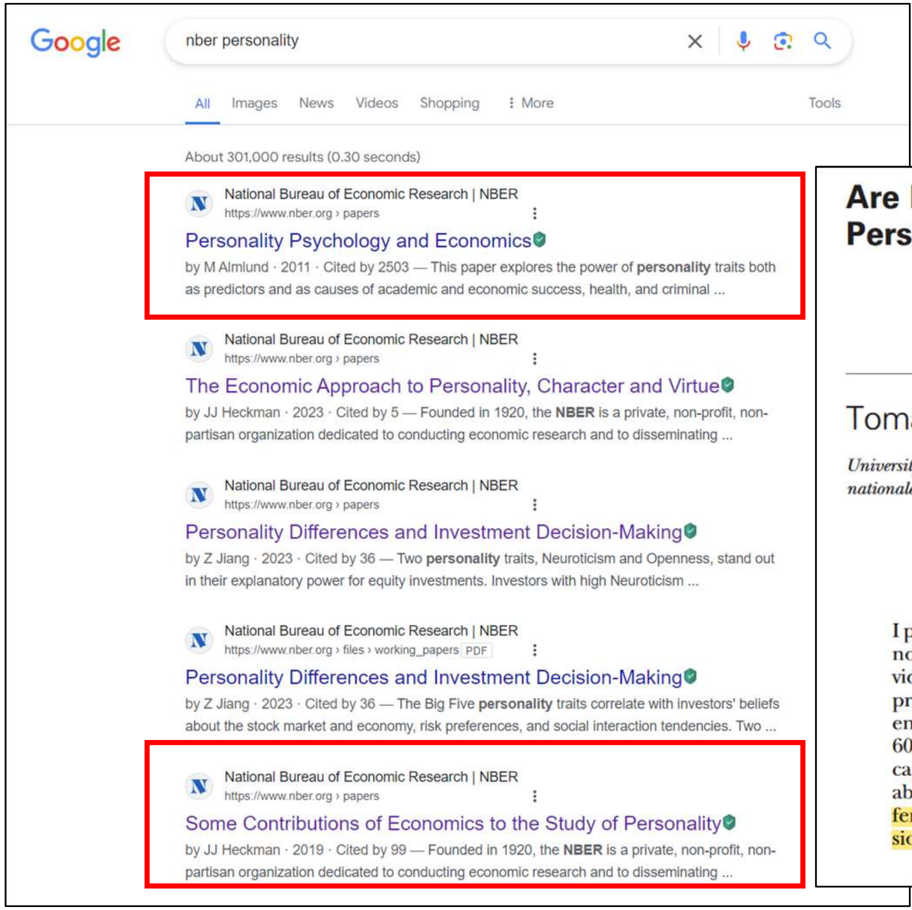
Accounting for Individual-Specific Reliability of Self-Assessed Measures of Economic Preferences and Personality Traits

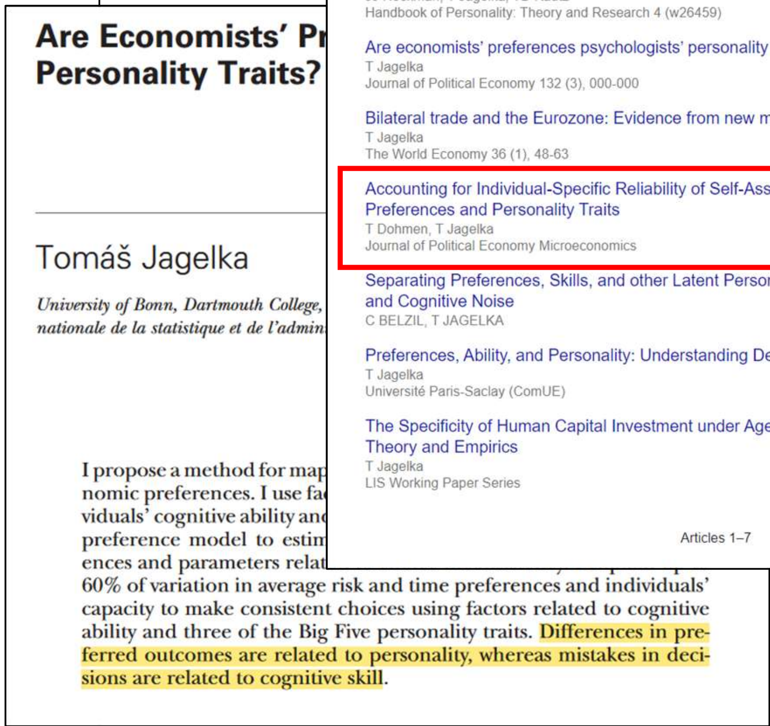
Thomas DOHMEN and Tomáš JAGELKA (2023, forthcoming Journal of Political Economy Microeconomics)

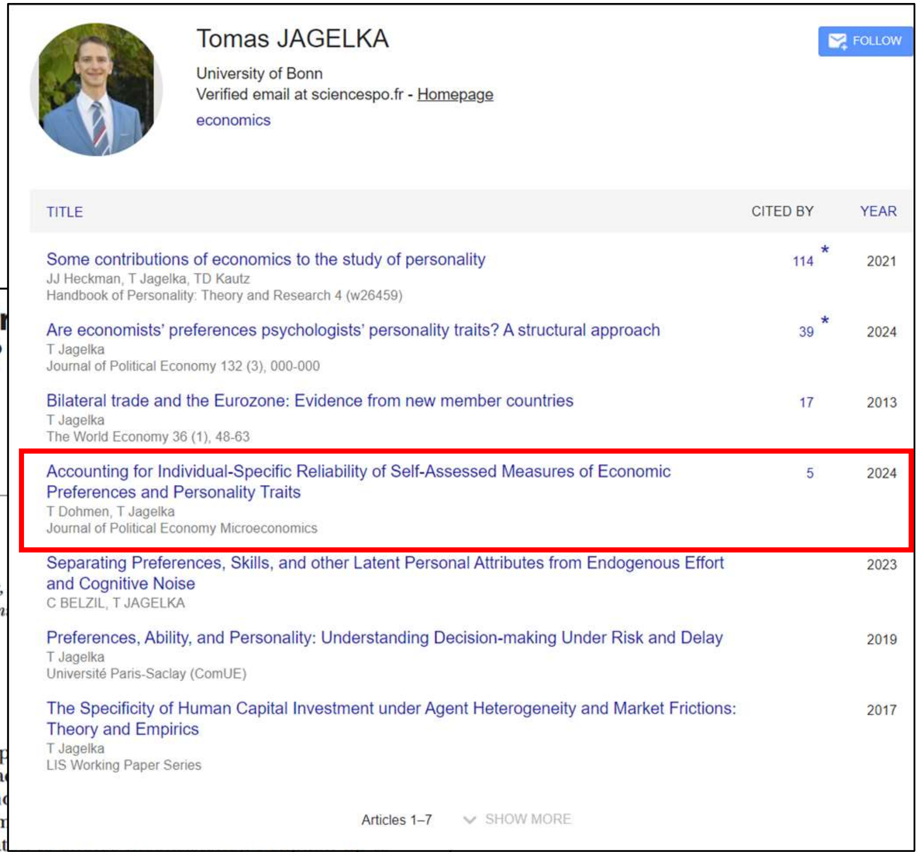
Self reports about self report

Journal club presentation by Helena Luo (16/04/2024)

How did I found this paper







Why this paper?

- Survey question to distinguish highly reliable “respondents” from not reliable

- Easy to add into survey or experiment
- Allows researchers to account for measurement error using cross-sectional data only

Abstract

We measure both revealed and self-reported reliability of individuals' answers on self-reports of latent characteristics. We propose a straightforward survey question which allows to distinguish individuals who give highly reliable answers from those who do not. Our novel indicator can be used to cost-effectively reduce attenuation bias in estimates of cognitive and non-cognitive determinants of key life outcomes. Without requiring panel data, the achieved correction is similar to 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.

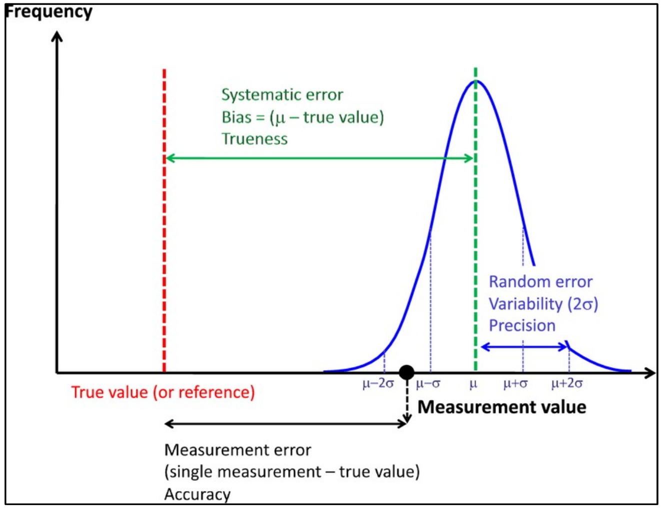
Self-assessments and measurement error

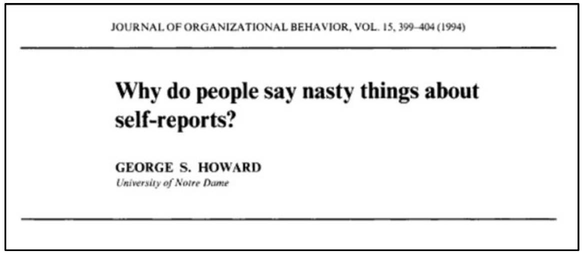
Self-assessments and measurement error

- Important source of information (constructs not typically observable)
  - Measurement error (effort and self-knowledge)
  - Hard to incentivize properly

- Error reduces precision of measurements → harder to find true relationship between constructs and outcomes

- Usually use panel data or other cumbersome methods to overcome measure error





JOURNAL OF ORGANIZATIONAL BEHAVIOR, VOL. 15, 399–404 (1994)

Why do people say nasty things about self-reports?

GEORGE S. HOWARD
University of Notre Dame

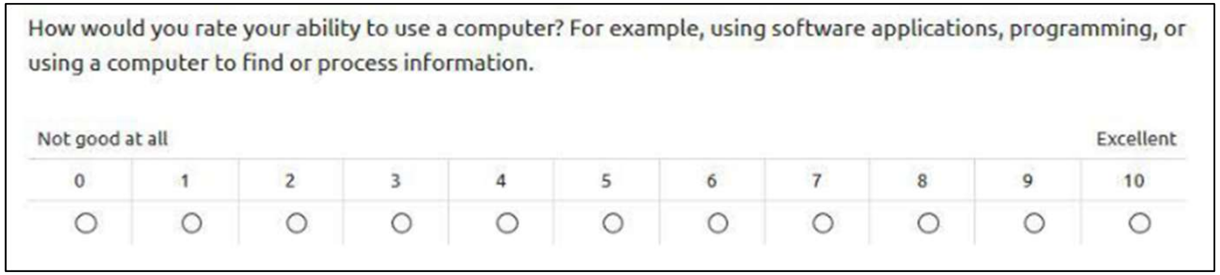
How to measure something?

History of definitions since 1798 [edit]

Definitions of the metre since 1798[189]

Basis of definition	Date	Absolute uncertainty	Relative uncertainty
$^{1}/_{10,000,000}$ part of one half of a meridian, measurement by Delambre and Méchain	1798	0.5–0.1 mm	10^{-4}
First prototype *Mètre des Archives* platinum bar standard	1799	0.05–0.01 mm	10^{-5}
Platinum-iridium bar at melting point of ice (1st CGPM)	1889	0.2–0.1 µm	10^{-7}
Platinum-iridium bar at melting point of ice, atmospheric pressure, supported by two rollers (7th CGPM)	1927	n/a	n/a
1,650,763.73 wavelengths of light from a specified transition in krypton-86 (11th CGPM)	1960	0.01–0.005 µm	10^{-8}
Length of the path travelled by light in a vacuum in $^{1}/_{299,792,458}$ of a second (17th CGPM)	1983	0.1 nm	10^{-10}

Where does measurement error come from? (willing and able)


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.

Not good at all										Excellent
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

- “Insufficient effort responding" (IER) concept of Huang et al. (2012)
    - Time = effort?
- Imperfect self-knowledge or cognitive uncertainty impacts
    - Impact both qualitative survey questions and on incentivized choice tasks (e.g., Jagelka, 2023; Enke and Graeber, 2021; Falk, Neuber, and Strack, 2021)
- Perception of the tasks and their attributes might be changing
    - HL: Additional thoughts about (state/trait) and using the right instrument

How do you deal with measurement error?

1. Id problematic individuals (screening out)
   - Attention check, time, response patterns (e.g. streamlining a survey), answer consistency, or outliers (Meade and Craig (2012), Stantcheva (2022), Read, Wolters, and Berinsky (2022))
   - Theory-based model for low self-knowledge (Falk, Neuber, and Strack (2021))
   - Potentially ask individuals? Mostly focused on effort or attention (Wise and Kong (2005), Meade and Craig (2012), and Alesina, Miano, and Stantcheva (2023).
2. Using fancy econometrics
   - Asking multiple tasks related to a given construct and averaging over the responses ((e.g., Soto and John, 2017; Falk et al., 2022)
   - Obviously related instrumental variables (ORIV) of panel data using given measure as instrument for another (Gillen, Snowberg, and Yariv (2019))
   - Structural modeling of mental processes (e.g., Cunha, Heckman, and Schennach, 2010; Jagelka, 2023; Belzil and Jagelka, 2020)

Data

- Two-wave study online study with 651 respondents
  - Waves varies across 2 to 11 weeks between the two waves
  - English-speaking individuals between 18 and 25 years old from Australia, Canada, the United Kingdom, and the United States
- Selection into second survey wave
  - Main analysis includes 1,400 individuals, 651 of whom completed both survey waves
- Survey question on well-being, personality, and economic preferences, intellectual ability and reliability
  - Panel data to construct quantitative person-specific measures of revealed reliability (test-retest correlation as benchmark to test self-report of reliability)
- Treatment of magnitude of monetary incentives for participation and order of questions

Test (and then retest) “Revealed reliability”

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).

Effects on test-retest

- More items: increases by approximately 50% from 0.56 to 0.82 (1 item to 12 item)
  - Qualitative measure of an economic preference often has only one dedicated survey question
- Time: initial increase but plateaus
  - Threshold of 5th percentile of distribution of survey times
- Other variables: mostly stable across country, time between waves, basic demographic variables, survey conditions, extra incentives (treatment) and question order (treatment)

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

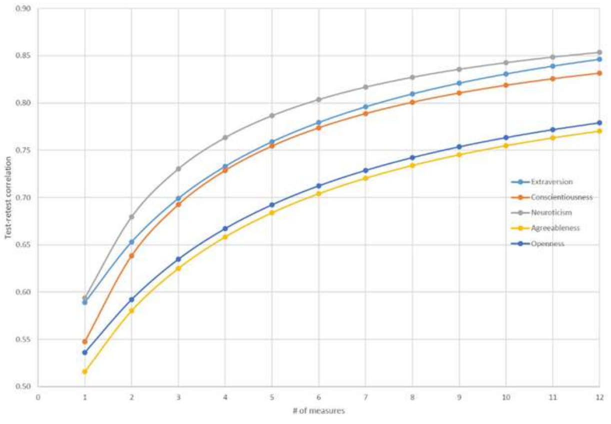
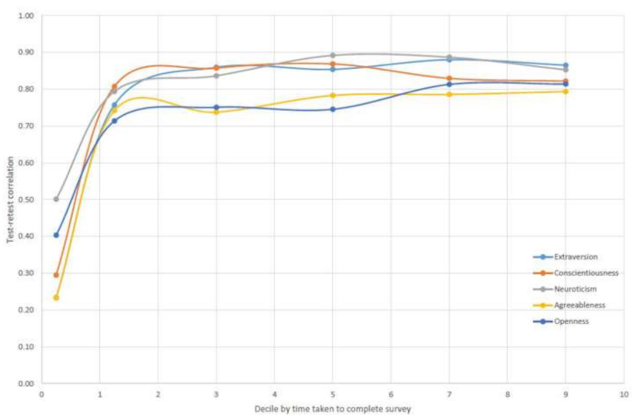


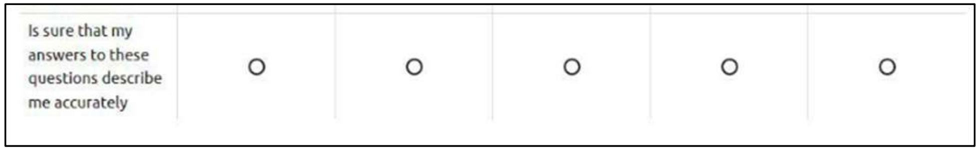
Figure 2: BFI-2 Test-Retest Correlations by Time Decile

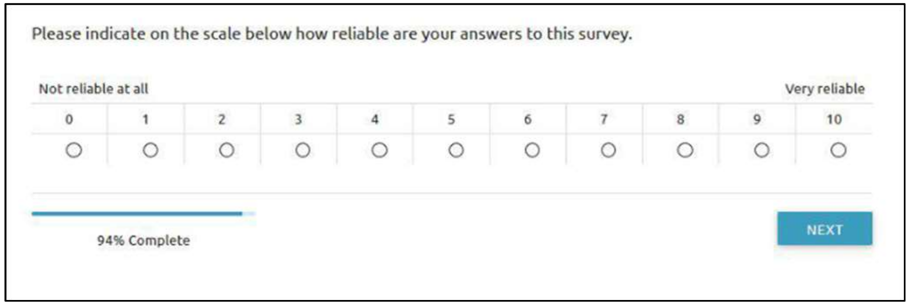


Note: The first data point represents individuals below the 5th percentile in terms of time taken to complete the survey. The remaining data points move by quintile.

The solution: Self report about self reports?

- Two measures: after the end of the BFI-2 questionnaire ("BFI reliability") and at the very end of the survey ("overall reliability")
- 70% reported a reliability of 9 or above on our 11-point Likert scale (average)
- Over 50% indicated "strongly agree" that the answers which they provided describe them accurately





Measurement validity

- Construct validity: individuals' self-reported answer reliability is related to measured answer reliability
  - Same as increasing the number of items (from 1 to 5 or from 3 to 12)
  - BFI reliability has larger impact on revealed individual reliability
  - BFI reliability relevant with inclusion of other controls

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).

Estimating relationships (internal validity)

- Attenuation bias

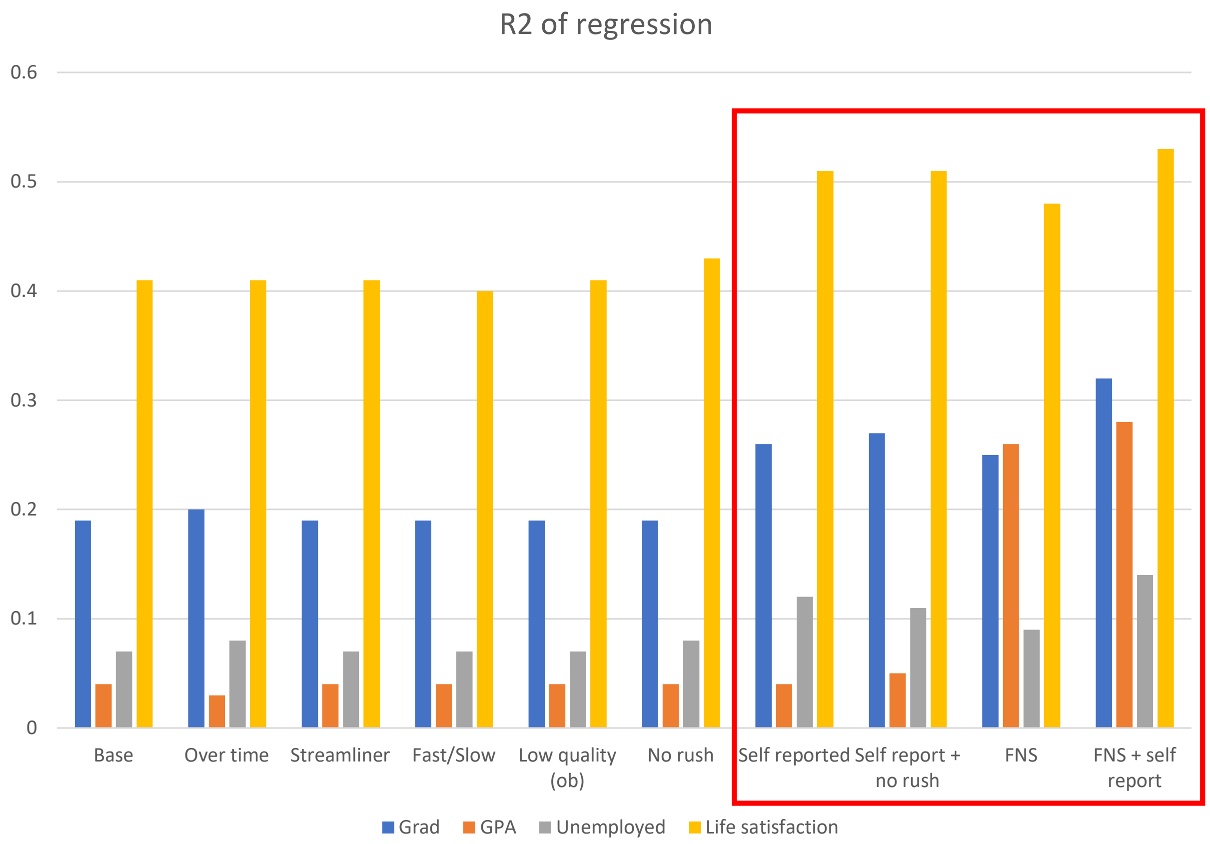
Consider a regression equation:

$y = \beta_0 + \beta_1 x + \epsilon$

If x is measured with error, the observed x would be $x^* = x + u$, where u represents the measurement error. Assuming that u is independent of x and ϵ, and has a mean of zero, the regression of y on x^* instead of x will generally lead to a biased estimate of β_1 (specifically, $\hat{\beta}_1$ will be biased towards zero).

Estimating relationships (internal validity)

- R2 of regressions (includes all controls)
- Statistical significance of estimates:
  1. 70% using ORIV (GSY)
  2. 48% using self report
  3. 42% using self-knowledge (FNS)
- HL thought: Is this p-hacking though?



Note that this is bar chart was recreated by Helena using data from Table 6

Representative? (external validity)

- No evidence that the:
  - reliable sample was substantially different from the unreliable sample
  - relationship of interest was highly non-linear in the dimensions on which the two samples differed

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.

Key takeaways

- Validity of self reports
- Add in the self-reliability question

Is sure that my answers to these questions describe me accurately	○	○	○	○	○

- Add in multiple self-reliability questions
  - Wood, Dustin, et al. "Response speed and response consistency as mutually validating indicators of data quality in online samples." Social Psychological and Personality Science 8.4 (2017): 454-464
  - HL: This paper was 6 years earlier and was not referenced at all

RESEARCH METHODS IN PSYCHOLOGY – 2ND CANADIAN EDITION

Chapter 5: Psychological Measurement

Reliability and Validity of Measurement

Please indicate on the scale below how reliable are your answers to this survey.

Not reliable at all										Very reliable
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

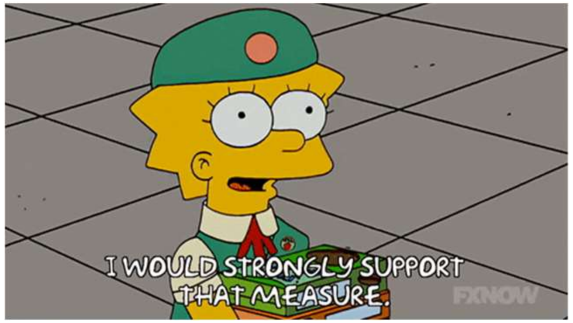
94% Complete

NEXT

Self-reported response quality
My responses to items on this survey are accurate.
I exerted sufficient effort on this survey.
I answered items on this survey without reading them (*R*).
I randomly responded to some survey items (*R*).
I rushed through this survey (*R*).
I thought carefully about each of my responses on this survey.
The researchers should include my data in the results.

In practice?

1. ‘Straight-lining’ (i.e., highly invariant responses to questionnaire items);
2. Implausibly fast responses (i.e., <1sec per questionnaire item);
3. Failing to respond to a simple open-ended question; and
4. (Low) self-reported data quality (e.g., “I rushed through this survey”)
   - Specifically, the self-reported data quality was scored from 1= low quality to 7 = high quality, and we removed anyone who didn’t get at least an average of 4
   - Cronbach’s alpha for the 7 items on that scale is always high (~.80)



Self-reported response quality
My responses to items on this survey are accurate.
I exerted sufficient effort on this survey.
I answered items on this survey without reading them (*R*).
I randomly responded to some survey items (*R*).
I rushed through this survey (*R*).
I thought carefully about each of my responses on this survey.
The researchers should include my data in the results.

Extra slides

Determinants of measurement error: Willing and able

- Willingness to provide highly reliable answers (amount of effort)
  - Hedge bets (high risk aversion and neuroticism and low trust), help researchers (prosocial – high agreeableness, positive reciprocity and low negative reciprocity)
  - Minimal amount of time - effort threshold contain 70% less noise relative to those who are below (comparable to increasing construct from 1 to 12)
- Ability to provide good answers requires good self-knowledge, precise understanding of the tasks at hand and ability to accurately self the most appropriate answer
  - Usually high conscientiousness, openness to experience, and cognitive ability
  - Self knowledge of trait-like features (state vs trait)
  - Self reliability as individual differences trait (reliable at the beginning or end of the survey, across and between waves, likely tied to conscientiousness and cognitive uncertainty correlated across domains of risk and time preferences

Treatments

- Treatment of magnitude of monetary incentives for participation and order of questions
  - Extra incentive treatment condition (extra 3 euros) in wave 1, then subgroup randomize into same (extra 3 euros) or double (extra 6 euros)
  - Order of questions condition: BFI-2 before or after ability section, GPA first or qualitatively evaluate intellectual ability first (same order maintained across waves)

Model to account for these findings

- Individuals first choose whether 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

Two types: reliable one and an unreliable one

- Self-reported reliability is the single best predictor of revealed reliability in our sample and captures measurement error due to both lack of effort and imperfect self-knowledge (40% less noise content relative to those who do not)
- Reduction in measurement error achieved by increasing the number of items per construct from 1 to 5

Comparison to existing methods

- Approach largely outperforms standard screening criteria based on response patterns and time outliers
- Reduction in attenuation bias is similar to that achieved by theory-based alternatives require repeated and/or multiple measurement instruments as well as nontrivial computation
  - Gillen, Snowberg, and Yariv (2019, GSY) and Falk, Neuber, and Strack (2021, FNS)
- Cost effective means to account for individual-specific measurement error

Gillen, Snowberg, and Yariv (2019, GSY)

We use a different approach when estimating causal effects or correlations, drawing inspiration from instrumental variables. Our approach, which we call *Obviously Related Instrumental Variables* (ORIV) uses duplicate elicitations of X and Y as instruments. Specifically, we obtain duplicate measures of X, denoted X^a and X^b, which are both proxies for X^* that are measured with error. By regressing X^a and X^b, we extract the information contained in X^b that can explain X^a. If the measurement error in the two elicitations is orthogonal—as we assume—then the resulting predicted values $\hat{X}^a(X^b)$ contain only information about X^*. We then use a stacked regression to combine the information from both $\hat{X}^a(X^b)$ and $\hat{X}^b(X^a)$, resulting in an efficient use of the data.

This technique is easily extended to allow for multiple measures of the outcome Y. This is particularly useful in estimating correlations, where there is no clear distinction between outcome and explanatory variables, and measurement error in either can attenuate estimates.

ORIV produces consistent coefficients, correlations, and standard errors. This technique is applied, in Section 4, to show that various risk elicitation methods are more correlated than previously thought, and that the patterns of correlations between them are indicative of phenomena outside the lab. We further use ORIV to show, in Section 5, that ambiguity aversion and reaction to compound lotteries are very close to perfectly correlated—once we account for measurement error. This leads us to conclude, in Section 6 that failing to correct for measurement error has lead the field to "over-identify" new phenomena.

Falk, Neuber, and Strack (2021, FNS)

2.2.2 Subjective Self-knowledge

The basic framework assumes that the respondent knows the relative precision τ of her signal x. In other words, she perfectly knows how well she knows herself and weighs her signals accordingly. However, a large body of evidence has shown that individuals often misperceive their own knowledge and skills (Camerer and Lovallo, 1999; Malmendier and Tate, 2005). Applied to our context, respondents may be over-confident and place too much weight on their signal x, or they are under-confident and place too much weight on the prior. In either case, this will result in a wedge between the optimal and the actual response, again potentially complicating inference about respondents' true types.

[4]The former could reflect, e.g., monetary or social approval incentives, while the latter may capture motives such as a desire to respond truthfully and accurately or simply an interest in (promoting) research.

To model potential biases in perceived self-knowledge, we introduce subjective self-knowledge $\tilde{\tau}$. A respondent has correct beliefs about her self-knowledge if $\tilde{\tau} = \tau$, she is under-confident if $\tilde{\tau} < \tau$, and she is over-confident if $\tilde{\tau} > \tau$. We assume that the agent is naive and that when determining her survey response, she applies relative weights according to her subjective self-knowledge $\tilde{\tau}$. Equation (2) changes as follows:

$$r = \frac{\bar{\theta} + \tilde{\tau}\, x}{1 + \tilde{\tau}}$$

Corresponding to Equation (4), the between-variance becomes

$$\sigma^2_{\text{between}} = \text{var}(\mathbb{E}[R \mid \theta]) = \left(\frac{\tilde{\tau}}{1+\tilde{\tau}}\right)^2 \sigma^2 .$$

Hence, the variability in answers between different items reflects the respondent's subjective self-knowledge but is independent of self-knowledge itself. Intuitively, as the between-variance is based only on the expected response, which is independent of the true precision of the agent's signal τ, the variance is also independent of the true precision of the agent's signal.

This is different for the within-variance, corresponding to Equation (5).

$$\sigma^2_{\text{within}} = \text{var}(r \mid \theta) = \left(\frac{\tilde{\tau}}{1+\tilde{\tau}}\right)^2 \frac{\sigma^2}{\tau} .$$

Limited Self-Knowledge and Survey Response Behavior

Abstract

We study response behavior in surveys and show how the explanatory power of self-reports can be improved. First, we develop a choice model of survey response behavior under the assumption that the respondent has imperfect self-knowledge about her individual characteristics. In panel data, the model predicts that the variance in responses for different characteristics increases in self-knowledge and that the variance for a given characteristic over time is non-monotonic in self-knowledge. Importantly, the ratio of these variances identifies an individual's level of self-knowledge, i.e. the latter can be inferred from observed response patterns. Second, we develop a consistent and unbiased estimator for self-knowledge based on the model. Third, we run an experiment to test the model's main predictions in a context where the researcher knows the true underlying characteristics. The data confirm the model's predictions as well as the estimator's validity. Finally, we turn to a large panel data set, estimate individual levels of self-knowledge, and show that accounting for differences in self-knowledge significantly increases the explanatory power of regression models. Using a median split in self-knowledge and regressing risky behaviors on self-reported risk attitudes, we find that the R2 can be multiple times larger for above- than below-median subjects. Similarly, gender differences in risk attitudes are considerably larger when restricting samples to subjects with high self-knowledge. These examples illustrate how using the estimator may improve inference from survey data.

3 Estimator

In this section, we derive an estimator for an individual's level of self-knowledge that is based on the insights from Section 2. We consider a panel data set comprising $I > 1$ agents and $T > 1$ waves. In each wave t, each agent i answers an identical set of $K > 1$ questions about distinct, time-invariant characteristics, traits, or beliefs. We denote by θ_{ik} the value of the k^{th} characteristic for agent i and assume that characteristics are independently normally distributed in the population with mean $\bar{\theta}$ and variance σ^2. In contemplating the answer to question k in wave t, agent i generates a signal x_{ikt} that she uses to form her answer r_{ikt}. The signal x_{ikt} is normally distributed with mean θ_i and variance σ^2/τ_i, independent of all other signals, such that the optimal response is given by

$$r_{ikt} = \frac{\bar{\theta} + \tau_i\, x_{ikt}}{1 + \tau_i} .$$

Given the $K \times T$ answers observed for each agent i, the objective of a researcher is to estimate agents' levels of self-knowledge τ_i. In Section 2, we have shown that τ equals the (the-

4 Experimental Evidence

5 Applications

In this section, we apply our estimator to data from the German Socio-economic Panel (SOEP), a large, representative panel data set. The main objective is to show that by using estimates of self-knowledge, $\hat{\tau}$, we can increase the explanatory power of regressions that involve self-reports. In particular, we estimate τ using answers to the Big Five personality inventory from multiple waves and split the relevant samples by the respective