other hand, if a given survey task t has n options, we will record some "rounded" version of s_i. One can then define a series of *n-1* ordered thresholds which map the underlying latent variable into the observed discrete choice values.

[40] The reliable type from Section 5.b corresponds to an individual with $E_i \geq \underline{E}$ and $\zeta_i = 1$. The unreliable type corresponds to an individual with $E_i < \underline{E}$ and/or $\zeta_i < 1$.

[41] We subtract 1 to ensure that as an individual nears perfect self-knowledge, σ_i tends to 0. Thus $\sigma_i \in (0, inf)$

On repeat elicitation when an individual faces the same task multiple times within a sufficiently short interval of time,[42] we expect answers to be more consistent when: a) an individual puts in sufficient effort; b) he has better self-knowledge; and c) the task is designed such that the available options have strong discriminatory power i.e. the individual's true choice on the task lies far away from the threshold separating it from the next best option.

7.a Test-retest Correlations

Test-retest correlations are a simple and easily obtainable indicator of the consistency of responses. They are equal to the fraction of the overall cross-sectional variation in *stated* responses which is attributable to the cross-sectional variation in individuals' *true* differences on the underlying characteristic of interest. In our setting where we elicit the measures of interest in two waves, the test-retest correlation can be written as:

$$corr(s1, s2) = \frac{var(c)}{var(c) + var(\epsilon)} = \frac{var(c)}{var(c) + \sigma_i^2} \tag{5}$$

where *s1* and *s2* represent the wave 1 and wave 2 measures respectively of characteristic c which has a cross-sectional variation *var(c)* in the sample and is perceived by individuals with imprecision σ^2 which corresponds to self-knowledge $\zeta = \frac{1}{1+\sigma}$.[43]

If self-knowledge is an individual characteristic and applies across elicited constructs, we can easily extend the test-retest correlation concept to obtain the individual revealed consistency measure which we present in Section 5.a.i.[44] The relevant cross-section then becomes the *available measures for a given* individual rather than measured values of a given construct *across individuals*.[45]

8 Concluding Remarks

This paper presents the first attempt at measuring both *revealed* and *self-reported* reliability of individuals' answers on self-reports of latent characteristics. We show that measurement error

[42] Such that the target latent variable can plausibly be assumed constant.

[43] The underlying assumption is that the wave 1 and wave 2 error terms are uncorrelated.

[44] This assumption is supported in our survey by the fact that self-reported reliability is highly predictive of both revealed individual reliability and of general test-retest reliability across the various constructs contained in our survey (personality, preferences, cognitive ability, life satisfaction). See Tables 2 and 3, and also Tables A.11 and A.12 of the Appendix.

[45] In empirical implementation it is important to standardize the values of the available measures in case they are not on the same scale.

on self-reports relevant to economists is heterogeneous across individuals and can be reasonably approximated by a distribution with two unobserved types. We propose a straightforward survey question which allows to distinguish individuals who give highly reliable answers from those who do not using cross-sectional data. We demonstrate that it predicts revealed individual reliability over and above all measured characterises, survey conditions, and experimental treatments. We show how our simple self-reported reliability measure can be used to cost-effectively reduce attenuation bias in estimates of cognitive and non-cognitive determinants of high school GPA, college graduation, unemployment, and life satisfaction. Without requiring panel data, the achieved correction is similar to some of the most effective reduced-form theory-based approaches in the existing literature.

Selecting or over-weighting individuals based on the reliability of their answers is appropriate when the reliable sample is reasonably representative and/or true relationships between variables of interest are not highly non-linear. We show that in our sample these assumptions appear reasonable. Estimated relationships using our proposed method are in general simply stronger than those obtained without accounting for measurement error. This is exactly what one would expect to see when eliminating attenuation bias. Nevertheless, we encourage researchers to carefully evaluate the appropriateness of these assumptions on a case by case basis.

We demonstrate the construct validity of our measure by examining its link with objective measures of measurement error. We show that individuals who self-identify as reliable have higher test-retest correlations. We benchmark the increase in test-retest correlations when relying on individuals who self-report to be reliable against the increase in test-retest correlations achieved by averaging over multiple measures. We find that the resulting increase in measurement accuracy achieved by our method is comparable to increasing the number of measures per construct from 1 to 5 or from 3 to 12 (i.e. equivalent to the difference in precision between the 3-item-per-trait SOEP personality questionnaire and the more extensive 12-item-per-trait BFI-2 personality questionnaire). Our proposed method is, however, much more cost effective.

Besides offering immediately actionable advice to practitioners, our work opens avenues for future research. We propose an effective indicator for self-reported reliability. It is important to understand the usefulness of this indicator for eliciting an expanded set of latent constructs and in various circumstances. For example, it should also be added to incentivized experiments to test its effectiveness in that context.

As a side benefit, we document test-retest correlations for latent constructs most relevant to

economists: economic preferences, cognitive ability, personality, and life satisfaction. We find that they are substantially below 1. This may explain a number of apparent puzzles in the literature and should be taken into account in empirical research which uses these unobserved characteristics. An analysis of the temporal stability of latent characteristics should take documented test-retest correlations as the relevant baseline comparison, instead of the ideal test-retest correlation of 1. Researchers analyzing the importance of such characteristics in predicting life outcomes need to correct for attenuation bias. Comparisons of the performance of various characteristics should take into account the fact that attenuation bias may affect particular characteristics to different degrees. Extra attention is needed when the characteristics in question are measured using varying numbers of dedicated indicators, as higher estimated impacts may simply be a reflection of more precisely measured characteristics. Time taken to complete experimental tasks or to answer survey questions should be measured and individuals at the low end of the distribution of times should be excluded, at least as a robustness check.

We use our quantitative measure of revealed answer reliability to examine the sources of measurement error for self-reports of latent traits. We find that both an individual's *willingness* and *ability* to provide reliable answers are important. The former depends on the individual's decision to exert effort and engage in genuine introspection, while the latter depends on his degree of self-knowledge.

Effort exhibits threshold behavior. Rushing the survey, characterized by spending less than 2 seconds per survey question, renders answers virtually uninformative. Responses of individuals above this threshold contain 70% less noise relative to those who are below. This is comparable to a reduction in measurement error achieved by increasing the number of indicators per construct from 1 to 12. The reliability of answers provided by individuals who exert sufficient effort is constrained by their level of self-knowledge. We propose a simple model which rationalizes our empirical findings.

9 Bibliography

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for truth-telling." *Econometrica* 87 (4): 1115–1153. [27]

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva. 2023. "Immigration and redistribution." *The Review of Economic Studies* 90 (1): 1–39. [6]

Alós-Ferrer, Carlos, Ernst Fehr, and Nick Netzer. 2021. "Time will tell: Recovering preferences when choices are noisy." *Journal of Political Economy* 129 (6): 1828–1877. [4]

Alós-Ferrer, Carlos, Đura-Georg Granić, Johannes Kern, and Alexander K Wagner. 2016. "Preference reversals: Time and again." *Journal of Risk and Uncertainty* 52:65–97. [4, 13]

Barrick, Murray R, and Michael K Mount. 1991. "The big five personality dimensions and job performance: a meta-analysis." *Personnel Psychology* 44 (1): 1–26. [1]

Barrick, Murray R, Michael K Mount, and Timothy A Judge. 2001. "Personality and performance at the beginning of the new millennium: What do we know and where do we go next?" *International Journal of Selection and assessment* 9 (1-2): 9–30. [1]

Beauchamp, Jonathan P, David Cesarini, and Magnus Johannesson. 2017. "The psychometric and empirical properties of measures of risk preferences." *Journal of Risk and Uncertainty* 54:203–237. [1]

Belzil, Christian, and Tomáš Jagelka. 2020. "Separating True Preferences from Noise and Endogenous Effort." Working paper. University of Bonn. https://drive.google.com/file/d/1JUlXF3lSoFWmXOPwbyrkqwZbI9UOiBQw/view. [6, 28]

Bjørnskov, Christian. 2010. "How comparable are the Gallup World Poll life satisfaction data?" *Journal of happiness Studies* 11:41–60. [7]

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement error in survey data." In *Handbook of Econometrics,* 5:3705–3843. Elsevier. [1]

Caliendo, Marco, Deborah A Cobb-Clark, and Arne Uhlendorff. 2015. "Locus of control and job search strategies." *Review of Economics and Statistics* 97 (1): 88–103. [1]

Carpenter, Jeffrey P, and David Munro. 2022. "Do Losses Trigger Deliberative Reasoning?" IZA Discussion Paper No. 15292. https://docs.iza.org/dp15292.pdf. [28]

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2023. "Econographics." *Journal of Political Economy Microeconomics* Ahead of Print. https://doi.org/doi.org/10.1086/723044. [6]

Cobb-Clark, Deborah A, Sarah C Dahmann, Daniel Kamhöfer, and Hannah Schildberg-Hörisch. 2019. "Self-control: Determinants, life outcomes and intergenerational implications." IZA Discussion Paper. [12]

Cobb-Clark, Deborah A, and Michelle Tan. 2011. "Noncognitive skills, occupational attainment, and relative wages." *Labour Economics* 18 (1): 1–13. [1]

Cozby, Paul C, and Scott C Bates. 2012. *Methods in Behavioral Research.* 11th ed. New York, NY: McGraw-Hill. [9]

Cunha, Flavio, James J Heckman, and Susanne M Schennach. 2010. "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica* 78 (3): 883–931. [6]

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. 2011. "Individual risk attitudes: Measurement, determinants, and behavioral consequences." *Journal of the European Economic Association* 9 (3): 522–550. [1]

Enke, Benjamin, and Thomas Graeber. 2021. "Cognitive Uncertainty in Intertemporal Choice." NBER Working Paper. National Bureau of Economic Research. [3, 5, 17]

Epstein, Seymour. 1979. "The stability of behavior: I. On predicting most of the people much of the time." *Journal of Personality and Social Psychology* 37 (7): 1097. [1, 11]

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global evidence on economic preferences." *The Quarterly Journal of Economics* 133 (4): 1645–1692. [1, 7]

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde. 2022. "The preference survey module: A validated instrument for measuring risk, time, and social preferences." *Management Science.* [6]

Falk, Armin, Thomas Neuber, and Philipp Strack. 2021. "Limited self-knowledge and survey response behavior." CEPR Discussion Paper No. 16345. https://cepr.org/publications/dp16345. [1, 2, 5]

Fehr, Ernst, and Antonio Rangel. 2011. "Neuroeconomic foundations of economic choice—recent advances." *Journal of Economic Perspectives* 25 (4): 3–30. [5]

Fiala, Lenka, John Eric Humphries, Juanna Schrøter Joensen, Uditi Karna, John A List, and Gregory F Veramendi. 2022. "How early adolescent skills and preferences shape economics education choices." In *AEA Papers and Proceedings,* 112:609–13. [1]

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki. 2018. "Speed, accuracy, and the optimal timing of choices." *American Economic Review* 108 (12): 3651–84. [4, 13]

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with measurement error: Techniques with applications to the caltech cohort study." *Journal of Political Economy* 127 (4): 1826–1863. [1, 2, 6, 21, 22]

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes." *American Economic Review* 103 (6): 2052–2086. [1]

Heckman, James J, John Eric Humphries, and Tim Kautz. 2014. *The myth of achievement tests: The GED and the role of character in American life.* University of Chicago Press. [1]

Heckman, James J, Tomáš Jagelka, and Tim Kautz. 2021. "Some contributions of economics to the study of personality." In *Handbook of Personality: Theory and Research,* edited by O. P. John and R. W. Robins, 853–892. The Guilford Press. [1, 4, 27]

Heckman, James J, Seong Hyeok Moon, Rodrigo Pinto, Peter A Savelyev, and Adam Yavitz. 2010. "The rate of return to the HighScope Perry Preschool Program." *Journal of Public Economics* 94 (1-2): 114–128. [1]

Heckman, James J, Jora Stixrud, and Sergio Urzua. 2006. "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior." *Journal of Labor Economics* 24 (3): 411–482. [1]

Huang, Jason L, Paul G Curran, Jessica Keeney, Elizabeth M Poposki, and Richard P DeShon. 2012. "Detecting and deterring insufficient effort responding to surveys." *Journal of Business and Psychology* 27:99–114. [4]

Jagelka, Tomáš. 2020. "Are economists' preferences psychologists' personality traits? A structural approach." IZA Discussion Paper. [1, 3, 5, 6, 17, 26]

List, Greta, John A List, Lina M Ramirez, and Anya Samek. 2022. "Time and risk preferences of children predict health behaviors but not BMI." *Economics Letters* 218:110724. [1]

Liu, Shuo, and Nick Netzer. 2021. "Happy times: Identification from ordered response data." University of Zurich, Department of Economics, Working Paper 371. [4, 13]

Loomes, Graham, and Robert Sugden. 1995. "Incorporating a stochastic element into decision theories." *European Economic Review* 39 (3-4): 641–648. [5]

Lord, Frederic M, and Melvin R Novick. 1968. *Statistical Theories of Mental Test Scores.* With contributions by Allan Birnbaum. Reading, MA: Addison-Wesley Publishing Co. [1]

Meade, Adam W, and S Bartholomew Craig. 2012. "Identifying careless responses in survey data." *Psychological Methods* 17 (3): 437. [4–6, 13]

Mischel, Walter, Ozlem Ayduk, Marc G Berman, BJ Casey, Ian H Gotlib, John Jonides, Ethan Kross, Theresa Teslovich, Nicole L Wilson, Vivian Zayas, et al. 2011. "'Willpower' over the life span: decomposing self-regulation." *Social Cognitive and Affective Neuroscience* 6 (2): 252–256. [1]

Nyhus, Ellen K, and Empar Pons. 2005. "The effects of personality on earnings." *Journal of Economic Psychology* 26 (3): 363–384. [1]

Ratcliff, Roger, and Jeffrey N Rouder. 1998. "Modeling response times for two-choice decisions." *Psychological Science* 9 (5): 347–356. [4]

Read, Blair, Lukas Wolters, and Adam J Berinsky. 2022. "Racing the clock: Using response time as a proxy for attentiveness on self-administered surveys." *Political Analysis* 30 (4): 550–569. [6, 14]

Roberts, Brent W, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. 2007. "The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes." *Perspectives on Psychological Science* 2 (4): 313–345. [1]

Salgado, Jesus F. 1997. "The Five Factor Model of personality and job performance in the European Community." *Journal of Applied Psychology* 82 (1): 30. [1]

Schildberg-Hörisch, Hannah. 2018. "Are risk preferences stable?" *Journal of Economic Perspectives* 32 (2): 135–54. [12]

Schmidt, Frank L, and John E Hunter. 1996. "Measurement error in psychological research: Lessons from 26 research scenarios." *Psychological Methods* 1 (2): 199. [1]

Shadlen, Michael N, and Roozbeh Kiani. 2013. "Decision making as a window on cognition." *Neuron* 80 (3): 791–806. [4]

Smith, Stephanie M, Ian Krajbich, and Ryan Webb. 2019. "Estimating the dynamic role of attention via random utility." *Journal of the Economic Science Association* 5:97–111. [4]

Soto, Christopher J, and Oliver P John. 2017. "The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power." *Journal of Personality and Social Psychology* 113 (1): 117. [6, 7, 9]

Stango, Victor, and Jonathan Zinman. 2020. "We are all behavioral, more or less: A taxonomy of consumer decision making." NBER Working Paper. National Bureau of Economic Research. [1, 6]

Stantcheva, Stefanie. 2022. "How to run surveys: A guide to creating your own identifying variation and revealing the invisible." NBER Working Paper. National Bureau of Economic Research. [6]

Todd, Petra E, and Weilong Zhang. 2020. "A dynamic model of personality, schooling, and occupational choice." *Quantitative Economics* 11 (1): 231–275. [1]

Wise, Steven L, and Xiaojing Kong. 2005. “Response time effort: A new measure of examinee motivation in computer-based tests.” *Applied Measurement in Education* 18 (2): 163–183. [4, 6, 13]

A Appendix

Table A.1: Sample Descriptive Statistics

	Wave 1				Wave 2				
	# Observations	**%**	**Mean**	**Sd**	**# Observations**	**%**	**Mean**	**Sd**	**Predictor of Wave 2 Participation (**)**
Gender	1400				651				
Male		26%	NA	NA		27%	NA	NA	
Female		74%	NA	NA		73%	NA	NA	
Age	1400	NA	22.4	2.2	651		22.5	2.2	+
Country	1400				651				
Australia		26%	NA	NA		27%	NA	NA	
Canada		20%	NA	NA		18%	NA	NA	
United Kingdom		39%	NA	NA		43%	NA	NA	+
USA		16%	NA	NA		12%	NA	NA	-
Occupation	1400				651				
High School Student		4%	NA	NA		4%	NA	NA	
College Student		33%	NA	NA		29%	NA	NA	-
Employed		49%	NA	NA		53%	NA	NA	+
Unemployed		14%	NA	NA		13%	NA	NA	
Other		1%	NA	NA		1%	NA	NA	
Highest Level of Education Attained	1400				651				
Less Than High School		1%	NA	NA		1%	NA	NA	
High School		45%	NA	NA		45%	NA	NA	
Bachelors Degree		38%	NA	NA		40%	NA	NA	
Masters Degree		9%	NA	NA		8%	NA	NA	
Other		7%	NA	NA		6%	NA	NA	
High School GPA (%)	1359	NA	0.84	0.32	636	NA	0.84	0.44	
General Life Satisfaction (0-10)	606	NA	6.53	1.95	651	NA	6.74	2.04	
Mood at Beginning (0-10)	1400	NA	6.52	2.08	651	NA	6.59	2.06	
Mood at End (0-10)	1400	NA	6.64	2.02	651	NA	6.63	2.08	
Self-Reported Answer Reliability (0-10)	1400	NA	9.07	1.32	651	NA	iva	1.27	
Survey on PC	1400	NA	0.52	0.50	651	NA	0.53	0.50	
Length Between Wave1 and Wave2 (days)	651	NA	34.83	14.78	NA	NA	NA	NA	
Time Taken to Complete Experiment (min)	1400	NA	13.70	13.33	651	NA	18.24	16.11	

Notes: The "+"/"-" signs denote a positive/negative correlation respectively, when significant at 5%, of each variable with an individual's participation in the second survey wave.

Table A.2: Sample Descriptive Statistics and Overlap Analysis by Self-Reported Reliability

	Full Sample Wave 1				Reliable Wave 1					
	# Obs	%	Mean	Sd	# Obs	%	Mean	Sd	Difference in Means/Proportions	Reliable Sample Full Coverage
Gender	1400				776					YES
Male		26%	NA	NA		21%	NA	NA	-0.05	YES
Female		74%	NA	NA		79%	NA	NA	0.05	YES
Age	1400	NA	22.4	2.2	776		22.6	2.2	0.19	YES
Country	1400				776					YES
Australia		26%	NA	NA		24%	NA	NA	-0.01	YES
Canada		20%	NA	NA		23%	NA	NA	0.03	YES
United Kingdom		39%	NA	NA		35%	NA	NA	-0.04	YES
USA		16%	NA	NA		18%	NA	NA	0.03	YES
Occupation	1400				776					YES
High School Student		4%	NA	NA		3%	NA	NA	-0.01	YES
College Student		33%	NA	NA		31%	NA	NA	-0.02	YES
Employed		49%	NA	NA		52%	NA	NA	0.03	YES
Unemployed		14%	NA	NA		14%	NA	NA	0.00	YES
Other		1%	NA	NA		1%	NA	NA	0.00	YES
Highest Level of Education Attained	1400				776					YES
Less Than High School		1%	NA	NA		1%	NA	NA	0.00	YES
High School		45%	NA	NA		44%	NA	NA	-0.01	YES
Bachelors Degree		38%	NA	NA		38%	NA	NA	0.00	YES
Masters Degree		9%	NA	NA		8%	NA	NA	-0.01	YES
Other		7%	NA	NA		9%	NA	NA	0.02	
High School GPA (%)	1331	NA	0.82	0.14	738	NA	0.82	0.14	0.00	YES
General Life Satisfaction (0-10)	606	NA	6.53	1.95	356	NA	6.50	2.01	-0.03	YES
Mood at Beginning (0-10)	1400	NA	6.52	2.08	776	NA	6.61	2.15	0.08	YES
Mood at End (0-10)	1400	NA	6.64	2.02	776	NA	6.70	2.12	0.07	YES
Self-Reported Answer Reliability (0-10)	1400	NA	4.33	0.88	776	NA	5.00	0.00	0.67	
Survey on PC	1400	NA	0.52	0.50	776	NA	0.48	0.50	-0.04	YES
Length Between Wave1 and Wave2 (days)	651	NA	34.83	14.78	364	NA	33.98	14.55	-0.85	
Time Taken to Complete Experiment (min)	1400	NA	13.70	13.33	3.6	NA	14.71	13.47	1.02	

Notes: The reliable sample has "full coverage" on a particular dimension if its minimum and maximum value on that dimension coincide with those of the full sample.

Table A.3: Predictors of Wave 2 Participation: Ability, Personality, and Preferences

	Wave 1			Wave 2			
	#Obs	**Mean**	**Sd**	**#Obs**	**Mean**	**Sd**	**Predictor of Wave 2 Participation (**)**
BFI2 Extraversion	1400	36.0	7.5	651	35.9	7.4	
Sociability	1400	11.3	3.5	651	11.3	3.6	
Assertiveness	1400	11.8	3.1	651	11.7	3.0	
Energy	1400	13.0	2.9	651	13.0	2.8	
BFI2 Conscientiousness	1400	41.3	7.8	651	42.1	8.1	+
Organization	1400	14.1	3.4	651	14.4	3.5	+
Productiveness	1400	13.3	3.0	651	13.6	3.1	+
Responsibility	1400	13.9	2.8	651	14.1	2.8	+
BFI2 Neuroticism	1400	38.0	8.8	651	37.6	8.9	
Anxiety	1400	13.9	3.3	651	13.8	3.3	
Depression	1400	11.9	3.5	651	11.8	3.5	
Emotional Volatility	1400	12.3	3.4	651	12.0	3.5	-
BFI2 Agreeableness	1400	42.4	6.9	651	42.6	6.9	
Compassion	1400	14.6	2.9	651	14.8	2.8	
Respectfulness	1400	15.1	3.0	651	15.2	2.9	
Trust	1400	12.6	2.7	651	12.6	2.7	
BFI2 Openness to Experience	1400	41.1	6.7	651	40.9	6.7	
Curiosity	1400	14.3	2.8	651	14.3	2.8	
Aesthetic_Sense	1400	13.2	3.0	651	13.1	3.1	
Imagination	1400	13.6	2.7	651	13.5	2.8	
SOEP Extraversion	1400	9.1	2.6	651	9.0	2.7	
SOEP Conscientiousness	1400	10.7	2.2	651	10.9	2.3	+
SOEP Neuroticism	1400	10.5	2.8	651	10.4	2.7	
SOEP Agreeableness	1400	11.1	2.2	651	11.3	2.1	
SOEP Openness to Experience	1400	15.0	2.7	651	14.9	2.8	
Risk Tolerance	1400	7.2	2.1	651	7.2	2.0	-
Patience	1400	8.1	1.8	651	8.1	1.8	
Present Bias	1400	6.9	2.5	651	6.8	2.5	-
Altruism	1400	8.4	1.9	651	8.3	2.0	
Trust	1400	6.6	2.5	651	6.5	2.4	
Positive Reciprocity	1400	9.3	1.6	651	9.3	1.6	
Neg Reciprocity Self	1400	6.2	2.5	651	6.0	2.5	
Negative Reciprocity Self2	1400	5.3	2.7	651	5.4	2.7	
Neg Reciprocity Other	1400	6.9	2.3	651	6.6	2.3	
Gallup General Life Satisfaction	606	7.5	2.0	651	7.7	2.0	
SWLS	606	6.8	2.0	651	7.1	2.2	
Mood at Beginning of Survey	1400	7.5	2.1	651	7.6	2.1	
Mood at End of Survey	1400	7.6	2.0	651	7.6	2.1	
Ability Computer	1400	8.7	1.8	651	8.8	1.7	
Ability Writing	1400	8.6	1.8	651	8.7	1.8	
Ability Reading	1400	9.0	1.7	651	9.2	1.6	
Ability Communication	1400	8.0	2.0	651	8.1	2.0	
Ability Problem-Solving	1400	8.2	1.7	651	8.2	1.6	
Ability Math	1400	7.5	2.3	651	7.7	2.2	

Notes: The "+"/"-" signs denote a positive/negative correlation respectively, when significant at 5%, of each variable with an individual's participation in the second survey wave.

Table A.4: Latent Trait Overlap Analysis by Self-Reported Reliability

	Full Sample Wave 1			Reliable Wave 1				
	# Obs	Mean	Sd	# Obs	Mean	Sd	Difference in Means	Reliable Sample Full Coverage
BFI2 Extraversion (out of 60)	1400	36.0	7.5	776	36.3	8.7	0.2	YES
BFI2 Conscientiousness (out of 60)	1400	41.3	7.8	776	43.3	8.2	2.0	YES
BFI2 Neuroticism (out of 60)	1400	38.0	8.8	776	38.4	10.1	0.3	YES
BFI2 Agreeableness (out of 60)	1400	42.4	6.9	776	44.2	7.1	1.9	YES
BFI2 Openness to Experience (out of 60)	1400	41.1	6.7	776	42.7	7.2	1.5	YES
Risk Tolerance (out of 10)	1400	6.2	2.1	776	6.1	2.3	-0.1	YES
Patience (out of 10)	1400	7.1	1.8	776	7.2	1.8	0.2	YES
Present Bias (out of 10)	1400	5.9	2.5	776	5.7	2.8	-0.2	YES
Altruism (out of 10)	1400	7.4	1.9	776	7.7	2.0	0.4	YES
Trust (out of 10)	1400	5.6	2.5	776	5.4	2.7	-0.2	YES
Positive Reciprocity (out of 10)	1400	8.3	1.6	776	8.8	1.4	0.5	YES
Neg Reciprocity Self (out of 10)	1400	5.2	2.5	776	5.0	2.7	-0.2	YES
Negative Reciprocity Self2 (out of 10)	1400	4.3	2.7	776	3.8	2.8	-0.6	YES
Neg Reciprocity Other (out of 10)	1400	5.9	2.3	776	5.9	2.5	0.0	YES
Gallup General Life Satisfaction (out of 10)	606	6.5	2.0	356	6.5	2.0	0.0	YES
SWLS (out of 10)	606	5.8	2.0	356	5.6	2.2	-0.2	YES
Cognitive (out of 50)	1400	50.0	8.2	776	51.3	8.2	1.3	YES

Notes: The reliable sample has "full coverage" on a particular dimension if its minimum and maximum value on that dimension coincide with those of the full sample.

Table A.5: Test-retest Correlations of Standardly Used Qualitative Behavioral Measures: Impact of Demographics, Treatments, and Survey Conditions

				Impact on Test-Retest Correlations Significant at 5%							
Group Personality	Instrument	Construct	Test-Retest Correlation	Male	Young	Recontact time > median (5 weeks)	PC	Extra Incentives (3 Euros)	Double Extra Incentives	BFI First	GPA First
Personality	BFI-2 60-item	Extraversion	0.85	–							
		Conscientiousness	0.83								
		Neuroticism	0.85		–						
		Agreeableness	0.77								
		Openness to Experience	0.78						–		
	SOEP	Extraversion	0.79		–						
		Conscientiousness	0.68					–			
		Neuroticism	0.78								
		Agreeableness	0.65	–							
		Openness to Experience	0.68								
Economic Preference	Global Preference Survey	Risk Tolerance	0.71								
	1-item	Patience	0.42							–	
		Present Bias	0.58		–						
		Altruism	0.57			–		–			+
		Trust	0.6			–					
		Positive Reciprocity	0.53								–
		Neg Reciprosity Self	0.56								
		Negative Reciprosity Self2	0.61								
		Neg Reciprosity Other	0.48	+							
Well-being	Gallup 1-item	Life Satisfaction	0.77								+
	SWLS 5-item	Life Satisfaction	0.72				–			–	
Cognitive Ability	Qualitative Assessment 6-item	Cognitive Ability	0.72	+				–	–		

Notes: The "+"/"-" signs denote a positive/negative coefficient respectively, when significant at 5%, of each variable in the column header on the test-retest correlation of a particular construct.

Table A.6: Test-retest Correlations of Standardly Used Qualitative Behavioral Measures: Existing Literature

Group	Instrument	Construct	Dohmen and Jagelka (2022)	Soto and John (2017)	Lang et al. (2011)	Beauchamp et al. (2017)	Krueger & Schkade (2008)
Personality	BFI-2	Extraversion	**0.85**	0.84			
		- Sociability	**0.82**	0.83			
		- Assertiveness	**0.74**	0.80			
		- Energy	**0.68**	0.74			
		Conscientiousness	**0.83**	0.83			
		- Organization	**0.77**	0.76			
		- Productiveness	**0.75**	0.74			
		- Responsibility	**0.71**	0.68			
		Neuroticism	**0.85**	0.81			
		- Anxiety	**0.77**	0.79			
		- Depression	**0.79**	0.74			
		- Emotional Volatility	**0.75**	0.70			
		Agreeableness	**0.77**	0.76			
		- Compassion	**0.67**	0.68			
		- Respectfulness	**0.69**	0.66			
		- Trust	**0.64**	0.75			
		Openness to Experience	**0.78**	0.76			
		- Curiosity	**0.67**	0.78			
		- Aesthetic_Sense	**0.72**	0.67			
		- Imagination	**0.69**	0.67			
	SOEP	Extraversion	**0.79**		0.81 / 0.87 / 0.79		
		Conscientiousness	**0.68**		0.70 / 0.70 / 0.66		
		Neuroticism	**0.78**		0.81 / 0.84 / 0.80		
		Agreeableness	**0.65**		0.75 / 0.85 / 0.74		
		Openness to Experience	**0.68**		0.72 / 0.75 / 0.73		
Economic Preference	Global Preference Survey	Risk Tolerance	**0.71**			0.633	
		Patience	**0.42**				
		Present Bias	**0.58**				
		Altruism	**0.57**				
		Trust	**0.60**				
		Positive Reciprocity	**0.53**				
		Neg Reciprocity Self	**0.56**				
		Negative Reciprocity Self2	**0.61**				
		Neg Reciprocity Other	**0.48**				
Well-being	Gallup 1-item	Life Satisfaction	**0.77**				0.40–0.66
	SWLS 5-item	Life Satisfaction	**0.72**				0.50-0.84
	Current Mood	Mood at Beginning of Survey	**0.61**				
		Mood at End of Survey	**0.65**				
Cognitive Ability	Qualitative Assesment	Ability Computer	**0.62**				
		Ability Writing	**0.68**				
		Ability Reading	**0.60**				
		Ability Communication	**0.64**				
		Ability Problem-Solving	**0.58**				
		Ability Math	**0.72**				

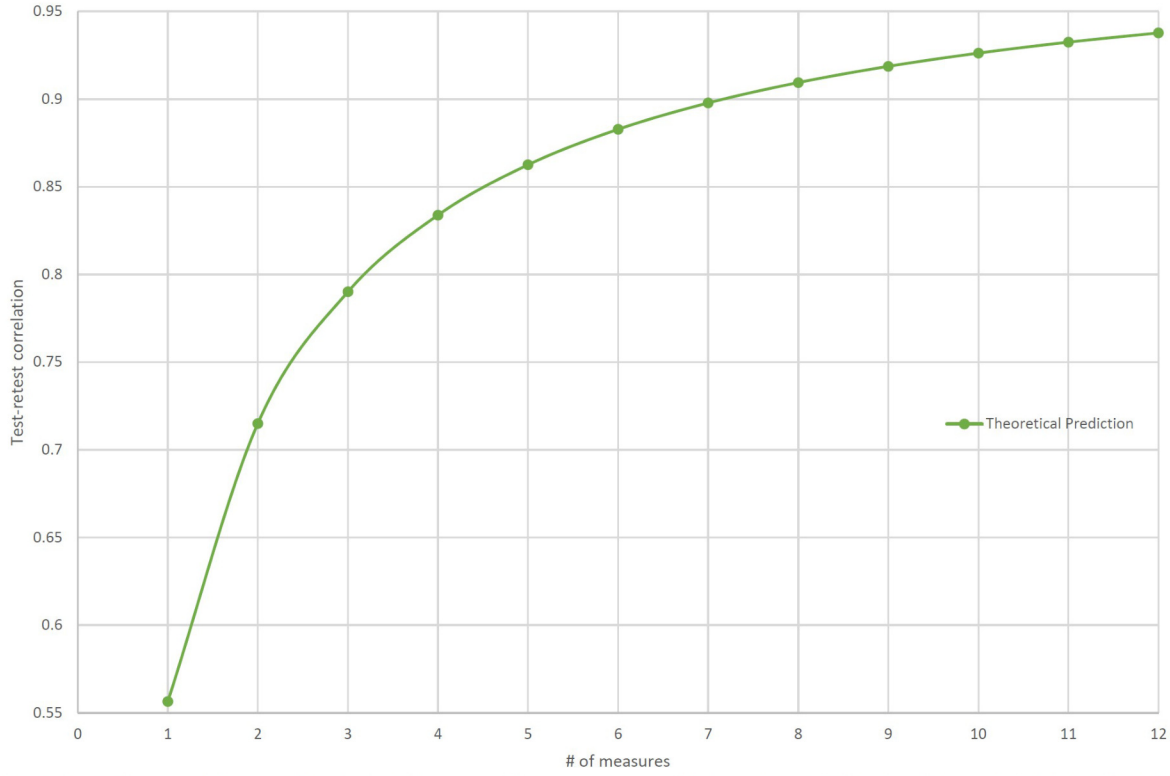
Notes: The test-retest correlations from Lang et al. (2011) pertain respectively to a sample of Young Adults (N=4,232) / Middle-Aged Adults (N=5,503) / Older Adults (N=3,724).

Table A.7: Test-retest Correlations of Standardly Used Qualitative Behavioral Measures: Single Items

Group	Instrument	Item Text	Test-Retest Correlation
Personality	BFI-2 60-item	Is outgoing, sociable	0.74
		Is compassionate, has a soft heart	0.58
		Tends to be disorganized	0.65
		Is relaxed, handles stress well	0.66
		Has few artistic interests	0.41
		Has an assertive personality	0.59
		Is respectful, treats others with respect	0.45
		Tends to be lazy	0.65
		Stays optimistic after experiencing a setback	0.62
		Is curious about many different things	0.62
		Rarely feels excited or eager	0.49
		Tends to find fault with others	0.49
		Is dependable, steady	0.41
		Is moody, has up and down mood swings	0.60
		Is inventive, finds clever ways to do things	0.47
		Tends to be quiet	0.62
		Feels little sympathy for others	0.46
		Is systematic, likes to keep things in order	0.48
		Can be tense	0.53
		Is fascinated by art, music, or literature	0.67
		Is dominant, acts as a leader	0.67
		Starts arguments with others	0.54
		Has difficulty getting started on tasks	0.53
		Feels secure, comfortable with self	0.58
		Avoids intellectual, philosophical discussions	0.53
		Is less active than other people	0.58
		Has a forgiving nature	0.51
		Can be somewhat careless	0.59
		Is emotionally stable, not easily upset	0.60
		Has little creativity	0.52
		Is sometimes shy, introverted	0.57
		Is helpful and unselfish with others	0.39
		Keeps things neat and tidy	0.61
		Worries a lot	0.63
		Values art and beauty	0.58
		Finds it hard to influence people	0.49
		Is sometimes rude to others	0.57
		Is efficient, gets things done	0.49
		Often feels sad	0.63
		Is complex, a deep thinker	0.48
		Is full of energy	0.57
		Is suspicious of others' intentions	0.54
		Is reliable, can always be counted on	0.45
		Keeps their emotions under control	0.53
		Has difficulty imagining things	0.56
		Is talkative	0.65
		Can be cold and uncaring	0.65
		Leaves a mess, doesn't clean up	0.60
		Rarely feels anxious or afraid	0.49
		Thinks poetry and plays are boring	0.60
		Prefers to have others take charge	0.54
		Is polite, courteous to others	0.47
		Is persistent, works until the task is finished	0.51
		Tends to feel depressed, blue	0.65
		Has little interest in abstract ideas	0.50
		Shows a lot of enthusiasm	0.55
		Assumes the best about people	0.53
		Sometimes behaves irresponsibly	0.59
		Is temperamental, gets emotional easily	0.60
		Is original, comes up with new ideas	0.49

	SOEP	Does a thorough job	0.50
		Is reserved	0.55
		Has an active imagination	0.56
		Gets nervous easily	0.64
		Is eager for knowledge	0.49
		Is considerate and kind to others	0.46
Economic Preference	Global Preference Survey	Please tell us, in general, how willing or unwilling you are to take risks.	0.71
		How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?	0.42
		I tend to postpone tasks even if I know it would be better to do them right away.	0.58
		How willing are you to give to good causes without expecting anything in return?	0.57
		I assume that people have only the best intentions.	0.60
		When someone does me a favor I am willing to return it.	0.53
		How willing are you to punish someone who treats you unfairly, even if there may be costs for you?	0.56
		If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so.	0.61
		How willing are you to punish someone who treats others unfairly, even if there may be costs for you?	0.48
Well-being	Gallup 1-item	All things considered, how satisfied are you with your life?	0.77
	SWLS 5-item	In most ways my life is close to my ideal.	0.64
		The conditions of my life are excellent.	0.64
		I am satisfied with my life.	0.71
		So far I have gotten the important things I want in life.	0.50
		If I could live my life over, I would change almost nothing.	0.64
Cognitive Ability	Qualitative Assessment	How would you rate your ability to use a computer? For example, using software applications, programming, or using a computer to find or process information.	0.62
		How would you rate your writing abilities? For example, writing to get across information or ideas to others, or editing writing to improve it.	0.68
		How would you rate your reading abilities? For example, understanding what you read and identifying the most important issues, or using written material to find information.	0.60
		How would you rate your oral communication abilities? For example, explaining ideas to others, speaking to an audience, or participating in discussions.	0.64
		How would you rate your ability to solve new problems? For example, identifying problems and possible causes, planning strategies to solve problems, or thinking of new ways to solve problems.	0.58
		How would you rate your mathematical abilities? For example, using formulas to solve problems, interpreting graphs or tables, or using math to figure out practical things in everyday life.	0.72

Figure A.1: Test-retest Correlations of Big 5 Personality Traits as a Function of the Number of Measures Used: Comparison with a Theoretical Benchmark



Notes: Equation 2 of the paper is the starting point for predicting what happens as multiple measures become available for a noisy construct. It directly yields the formula for the test-retest correlation of a 1-measure construct: *tr(1)= var(c)/[var(c)+var(ε)]*. When *n* measures of the same construct are available, the test-retest correlation of the construct obtained as the average of its *n* measures can be expressed as: *tr(n)= var(c)/[var(c)+var(ε)/n]*.

Figure A.2: Distribution of Self-Reported Overall Survey Reliability

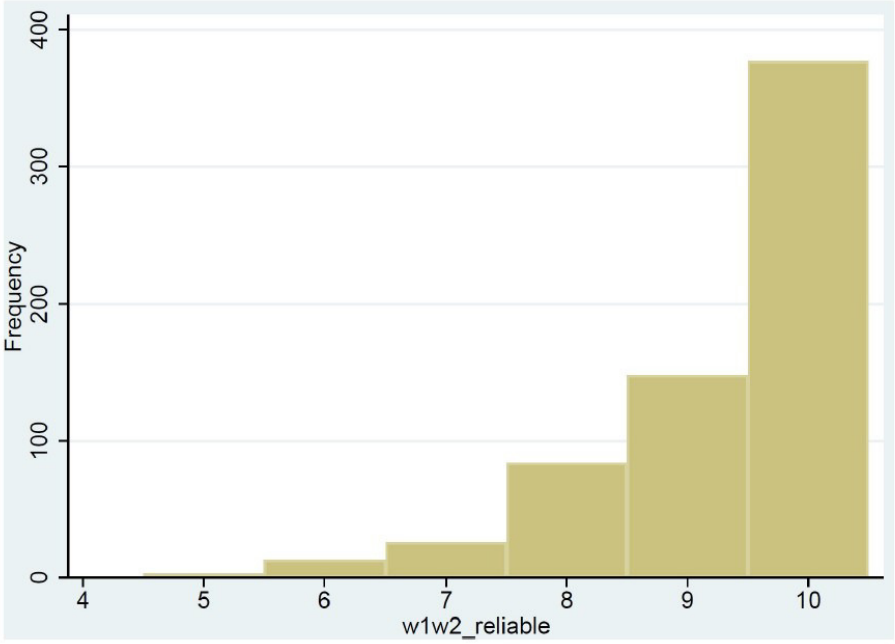


Figure A.3: Distribution of Self-Reported BFI Reliability

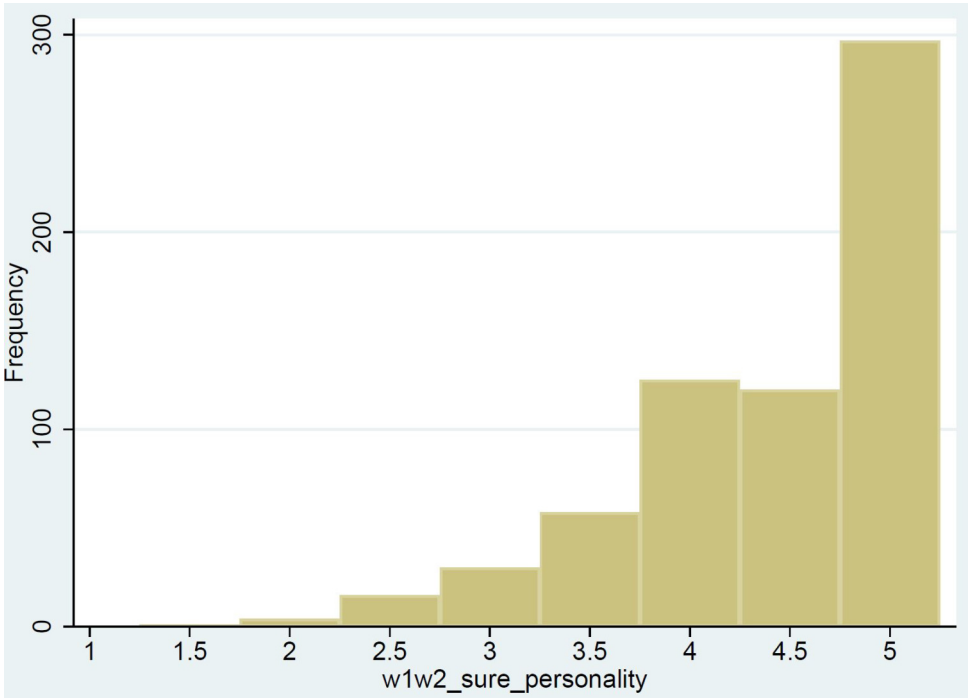


Table A.8: Impact of Delay Between Wave1 and Wave2 on Revealed Individual Reliability

	(1)	(2)	(3)	(4)	(5)	(6)
			Individual Revealed Reliability			
VARIABLE	All Individuals		Self-Reported Unreliable		Self-Reported Reliable	
Recontact Time (days)	0.00		-0.00		-0.00	
	(0.00)		(0.00)		(0.00)	
Recontact Time >6 weeks		-0.02		0.00		-0.03
		(0.02)		(0.03)		(0.02)
Constant	0.55***	0.54***	0.46***	0.44***	0.65***	0.65***
	(0.02)	(0.01)	(0.03)	(0.02)	(0.03)	(0.01)
Observations	651	651	354	354	297	297
R-squared	0.00	0.00	0.00	0.00	0.00	0.01

Standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A.9: Test-retest Correlations by Self-Reported Reliability Alternatively Using Only Reliability Information from the First Wave of Data Collection and Excluding Individuals who "Rushed" the Survey

	Excluding Individuals who "Rushed" the Survey				Only Using Wave 1 Answers to the Determine Self-Reported Reliability			
Trait	Sure about BFI	Unsure about BFI	Reliable Survey Answers	Unreliable Survey Answers	Sure about BFI	Unsure about BFI	Reliable Survey Answers	Unreliable Survey Answers
BFI-2 Extraversion	0.89	0.79	0.89	0.75	0.88	0.75	0.87	0.70
BFI-2 Conscientiousness	0.88	0.78	0.87	0.78	0.85	0.75	0.84	0.75
BFI-2 Neuroticism	0.88	0.83	0.87	0.83	0.87	0.82	0.87	0.77
BFI-2 Agreeableness	0.80	0.71	0.80	0.68	0.79	0.66	0.79	0.63
BFI-2 Openness to Experience	0.81	0.73	0.80	0.74	0.79	0.73	0.78	0.75
SOEP Extraversion	0.87	0.69	0.84	0.68	0.84	0.67	0.81	0.69
SOEP Conscientiousness	0.79	0.58	0.75	0.61	0.74	0.52	0.71	0.51
SOEP Neuroticism	0.84	0.74	0.82	0.74	0.83	0.68	0.80	0.66
SOEP Agreeableness	0.74	0.55	0.68	0.61	0.71	0.47	0.65	0.57
SOEP Openness to Experience	0.78	0.55	0.75	0.55	0.73	0.57	0.71	0.54
GPS Risk Tolerance	0.77	0.64	0.74	0.64	0.75	0.64	0.72	0.67
GPS Patience	0.45	0.35	0.43	0.35	0.44	0.39	0.41	0.49
GPS Present Bias	0.64	0.47	0.61	0.48	0.61	0.49	0.60	0.46
GPS Altruism	0.63	0.50	0.60	0.49	0.62	0.48	0.59	0.45
GPS Trust	0.64	0.53	0.61	0.55	0.62	0.56	0.62	0.50
GPS Pos Reciprocity	0.56	0.43	0.52	0.44	0.54	0.46	0.52	0.43
GPS Neg Reciprocity Self	0.62	0.44	0.59	0.44	0.58	0.50	0.57	0.51
GPS Neg Reciprocity Self2	0.64	0.53	0.65	0.46	0.61	0.56	0.64	0.47
GPS Neg Reciprocity Other	0.51	0.43	0.51	0.41	0.50	0.43	0.49	0.43
Gallup General Life Satisfaction	0.83	0.72	0.80	0.72	0.80	0.73	0.77	0.75
SWLS	0.73	0.71	0.72	0.73	0.72	0.74	0.72	0.70
Ability Computer	0.63	0.57	0.59	0.59	0.63	0.59	0.59	0.65
Ability Writing	0.75	0.60	0.74	0.52	0.71	0.61	0.65	0.65
Ability Reading	0.70	0.49	0.65	0.45	0.65	0.51	0.59	0.54
Ability Communication	0.73	0.55	0.68	0.55	0.70	0.54	0.65	0.58
Ability Problem-Solving	0.67	0.54	0.62	0.58	0.62	0.53	0.59	0.54
Ability Math	0.83	0.60	0.78	0.59	0.79	0.58	0.74	0.62
Observations	293	302	389	206	364	287	485	166

Notes: In the first four columns, Test-retest correlations highlighted in red are higher for individuals who reported a high reliability of answers (i.e. overall self-reported reliability >= 9/10 on both survey waves; self-reported BFI reliability = 5/5 on both survey waves).

In the last four columns, test-retest correlations highlighted in red are higher for individuals who reported a high reliability of answers (i.e. overall self-reported reliability >= 9/10 on the first survey wave; self-reported BFI reliability = 5/5 on the first survey wave).

Table A.10: Impact of Survey Time on Revealed Reliability: Effort Threshold Robustness

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Revealed Individual Reliability				
Wave 1 Survey Time	0.00				
	(0.00)				
Wave 2 Survey Time	0.00				
	(0.00)				
Combined Survey Time		0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)
Rush, fastest 5%	−0.21***	−0.21***			
	(0.03)	(0.03)			
Rush, fastest 10%			−0.2***		
			(0.02)		
Rush, fastest 15%				−0.18***	
				(0.02)	
Rush, fastest 20%					−0.16***
					(0.02)
Constant	0.54***	0.54***	0.57***	0.58***	0.59***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Observations	651	651	651	651	651
R-squared	0.07	0.07	0.10	0.10	0.10

Standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes: The "rush" dummies indicate whether an individual was below the 5th/10th/15th/20th percentile respectively in survey times on either wave.

Table A.11: Revealed vs. Self-Reported Individual Survey Reliability: Personality Measures

	(1)	(2)	(3)
VARIABLES	Revealed Individual Reliability		
Self-Reported BFI Reliability	0.12***		0.11***
	(0.01)		(0.01)
Self-Reported Overall Reliability		0.08***	0.04***
		(0.01)	(0.01)
Self-Reported BFI Reliability#Self-Reported Overall Reliability			0.04***
			(0.01)
Constant	0.51***	0.51***	0.49***
	(0.01)	(0.01)	(0.01)
Observations	651	651	651
R-squared	0.18	0.08	0.20

Standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes: Revealed individual reliability represents the test-retest correlation of an individual's responses on 66 personality measures included in the survey.

Table A.12: Revealed vs. Self-Reported Individual Survey Reliability: Non-Personality Measures

	(1)	(2)	(3)
VARIABLES	Revealed Individual Reliability		
Self-Reported BFI Reliability	0.07***		0.07***
	(0.01)		(0.02)
Self-Reported Overall Reliability		0.03**	0.03*
		(0.01)	(0.02)
Self-Reported BFI Reliability#Self-Reported Overall Reliability			0.07***
			(0.02)
Constant	0.52***	0.52***	0.49***
	(0.01)	(0.01)	(0.01)
Observations	651	651	651
R-squared	0.03	0.01	0.06

Standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes: Revealed individual reliability represents the test-retest correlation of an individual's responses on 21 non-personality measures included in the survey.

Table A.13: Revealed Individual vs. Combined Self-Reported Reliability: Controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Revealed Individual Reliability							
Self-Reported Combined Reliability	0.10***	0.09***	0.10***	0.11***	0.11***	0.08***	0.07***	0.06***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Rushed Survey		x						x
Experimental Treatments			x					x
Demographics				x				x
Cognitive Ability					x			x
Personality						x		x
Economic Preferences							x	x
Constant	0.52***	0.53***	0.52***	0.56***	0.52***	0.52***	0.52***	0.59***
	(0.01)	(0.01)	(0.02)	(0.03)	(0.01)	(0.01)	(0.01)	(0.04)
Observations	651	651	651	613	651	651	651	613
R-squared	0.15	0.17	0.15	0.18	0.15	0.20	0.22	0.28

Standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes: The combined reliability indicator combines information from both the "BFI reliability" and "Overall Reliability" self-reports.

Experimental Treatments include: extra incentives, order of personality section, order of ability section.

Demographics include: sex, age, country, current profession, highest achieved education, and using a pc/handheld device for this survey.

Rushed Survey includes a dummy variable indicating whether an individual was below the 5th percentile in survey times on either survey wave.

Cognitive Ability includes qualitative questions regarding ability.

Personality includes 60 BFI2 questions.

Economic Preferences include qualitative questions regarding preferences for risk, time (including present bias), reciprocity, altruism, and trust.

Regressions which include demographics exclude marginal categories which have near 0 mass. This excludes 38 individuals.

Table A.14: Correlates of Revealed Individual Reliability

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Revealed Individual Reliability						
Extra Income	−0.01						−0.01
	(0.02)						(0.02)
BFI2 Section First	-0.00						−0.01
	(0.02)						(0.02)
GPA Question First	0.03						0.04**
	(0.02)						(0.02)
Rushed Survey		−0.22***					−0.13***
		(0.03)					(0.03)
BFI-2 Extraversion					−0.02**		−0.01
					(0.01)		(0.01)
BFI-2 Conscientiousness					0.04***		0.00
					(0.01)		(0.01)
BFI-2 Neuroticism					0.03**		0.00
					(0.01)		(0.01)
BFI-2 Agreeableness					0.04***		0.02
					(0.01)		(0.01)
BFI-2 Openness to Experience					0.04***		0.02*
					(0.01)		(0.01)
Cognitive Ability				0.02**			0.02*
				(0.01)			(0.01)
GPS Risk Tolerance						−0.03***	−0.03***
						(0.01)	(0.01)
GPS Patience						−0.00	−0.00
						(0.01)	(0.01)
GPS Present Bias						−0.01	−0.00
						(0.01)	(0.01)
GPS Neg Reciprocity Self						0.01	0.01
						(0.02)	(0.02)
GPS Neg Reciprocity Self2						−0.05***	−0.03**
						(0.01)	(0.02)
GPS Neg Reciprocity Other						0.00	−0.00
						(0.01)	(0.01)
GPS Altruism						0.01	0.00
						(0.01)	(0.01)
GPS Pos Reciprocity						0.06***	0.04***
						(0.01)	(0.01)
GPS Trust						−0.04***	-0.04***
						(0.01)	(0.01)
Male			−0.03				0.02
			(0.02)				(0.02)
Age ≤21			−0.07***				−0.06***
			(0.02)				(0.02)
Took Survey on PC			−0.02				−0.02
			(0.02)				(0.02)
Country: Canada			0.01				0.02
			(0.03)				(0.03)
Country: UK			−0.02				0.00
			(0.02)				(0.02)
Country: USA			0.02				0.01
			(0.03)				(0.03)
Profession: Employed			−0.03				−0.03
			(0.02)				(0.02)
Profession: Unemployed			0.01				−0.02
			(0.03)				(0.03)
Highest Edu: Bachelors			−0.02				0.00
			(0.02)				(0.02)
Highest Edu: Masters			0.02				0.05
			(0.04)				(0.03)
Highest Edu: Other			0.07				0.04
			(0.04)				(0.04)
Constant	0.52***	0.55***	0.59***	0.53***	0.53***	0.53***	0.56***
	(0.02)	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)	(0.03)
Observations	651	651	613	651	651	651	613
R-squared	0.00	0.07	0.04	0.01	0.13	0.17	0.26

Standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$