Limited Self-knowledge and Survey Response Behavior

Armin Falk Luca Henkel Thomas Neuber Philipp Strack

June 15, 2025

Abstract

We study response behavior in surveys and propose a method to identify and improve the informativeness of survey evidence. First, we develop a choice model of survey response behavior under the assumption that responses imperfectly reveal respondents' characteristics due to limited self-knowledge, inattention, or lack of engagement. Respondents receive individual-specific signals about their characteristics and choose their responses accordingly. We identify the conditions under which this process leads to biased inference from survey evidence and demonstrate how focusing on respondents with high signal precision mitigates bias. Importantly, we show that a respondent's signal precision can be inferred from observed response patterns. Second, based on these insights, we develop a consistent and unbiased estimator for a respondent's signal precision. Third, we provide experimental and survey evidence concerning the performance of the model and estimator. We experimentally test the model's key predictions in a context where the researcher knows the true characteristics. The data confirm both the model's predictions and the estimator's validity. Using a large survey, we show how our estimator can be used to improve survey evidence. Our estimator significantly increases the explanatory power of self-assessments and their association with behavior, and performs well relative to alternative methods proposed in the literature.

Keywords: survey research, rational inattention, online experiment, non-cognitive skills, preferences
JEL Classification: C83, D83, C91, D91, J24

Contact: Armin Falk: University of Bonn; afalk@uni-bonn.de. Luca Henkel: Erasmus University Rotterdam; henkel@ese.eur.nl. Thomas Neuber: unaffiliated; thomas.e.neuber@gmail.com. Philipp Strack: Yale University; philipp.strack@yale.edu.

Acknowledgements: We are grateful to Roland Bénabou, Philipp Eisenhauer, Botond Kőszegi, John A. List, and participants at various conferences and seminars for helpful comments. We thank Markus Antony for excellent administrative support as well as Egshiglen Batbayar, Bruno Pelaccini, Uranchimeg Zamindii and in particular Juan Cruz Duhalde for excellent research assistance.
Funding: Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 – 390838866. Philipp Strack was supported by a Sloan fellowship.
Ethics approval: The study obtained ethics approval from the German Association for Experimental Economic Research (#rgKEIvFE)
Research transparency: The studies in this paper were preregistered at aspredicted.org (#140358 and #187998). See also Appendix K on research transparency.

1 Introduction

Survey evidence is a major source of knowledge in the social sciences, including economics. With growing interest in measuring cognitive and non-cognitive skills—such as economic preferences, beliefs, attitudes, and values—survey evidence is gaining increasing relevance (Almlund et al., 2011; Falk et al., 2018; Heckman, Stixrud, and Urzua, 2006; Stantcheva, 2023). Despite the growing use of surveys, there are concerns about the reliability and informativeness of survey responses given their hypothetical or low-stakes nature. Similarly, factors such as respondents' limited attention to survey questions, subjective interpretation of response scales, and lack of effort in providing answers threaten the quality of survey data.

This paper provides a method to empirically identify the informativeness of survey responses. The method is derived from a simple model of information acquisition and the resulting survey response behavior that allows identifying more vs. less informative respondents based only on *patterns* of their response behavior. In particular, we make two main contributions: First, we offer a comprehensive framework for modeling survey response behavior, which can be used to understand challenges to the validity of survey evidence such as limited attention or effort, subjective scale-use or social desirability bias. Second, we develop an easy-to-implement method to assess the informativeness of responses and provide empirical evidence that the method reduces biases and improves the explanatory power of survey measures.

Model. As a first step, we derive a choice model of survey response behavior. In the model, we take seriously the idea that providing survey responses is a *choice*: when being asked to report an individual characteristic such as a preference, belief, or some non-cognitive skill, a respondent has to make herself the object of her own self-assessment when choosing a response. We assume that there exists a true type (level of each characteristic) but that the respondent is not perfectly aware of her true type. This *limited self-knowledge* is modeled as an imperfect signal that the respondent receives about her true type. Generally, the signal's imprecision may result from costly information acquisition as commonly assumed in rational inattention models (e.g., Sims, 2003). More specifically, differences in self-knowledge may arise from the fact that individuals vary in their capacity to retrieve or memorize relevant information about themselves, engage more or less in reflecting who they are, or that some people simply lack life experience in the domain of interest. We further assume that the respondent wants to minimize the squared distance between her true type and report, i.e., the interests of the respondent and the researcher are aligned. Conditional on the signal's informativeness, our agent's Bayesian optimal report is a weighted sum of the population mean of the respective characteristic and her signal. The more informative the signal, the greater the weight placed on the signal relative to the population mean.

Using this setup, we analyze the expected variance of respondents' answering behavior conditional on the informativeness of the signal, both for repeated observations of a given characteristic as well as between different characteristics. We find that the variance *between* characteristics increases in the informativeness of the signal, which mirrors the fact that the more confident a respondent is about her answer, the more she deviates in expectation from the population mean. In contrast, the *within* variance—the variance of responses for a given characteristic over time—is non-monotonic in the signal precision. The intuition is that response behavior is stable over time if an individual knows herself either very well or not at all. This result cautions against the use of simple stability to measure the accuracy of signals and reports. Importantly, we show that the ratio of the variance between characteristics and the variance over time (for given characteristics) is equal to the informativeness of the signal. This key result implies that we can use observed variances to estimate individual differences in self-knowledge and the reliability of the respective reports.

We provide several extensions of the model and discuss their implications for expected response behavior. Our first extension relaxes the assumption that respondents are perfectly aware of the signal strength, i.e., how well they know themselves. Instead, we allow for subjective levels of self-knowledge that are higher or lower than actual self-knowledge. While subjective beliefs about self-knowledge affect the distribution of responses, we show that they do not impede the identification of differences in self-knowledge, simply because they cancel out. Second, we relax the assumption that the only objective of the respondent is to minimize the distance between true type and report. Specifically, we study social desirability and subjective scale use as potential strategic motives that lead respondents to distort their reports. We show that for specific parameterizations of social desirability and subjective scale use, identification of self-knowledge remains unchanged. Third, we explore the case in which responses are not only affected by limited self-knowledge but also by random errors in the form of a normal noise term. We show that such an error implies that the ratio of the variances underestimates the informativeness of the signal, i.e., it will be a conservative estimate.

We then turn to the consequences of imperfect signal precision or lack of self-knowledge for survey-based inference. We focus on a situation in which an analyst seeks to learn about the relationship between an outcome (e.g., investment behavior) and a characteristic measured by a survey item (e.g., self-assessed risk aversion) using linear regressions. We show that recovering the unbiased regression estimator represents a knife-edge case: it requires all agents to have perfect knowledge of their own signal precision or degree of self-knowledge. As soon as some agents have an imperfect assessment of their precision (e.g., if they are overconfident), linear regressions become biased, and we derive the exact condition determining the direction and extent of the bias. Importantly, we show that focusing on agents with high levels of self-knowledge reduces the bias, which vanishes in the limit. Accordingly, if an analyst were to observe the distribution of self-knowledge in a population, *subsetting on self-knowledge* can be used as a diagnostic tool to assess the extent and direction of the bias induced by limited self-knowledge. Moreover, we show that subsetting also increases the explanatory power of a regression as measured in terms of R^2. These insights motivate. These insights motivate the development of an estimator of self-knowledge, which we undertake in a second step.

Estimator. To derive an estimator of signal precision—or self-knowledge—from panel data, we consider the ratio between two sample variances: the between-variance (the variance of responses between items) and the within-variance (the variance for a given item over time). These are the sample analogs to our theoretically derived variances. We study the asymptotic properties of the estimator and formally show its consistency as well as unbiasedness. Using simulations, we illustrate the performance of the estimator for realistic sample sizes. We study various combinations of the number of respondents, survey items, and waves. The estimator generally performs well.

In the third part of the paper, we provide results from an experiment designed to test the main predictions of the model and to evaluate the estimator's performance. Subsequently, we analyze survey data to show how accounting for signal precision improves survey evidence and how our method compares to other methods proposed by the literature.

Experiment. To empirically test the main predictions of the model and our estimator's performance, it is crucial to (i) observe responses and compare them with respondents' *true* types and (ii) exogenously vary self-knowledge. However, this is difficult—if not impossible—with typical survey data. Therefore, we ran an experiment that created a panel data set with types that are imperfectly known to subjects but perfectly known to the researcher. Specifically, we use a classic psychometric dot estimation task: Subjects viewed 60 images, each on a separate screen, displaying varying numbers of dots. They were paid to accurately report the number of dots. This setup allows us to observe subjects' reports of an objective true type. Between subjects, we exogenously varied the time during which dots were displayed. In the Long-treatment dots were displayed for 7.5 seconds and with high visibility, while in the Short-treatment dots were displayed only for 0.5 seconds with low visibility. Results from the experiment confirm our main predictions. First, subjects' reports are linear in true types and biased towards the population average, i.e., images displaying below-mean numbers of dots are, on average, overestimated, and images with above-mean numbers are underestimated. Second, the bias is stronger in the Short compared to the Long treatment, showing that lower self-knowledge biases responses more strongly toward the population average. Turning to our estimator, we find that the estimator reliably detects the exogenous variation in self-knowledge between the two treatments. Importantly, we show that subjects to whom our estimator assigns high values of self-knowledge indeed have less bias towards the average and their responses are more predictive of their true types. Lastly, we demonstrate how a lack of self-knowledge biases regression estimates and how restricting the sample based on our estimator reduces this bias, as predicted by the model. This provides evidence for the usefulness of our estimator as a diagnostic tool to detect and assess biases introduced by limited self-knowledge.

Survey evidence. Finally, we apply our estimator to a large survey to investigate its effectiveness in improving survey evidence. To estimate self-knowledge, we suggest a simple survey module that can be used by any researcher interested in assessing respondents' self-knowledge. The module is a standard fifteen-item Big Five personality survey, which is repeated one time using a slightly rephrased version. Our application to measure the performance of our estimator is the relationship between self-assessments and behavior. We focus on two domains, risk attitudes and social preferences, and measure self-assessed risk attitudes and altruism using the general risk (Dohmen et al., 2011) and general altruism question (Falk et al., 2018), respectively. To measure behavior, subjects face an incentivized lottery choice and a dictator game. They also report a set of risk and altruistic behaviors, such as whether they own stocks or donated to charity in the past, providing us with ten measures of behavior in total. To quantify performance, we focus on three criteria: (i) the strength of the association between self-assessment and behavior, measured by OLS regression coefficients, (ii) the explanatory power of self-assessments for behavior, measured in terms of a regression's R^2, and (iii) the test-retest stability of self-assessments. These three criteria reflect different dimensions of survey quality.

We find that our estimator of self-knowledge significantly improves survey evidence across all three criteria. Focusing on subjects with high estimated levels of self-knowledge consistently increases the coefficients of self-assessments. For instance, relative to baseline, the OLS coefficient of self-assessed risk on lottery choices almost doubles when estimated among above-median self-knowledge subjects. Similarly, explained variance in terms of R^2 is significantly higher in regressions considering respondents with above-median levels of self-knowledge. To illustrate, in the case of altruism, the R^2 increases from 0.13 in baseline to 0.36 in the high self-knowledge sample. Moreover, subjects with an estimated high level of self-knowledge also display substantially higher test-retest correlations in self-assessments. For example, for self-reported risk preferences using the general risk question, the top 10% of subjects show a test-retest correlation of 0.94. Importantly, the improvements in coefficients, R^2, and test-retest correlations are fairly monotonic in estimated self-knowledge: the higher the level of self-knowledge in a subsample, the higher the coefficients, levels of R^2, and correlations, respectively.

To put the improvements of our self-knowledge estimator into perspective, we also study a comprehensive set of alternative methods that have been suggested to improve survey evidence. Among them are attention checks, effort measures, response time, and averaging, as well as instrumental variable strategies. We find that our estimator performs well compared to these alternative methods on both criteria. The OLS coefficients obtained from focusing on high levels of self-knowledge (e.g., above-median and top 20%) are among the highest across all methods. In fact, our estimator is the only method that consistently leads to improvements in all ten associations of self-assessment with behavior. Other methods, such as excluding subjects based on attention checks or response times, sometimes lead to stronger, and sometimes weaker associations, and the results are sensitive to the specific exclusion criteria used. Similarly, the improvements in R^2 when regressing behavior on self-assessments, and the test-retest correlations of self-assessments are among the highest across all methods when focusing on subjects with above-median self-knowledge and the highest when focusing on the top 20%. These results provide further evidence that our self-knowledge estimator offers a reliable and effective method to assess bias and to improve survey evidence.

Related literature. Our paper is related to multiple strands of the literature. As we take the informational constraints of the agent seriously and study their choice implications, we relate to the work on rational inattention (Caplin et al., 2020; Caplin and Dean, 2015; Matějka and McKay, 2015; Sims, 1998, 2003). This literature focuses on flexible information acquisition and studies what type of information is acquired in a single-agent setting. Our goal is different, and we analyze how to identify agents' levels of information in a situation with many agents who share a common prior. Our framework enables analyzing the provision of incentives in surveys as studied, for example, in Prelec (2004) and Cvitanić et al. (2019) as well as how contextual factors such as image or social desirability affect responses (see, e.g., Bénabou et al., 2023; Chen et al., 2020). The notion of limited self-knowledge and its economic consequences for the labor market has been studied in Falk, Huffman, and Sunde (2006a,b). The model is also related to work on preferences for consistency, as modeled and tested in Falk and Zimmermann (2017) and applied to survey methodology in Falk and Zimmermann (2013).

Moreover, the paper contributes to the literature on measurement error in surveys (for an overview, see Bound, Brown, and Mathiowetz, 2001). For the case of classical measurement error—where deviations in answers are independent of the respective true value— instrumental variables techniques are capable of removing bias. More recently, Gillen, Snowberg, and Yariv (2019) have suggested to measure duplicate instances and to use them as mutual instruments. Hyslop and Imbens (2001) consider a model that is related to ours where an agent observes a Normal signal and reports their best estimate of an underlying variable of interest. They analyze the effect of the resulting non-classical measurement error on regression coefficients but do not consider remedies. The focus of our paper is to estimate the precision of the agent's signal, which allows placing higher weight on subjects with better self-knowledge.

Drerup, Enke, and Gaudecker (2017) estimate a structural model of stock market participation that identifies individuals for whom relevant preferences and beliefs have increased explanatory power. Alternative approaches to deal with measurement error in subjective survey data use structural estimation techniques to recover underlying primitives and choice models, finding that accounting for measurement error yields greater predictive power (Beauchamp, Cesarini, and Johannesson, 2017; Kimball, Sahm, and Shapiro, 2008).[1] Another strand of the literature uses separate items to capture measures of quality, such as attention (Berinsky et al., 2021), reliability (Dohmen and Jagelka, 2023), effort (Meade and Craig, 2012), or response times (Curran, 2016). We add to this literature a systematic empirical evaluation of different methods' effectiveness in increasing the explanatory power of survey items for behavior. A related contribution comes from Beauchamp et al. (2020), who analyze incentivized behavior in experiments rather than self-reports in surveys. They argue that accounting for the "compromise effect" —whereby subjects' answers tend towards the center of the provided scale—, can improve estimates of risk preferences.

The remainder of the paper proceeds as follows. Section 2 develops the model with its basic framework, extensions, and consequences for inference. Building upon its insights, Section 3.1 introduces the estimator, presents its theoretical properties, and explores its performance in finite samples. Section 4 presents the stylized experiment. In Section 5, we apply the estimator to a survey and compare its performance to other commonly used methods to improve survey response behavior. Section 6 concludes.

2 Model

In this section, we first introduce a simple framework to model response behavior in surveys, based on limited self-knowledge. Second, we derive how patterns in answering behavior reveal the informational content of responses, providing the intuition for how we later estimate self-knowledge. Third, we present various extensions of the baseline model to study further important aspects of the answering process. Finally, we show how the presence of limited self-knowledge influences inference from survey responses.

Introspection and Self-knowledge. The context that we are interested in is a simple survey situation. A researcher asks a respondent (or agent) a question about a specific characteristic, e.g., some preference, personality trait, or belief.[2] The agent's true type is denoted by θ, and we assume that it is normally distributed in the population with mean $\bar{\theta}$ and variance σ^2. Agents act upon their true types but vary with respect to how well they know their type. Hence, when asked about her type

[1] In the psychology literature, processes that underlie response behavior have been studied under the label of *cognitive aspects of survey methodology* (see Bradburn, Sudman, and Wansink, 2004; Schwarz, 2007; Sudman, Bradburn, and Schwarz, 1996). Broadly, our paper is also related to classical test theory and item response theory (see, e.g., Bolsinova, de Boeck, and Tijmstra, 2017; Edwards, 2009; Kyllonen and Zu, 2016).

[2] For example, the researcher may ask the respondent to state her willingness to take risks, her level of agreeableness or conscientiousness, or her belief about her internal or external locus of control.

θ, the respondent does not perfectly know herself but instead engages in a process of introspection. The outcome of this process is an informative but noisy signal x about her true type. The signal is normally distributed with a mean equal to the agent's type θ and variance σ^2/τ. The parameter $\tau > 0$ hence indicates the precision of the signal relative to the variance in the population. The higher the value of τ, the more precise is the signal that an individual receives about herself. We refer to τ as *self-knowledge*.

Response Behavior. After reflecting on her true type θ, the respondent reports her answer. We assume that she seeks to provide a response r that is as precise as possible, i.e., the interests of the researcher and respondent are aligned.[3] Formally, the respondent uses her signal x to provide a response r that minimizes the expected quadratic distance to her unknown true type, i.e.,

$$u_\theta(r) = -\left(r-\theta\right)^2 \ . \tag{1}$$

This objective ensures that the respondent reports her best guess of her type $r = \mathbb{E}[\theta \,|\, x]$. The respondent's prior equals the distribution of types in the population with mean $\bar{\theta}$. Substituting for the expected value of her posterior belief about her type, we obtain by Bayes' Rule that

$$r = \frac{\bar{\theta} + \tau\, x}{1+\tau} \, . \tag{2}$$

Intuitively, the higher her self-knowledge τ, the more precise the respondent's signal, and the more weight she puts on her signal relative to the population mean $\bar{\theta}$. In the limit, if she knows nothing about herself, her best estimate is to report the mean of her prior, whereas if she knows herself perfectly, she disregards the prior completely. This concludes our basic framework. The model defines a mapping from true types to distributions over observable responses, taking into account the notion of limited self-knowledge.

2.1 Response Patterns

We now explore the implications of limited self-knowledge for response patterns. We are particularly interested in the variances in reports, both unconditional and conditional on an agent's type. These variances will allow us to identify differences in self-knowledge. In Section 3.1, we will build on these insights when we derive an estimator for an individual's level of self-knowledge in panel data.

Expected Report. It follows from Equation (2) that the expected report conditional on the true type θ equals

$$\mathbb{E}[r \,|\, \theta] = \frac{\bar{\theta} + \tau\, \theta}{1+\tau} \, . \tag{3}$$

For low values of self-knowledge τ, the expected report is close to the population mean $\bar{\theta}$, irrespective of the true type θ. For large values of τ, the expected report converges to the true type θ.

Between-variance. Consider now the variance of conditional expected reports. In the context of panel data, one can think of this theoretical quantity as an approximation of the variance in aver-

[3] For many interview situations, we think that this is a valid assumption. However, there are contexts in which respondents may want to strategically signal a specific type that is actually different from their belief about their true type for reputational or "social desirability" reasons. In Section 2.3, we provide extensions that relax the assumption.

age reports concerning different characteristics. Following this interpretation, we refer to it as the *between-variance*. It is given by

$$\begin{aligned}\sigma^2_{\text{between}} &:= \operatorname{var}(\mathbb{E}[r \mid \theta]) = \operatorname{var}\left(\frac{\bar{\theta} + \tau\,\theta}{1+\tau}\right) \\ &= \left(\frac{\tau}{1+\tau}\right)^2 \operatorname{var}(\theta) = \left(\frac{\tau}{1+\tau}\right)^2 \sigma^2 \,.\end{aligned} \tag{4}$$

The between-variance is strictly increasing in self-knowledge τ. This reflects the fact that agents with high levels of self-knowledge put relatively little weight on their prior. Instead, they provide reports that tend to deviate from the population mean.

Within-variance. Now consider the variance conditional on an agent's type. This theoretical quantity can be thought of as the variation in responses of an agent responding multiple times to questions about the *same* characteristic. We call this variation the *within-variance* of the agent's reports. It is given by

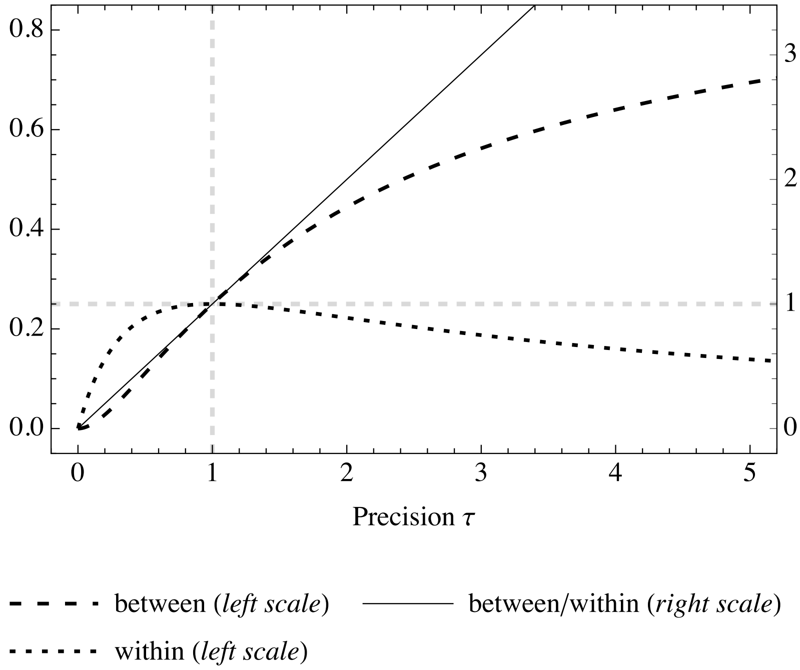
$$\sigma^2_{\text{within}} := \operatorname{var}(r \mid \theta) = \operatorname{var}\left(\frac{\bar{\theta} + \tau\,x}{1+\tau} \,\middle|\, \theta\right) = \left(\frac{\tau}{1+\tau}\right)^2 \operatorname{var}(x \mid \theta) = \frac{\tau}{(1+\tau)^2}\,\sigma^2 \,. \tag{5}$$

The relationship between self-knowledge τ and the within-variance is non-monotonic. For very low levels of τ, the variance is low, simply because the respondent refers to her prior. As τ increases, the variance increases as more weight is placed on the noisy signal. However, as τ further increases, the variance decreases because the signal about the true type becomes increasingly precise. From a researcher's perspective, this pattern implies that consistent responses—i.e., similar responses regarding the same characteristics over time—do not necessarily indicate high levels of self-knowledge and precision. The most stable responses come from respondents who know themselves perfectly—or who do not know themselves at all.

Figure 1 illustrates the relationship between the two variances and self-knowledge. It plots the between-variance (long dashes) and the within-variance (short dashes) as functions of self-knowledge τ. As τ goes to zero, both variances converge to zero. This means that the respondent provides the same answer (equal to the prior) to any question. As τ increases, the respondent places higher weight on her signal, which increases both the within- and between-variance. At $\tau = 1$, i.e., when the signal x is exactly as informative as the respondent's prior knowledge about the population, the within-variance reaches its maximum and is equal to the between-variance. Beyond this point, the between-variance further increases and ultimately converges to the variance of true types in the population, σ^2. At the same time, the within-variance strictly decreases and converges to zero, because a respondent with perfect self-knowledge will always provide exactly the same report for a given characteristic.

Both the between- and within-variance contain information about the respondent's level of self-knowledge τ. While a large between-variance is always "good news," indicating high levels of τ, a low within-variance can reflect either high or low levels of τ, respectively. However, considering both variances *jointly* perfectly reveals the level of self-knowledge. In fact, the ratio of the between- and

Figure 1: Theoretical variances



Note: Variances $\sigma^2_{\text{between}}$ and σ^2_{within} as functions of τ (values on the left axis). The solid line shows the ratio of the two variances, which is equal to τ (values on the right axis).

within-variance equals the degree of self-knowledge:

$$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{\left(\frac{\tau}{1+\tau}\right)^2 \sigma^2}{\frac{\tau}{(1+\tau)^2}\,\sigma^2} = \tau\,. \tag{6}$$

The respective relationship is also shown in Figure 1 where, for each level of τ, the thin solid line plots the ratio of the two variances.

Our paper builds on this insight. We show that the relationship between the variances and self-knowledge is robust to various extensions of the model, construct a finite sample estimator based on this relationship, and show that this estimator indeed predicts the informativeness of subjects' responses in experimental and survey data.

2.2 Response Patterns under Subjective Self-knowledge

Our framework so far assumed that the respondent knows the relative precision τ of her signal x. In other words, while she has imperfect knowledge about her characteristics, she has perfect knowledge about how well she knows herself and weighs her signals accordingly. In practice, this assumption may be restrictive, as imperfect knowledge about one's characteristics may coincide with imperfect knowledge about one's precision. Indeed, 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 may be 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.

Reporting behavior. 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:

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

Between-variance. 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.

Within-variance. The impact on the within-variance is different, as corresponding to Equation (5), the within-variance becomes:

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

The latter depends on both subjective self-knowledge as well as actual self-knowledge. Intuitively, the within-variance of responses is affected by the respondent's subjective self-knowledge $\tilde{\tau}$ through the weight that she places on her signal and by her self-knowledge τ through the variance of the signal.[4]

Ratio of variances. Importantly, the result from Equation (6) about the ratio of the two variances still holds.

$$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{\left(\frac{\tilde{\tau}}{1+\tilde{\tau}}\right)^2 \sigma^2}{\left(\frac{\tilde{\tau}}{1+\tilde{\tau}}\right)^2 \frac{\sigma^2}{\tau}} = \tau$$

Hence, while deviations from correct beliefs about the precision of one's signals affect expected response behavior in general, inference about τ remains feasible. This also means that the estimator we later develop in Section 3.1 will recover τ irrespective of subjective self-knowledge. However, the presence or absence of subjective self-knowledge greatly matters for inference from survey evidence, as we show in Section 2.4.

2.3 Relation to Rational (In)-Attention and Model Extensions

Rational (In)-Attention. Our model is equivalent to a rational inattention model where respondents choose the precision of their information and have heterogeneous information costs: Consider

[4] Observe that only for $\tilde{\tau} \to \infty$, the model predicts classical measurement error.

an agent who chooses how much effort to invest in introspection, determining τ at a cost $1/a\, c(\tau)$, where $a > 0$ captures the agent's ability. Her utility function equals $u_\theta(r, \tau) = -m\,(r-\theta)^2 - \frac{1}{a}c(\tau)$. Here, $m > 0$ measures the motivation to answer accurately, arising from intrinsic or extrinsic incentives.[5] We have the following observation:

Observation 1. *The agent answers as if their precision was exogenously fixed at the level* $\tau^* = \arg\max_{\tau \geq 0} am\frac{\sigma^2}{\tau+1} - c(\tau)$.

Thus, agents' responses in a rational inattention model where agents differ in their abilities a will be exactly as in our baseline model where the precisions τ are exogenously given. The key difference to our baseline model is that the rational inattention model predicts that agents will react to a change in incentives (captured here by a change in m) by adjusting the precision of their signals.

As before, differences in response quality can be inferred from observable patterns, but now these differences reflect variation in effort and ability. Higher incentives lead to more informative responses, aligning with practices like financial rewards in experiments (Camerer and Hogarth, 1999; Smith, 1976) or truth-telling mechanisms like the Bayesian Truth Serum (Prelec, 2004).

Social Desirability and Subjective Scale Use. One of our model's central assumptions is that respondents do not have a strategic motive to bias their responses in a particular direction. This assumption is the driving force behind our result that the ratio of between and within variance exactly equals τ. For specific survey items or environments, however, strategic motives that lead to biased responses may be present, e.g., in the form of social desirability effects. In this case, respondents have a preference to provide an answer that is deemed socially desirable, e.g., due to identity or image concerns. Another relevant case for systematically biased responses is subjective scale use. If agents interpret scales differently, they may provide different actual reports, although their *intended* reports are identical.

In Appendix B, we show how our baseline model can be extended to integrate these additional aspects of the survey response process. We show that for specific parameterizations of social desirability and subjective scale use, identification of τ as the ratio of variances is possible even though respondents have strategic motives to systematically bias their responses. In Section 5.4.2, we also provide a survey module that can be used to assess and correct for subjective scale use.

Trembling Hand Errors. Instead of assuming directional errors, another way to relax our central assumption is to assume that respondents make random errors when trying to provide a response. We model this as a noise component that is added to a respondent's intended response. That is, the observed response becomes $\tilde{r} = r + \epsilon_r$, with ϵ_r as an independent Normal shock ϵ_r. As we show in Appendix B.3, this extension will lead the ratio of between- and within-variance to underestimate the true level of τ.

2.4 The Implications of Limited Self-Knowledge for Estimating Regression Models

We have developed a framework to characterize how limited self-knowledge affects survey response behavior. We now apply this framework to study consequences for estimating regression models in a context where a particular outcome y (such as income or education) is regressed on an individual

[5] For instance, monetary incentives, social approval, or a desire to be truthful.

characteristic which is measured in terms of a survey response r (such as self-reported willingness to take risks). Does limited self-knowledge bias estimates in such a context, and can we use our framework to correct potential biases in order to improve inference from survey measures?

To address these questions, we focus on an analyst who estimates a regression model using survey response data consisting of pairs of reported characteristics r and the outcomes y. Participants in the survey differ in their unobserved characteristic θ, level of self-knowledge τ, and subjective self-knowledge $\tilde{\tau}$. Following our model, the reported characteristics are influenced by individual specific levels of subjective self-knowledge (see Section 2.2), i.e.,

$$r = \frac{\bar{\theta} + \tilde{\tau}\, x}{1 + \tilde{\tau}}\,, \tag{7}$$

where $x \sim \mathcal{N}(\theta, \sigma^2/\tau)$ is a signal the agent privately observes about their characteristic. Throughout, we assume that the characteristic θ is independent of the agent's level of objective and subjective self-knowledge.

The analyst is interested in understanding the relation between the outcome variable y and the characteristic θ. The characteristic θ affects outcome y through the linear relation

$$y = \beta_0 + \beta_1 \theta + \epsilon\,.$$

The analyst's goal is to learn $\beta = (\beta_0, \beta_1)$. For example, the analyst might observe income (corresponding to outcome y) and is interested in the relation to risk-aversion (corresponding to θ), but only observes a self-reported measure of risk aversion (corresponding to r). Hence, while y is observed by the analyst, she does not observe θ. She only observes the agent's report r, which is related to the agent's type through the signal x the agent privately observes. This poses a problem for the analyst as the agent's report is only a noisy signal of the true characteristic, which potentially biases any inference about the relationship between characteristics and outcome variables.

Note that here we study implications for regression estimates where limited self-knowledge affects the *independent variable*. In the Appendix Section H, we also study the implications of limited self-knowledge in a regression context where limited self-knowledge affects the *dependent* variable. This is the case, e.g., if the analyst is interested in studying the effect of gender on a measure of risk aversion.

2.4.1 Classical Regression Estimates

The common way of estimating the relation between θ and y is to ignore the fact that reports are only an imperfect signal about types and run a linear regression on a dataset $(r_i, y_i)_i$ of reports r_i and outcomes y_i for different subjects $i \in \{1, \ldots, N\}$ to obtain an estimate of β. The classical OLS-regression estimate is then given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (y_i - \bar{y})(r_i - \bar{r})}{\sum_{i=1}^N (r_i - \bar{r})^2} \qquad\qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{r}\,. \tag{8}$$

Because θ is not directly observed, we are in a situation with errors in variables, which potentially biases the estimates (Dougherty, 2016). Furthermore, due to the different degrees of self-knowledge we face heteroskedastic errors. Perhaps surprisingly, the regression estimate defined (8) is nevertheless a consistent estimator if

1. the degree of self-knowledge τ and the type θ are independent,
2. and every agent i estimates their self-knowledge correctly $\tau = \tilde{\tau}$.

Define $\gamma = \tau/(1+\tau)$ and $\tilde{\gamma} = \tilde{\tau}/(1+\tilde{\tau})$. The following proposition expresses this formally.

Proposition 1. *The linear regression estimate $\hat{\beta}$ defined in* (8) *satisfies*

$$\lim_{N\to\infty} \hat{\beta}_1 = \beta_1 \; \frac{1}{\mathbb{E}\left[\tilde{\gamma}/\gamma\right] + \frac{1}{\mathbb{E}[\tilde{\gamma}]} \times cov\left(\tilde{\gamma}, \tilde{\gamma}/\gamma\right)}$$

a.s. and thus it is a consistent estimator of β_1 if subjects' level of self-knowledge is correct, i.e., if $\tau = \tilde{\tau}$.

See Appendix Section A for the proof. This proposition implies that our regression coefficient is a consistent estimator if agent's characteristic and the degree of self-knowledge are independent, and the agent's level of subjective self-knowledge is correct. The intuitive reason for this result follows from our assumption that the incentives of the agent and the analyst are aligned. This assumption implies that each agent reports their best Bayesian estimate of their type. Hence, they weigh their signal relative to the prior exactly such that the effect of self-knowledge cancels out in the regression estimate.

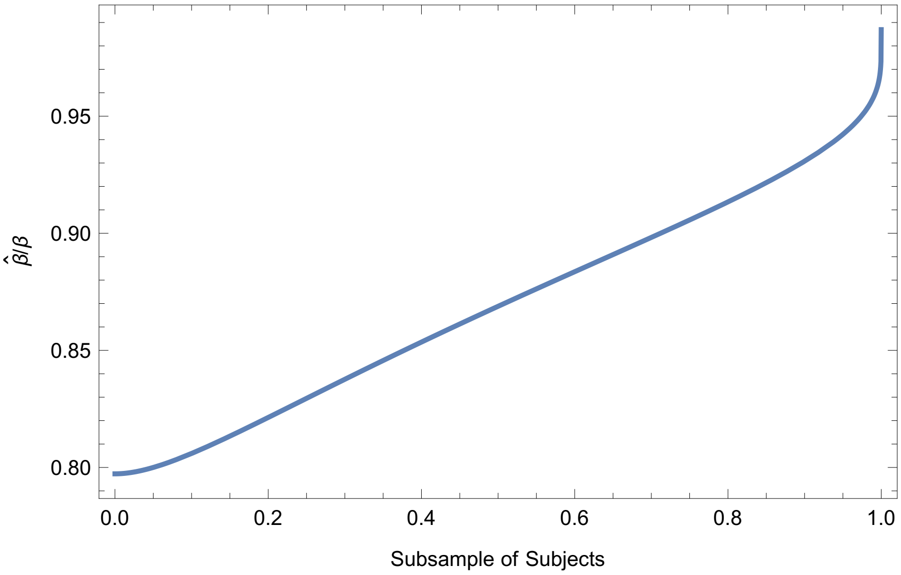
Importantly, Proposition 1 also provides the exact conditions that determine the extent and direction of the resulting bias if (some) agents do *not* correctly assess their level of subjective self-knowledge. For instance, if agents overestimates the precision of their signals, i.e., $\tilde{\gamma}_i/\gamma_i > 1$ and agents who overestimate themselves more are also more confident cov $(\tilde{\gamma}_i, \tilde{\gamma}/\gamma) > 0$, then the regression estimator will underestimate $\hat{\beta}$. Accordingly, an analyst will underestimate the effect of characteristic θ on the outcome variable y. Conversely, if agents underestimate their precision ($\tilde{\gamma}/\gamma < 1$) and, e.g., the extent of underestimation is uncorrelated with confidence cov $(\tilde{\gamma}, \tilde{\gamma}/\gamma) = 0$, an analyst will overestimate the effect of type θ on y. Put differently, Proposition 1 highlights that it is generally unlikely that the OLS estimate will be consistent.

2.4.2 Debiasing Regression Estimates by Sample Splitting

The previous section established that a consistent estimation of regression parameters is a knife-edge case in the presence of limited self-knowledge. Since the assumption required for consistency – all agents correctly know their level of self-knowledge – is likely violated in practice, estimates can be either over- or underestimated. Importantly, our framework suggests a simple way to reduce bias in this context. In the following, we show how restricting the sample to agents with high self-knowledge τ, debiases estimates.

We first illustrate the debiasing effect using a simulation. We simulate a scenario where $\hat{\tau} = 2\tau$. The figure shows that restricting attention to high self-knowledge individuals through subsampling brings the empirically estimated regression coefficient $\hat{\beta}$ closer to the true parameter β. In the limit, the empirical estimate is no longer biased. The intuition for this result is that agents with high levels of self-knowledge have limited potential to overestimate themselves, reducing the bias. That is, if τ is high, then $\gamma = \frac{\tau}{1+\tau}$ is close to 1, and as $\tilde{\gamma} \leq 1$, we have that $\tilde{\gamma}/\gamma$ will be close to 1.

Figure 2: The effect of restricting the sample to high self-knowledge subjects



Notes: The figure displays the ratio between an estimated regression coefficient $\lim_{N\to\infty} \hat{\beta}_1$ and the true coefficient β. Results are shown for a simulated population of subjects with $\hat{\tau} = 2\tau$.

The following proposition formally establishes the insight that focusing on agents with high self-knowledge improves regression estimates.

Proposition 2. *Suppose that $\tilde{\tau} = \alpha\tau$ for some constant $\alpha > 0$. We have that the bias*

$$\left|\beta_1 - \lim_{N\to\infty} \hat{\beta}_1\right|$$

is a.s. decreasing in $\underline{\tau}$ when restricting to subject's with $\tau \geq \underline{\tau}$.

See Appendix Section A for the proof. The reason we need to restrict attention to either overconfident or under-confident subjects is that they are biased in opposite directions. As a consequence, even though the bias is reduced within each group when restricting to higher self-knowledge individuals, the overall bias might not monotonically decrease if both under- and overconfident individuals are present in the population. However, in the limit, the bias will be eliminated independent of the distribution of under- and overconfident individuals.[6]

Effect of Sample Splitting on Explanatory Power. While Proposition 1 establishes that sample splitting leads to less biased regression estimates, this prediction is typically not directly testable. This is because the true β is typically not known to the analyst, and hence the direction in which a potential improvement should move is unknown. In contrast, for the explanatory power of a regression, the direction of an improvement is clear and measurable: an improvement means higher explanatory power. Hence, we next explore how the explanatory power of the estimated model reacts to focusing on high self-knowledge individuals. We measure explanatory power by the coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - [\hat{\beta}_0 + \hat{\beta}_1 r_i])^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

[6] Note that we have assumed that τ and types θ are independent. If they are correlated, subsetting may not (fully) eliminate regression bias in the limit.

The following Proposition shows that subsetting on τ also improves the coefficient of determination if agents are overconfident:

Proposition 3. *Suppose that $\tilde{\tau} = \alpha\tau$ for some constant $\alpha \geq 1$. We have that the coefficient of determination $\lim_{N\to\infty} R^2$ is a.s. increasing in $\underline{\tau}$ when restricting to subject's with $\tau \geq \underline{\tau}$.*

See Appendix Section A for the proof.

In sum, we have provided the conditions under which regression estimates are biased in the presence of limited self-knowledge. If an analyst were to observe subjects' levels of self-knowledge τ_i, then Proposition 2 shows that subsetting on τ_i reveals the direction and magnitude of this bias. Put differently, an estimator of self-knowledge can be used as a diagnostic tool to assess the extent and direction of potential bias. Moreover, a researcher can use the estimator to improve the explanatory power of their regressions. These insights motivate the next section, where we develop such an estimator for τ_i.

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.

3.1 Estimating Self-knowledge from Responses

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.[7] 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 (theoretical) variance of expected answers to different questions (between-variance), divided by the (theoretical) variance of answers to the same questions (within-variance). To construct an estimator $\hat{\tau}_i$, we use the sample variance between average answers for different characteristics as an approximation of the true between-variance and the average sample variance of answers for a given characteristic as an approximation of the true within-variance. Denote agent i's average answer to question k by $\bar{r}_{ik} = \frac{1}{T}\sum_{t=1}^{T} r_{ikt}$ and her average answer over all questions by $\bar{r}_i = \frac{1}{K}\sum_{k=1}^{K} \bar{r}_{ik}$. Our estimator $\hat{\tau}_i$

[7] See Appendix C.1 for a generalization to arbitrary and not necessarily identical distributions.

for the self-knowledge of agent i is given by

$$\hat{\tau}_i = \frac{\frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{r}_{ik} - \bar{r}_i\right)^2}{\frac{1}{K(T-1)-2}\sum_{k=1}^{K}\sum_{t=1}^{T}\left(r_{ikt} - \bar{r}_{ik}\right)^2} - \frac{1}{T}\,. \tag{9}$$

The numerator in the first summand of the expression captures the variation *between* the average answers of an agent for different characteristics, while the denominator expresses the average variation in answers *within* characteristics. Since the expected value of the ratio of two random variables is not the same as the ratio of their respective individual expected values, the denominator is adjusted by a constant factor relative to the unbiased estimator of the within-variance[8] and a correction term of $1/T$ is subtracted from the ratio. These two adjustments are necessary to ensure that the estimator is unbiased.

The following theorem establishes that $\hat{\tau}_i$ is a consistent and unbiased estimator of self-knowledge τ_i and describes its properties.

Theorem 1. *For every K, T that satisfy $K(T-1) > 4$.*

1. *The estimator $\hat{\tau}_i$ satisfies*

$$\hat{\tau}_i = \left(\tau_i + \frac{1}{T}\right)\frac{K(T-1)-2}{K(T-1)}F_i - \frac{1}{T} \tag{10}$$

for some random variable F_i that is F distributed with $K-1, K(T-1)$ degrees of freedom for every fixed vector of parameters $\tau_i, \sigma, \bar{\theta}$.

2. *$\hat{\tau}_i$ is an unbiased estimator for τ_i, i.e., $\mathbb{E}[\hat{\tau}_i \mid \tau_i] = \tau_i$.*

3. *The standard error of the estimator $\hat{\tau}_i$ is given by*

$$\sqrt{\mathbb{E}[(\hat{\tau}_i - \tau_i)^2 \mid \tau_i]} = \left(\tau_i + \frac{1}{T}\right)\sqrt{\frac{2((K-1) + K(T-1) - 2)}{(K-1)(K(T-1)-4)}}\,. \tag{11}$$

4. *$\hat{\tau}_i$ is a consistent estimator and converges to τ_i at the rate $1/\sqrt{K}$ in the number of attributes, and for all $K > 4$ it satisfies the following upper bound independent of the number of repeated observations T:*

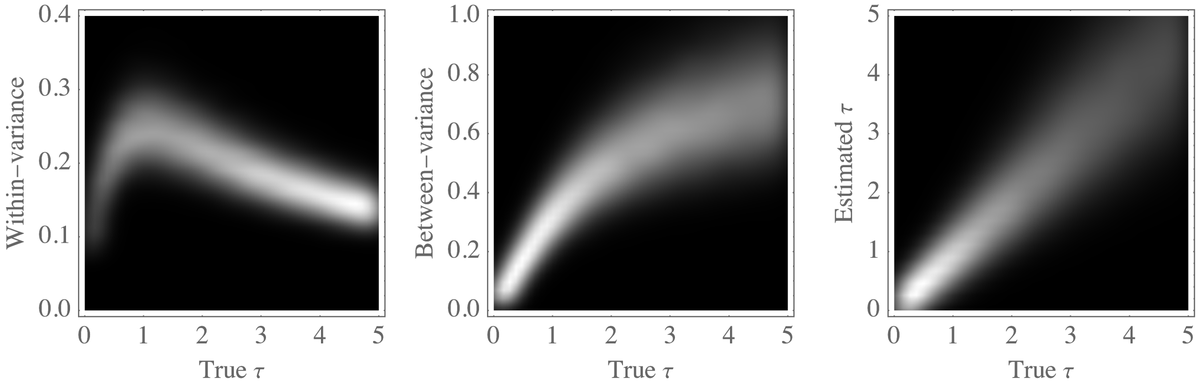
$$\sqrt{\mathbb{E}[(\hat{\tau}_i - \tau_i)^2 \mid \tau_i]} \leq \frac{2\tau_i + 1}{\sqrt{K-4}}$$

The proof of the theorem is provided in Appendix A. Part 4 of the theorem shows that for retrieving precise estimates, additional questions are more "valuable" than additional waves. This is the case because, intuitively, having additional questions adds to the precision of estimating both the between as well as the (average) within-variance, whereas additional waves only improve the precision of the estimated within-variance. Therefore, as K goes to infinity, the estimator converges to the true value even for just two waves, while the precision of the estimator is always limited for a finite number of questions.

Remark 1. *As we show in the proof of the theorem in Appendix A, the properties of the estimator extend unchanged to the model with subjective self-knowledge.*

[8] An unbiased estimator of the within-variance is given by $\frac{1}{K(T-1)}\sum_{k=1}^{K}\sum_{t=1}^{T}\left(r_{ikt} - \bar{r}_{ik}\right)^2$.

Figure 3: Simulations



(a) Within-variance

(b) Between-variance

(c) Estimated τ

Note: Kernel-density estimates, where lighter shading corresponds to a higher estimated density. Each panel is based on the same 100 simulations, each with I =1,000 hypothetical individuals, for whom reports about $K = 50$ characteristics are observed $T = 3$ times. The panels use Gaussian kernels with bandwidth selection according to Silverman's rule.

3.2 Simulating the Performance of the Estimator

Next, we illustrate our model and the behavior of the estimator using numerical simulations. For all illustrations, agents' levels of self-knowledge τ_i are drawn from a uniform distribution with support $[0.1, 5]$. The true average value of characteristics $\bar{\theta}$ is set to 5 and the true population variance σ^2 equals 1.

Figure 3 displays the joint distribution of the true level of self-knowledge τ_i and the sample within-variance, the sample between-variance, and estimated self-knowledge $\hat{\tau}_i$. For the within-variance, we observe the expected non-monotonic, hump-shaped relationship with the true level of self-knowledge (Figure 3a). The estimates for the between-variance increase in the true level of self-knowledge, but "fan out" for higher levels of true self-knowledge (Figure 3b). Our proposed estimator for self-knowledge is strongly concentrated around the 45-degree line and thus informative about agents' true levels of self-knowledge (Figure 3c).

In Table 1, we illustrate how the estimator performs for various sample specifications. We consider 100 or 10,000 agents, 15 or 50 characteristics, and 3 or 10 waves, respectively. For each scenario, we run 10,000 simulations and report the average value of three measures for the quality of the estimates: Pearson's correlation and Spearman's rank correlation between estimated and true self-knowledge and the proportion of simulated agents correctly identified as having a level of self-knowledge above or below the median. If our estimator had no informational value at all, we would expect a correlation and rank correlation of zero and 50% of correctly assigned agents in the median split.

The values of the correlation and the rank correlation coefficients of 0.68 and 0.76 shown in Column 1 for $I = 100$, $K = 15$, and $T = 3$ suggest that the estimator is already informative about self-knowledge for modest sample sizes. This is confirmed by 80% of hypothetical agents being assigned to the correct half of the sample in terms of self-knowledge. In Column 2, the number of hypothetical agents is increased to 10,000. The quality of predictions remains basically unchanged, reflecting the fact that our estimator does not use population information. However, as can be seen from Column 3, estimates strongly benefit from a larger number of characteristics (50 instead of

Table 1: Accuracy of estimates for different numbers of respondents, characteristics, and waves

	(1)	(2)	(3)	(4)	(5)
I (respondents)	100	10,000	100	100	100
K (characteristics)	15	15	50	15	50
T (waves)	3	3	3	10	10
Correlation τ and $\hat{\tau}$	0.68	0.68	0.87	0.76	0.91
Rank correlation τ and $\hat{\tau}$	0.76	0.77	0.90	0.82	0.93
Median split correct	80%	80%	88%	83%	90%

Notes: The table shows the results of simulating response behavior according to the model for different numbers of respondents, characteristics, and waves. For each resulting dataset, we estimate $\hat{\tau}$ and compare its correlation with the true τ. In the last row, we furthermore look at how many respondents are correctly classified as above or below median τ when performing a median split according to $\hat{\tau}$.

15), in line with Part 4 of the theorem. Relative to these increases, the increase in performance from a higher number of answers per characteristic in Column 4 (ten instead of three) is not quite as large (in line with Part 4 of the theorem, which shows that the standard error does not vanish in T). Column 5 combines the number of characteristics from Column 3 with the number of waves from Column 4, yielding the best performance, with correlation coefficients above 0.9 and a median split result of 90%. In sum, we find that the estimator performs reasonably well with a modest number of fifteen characteristics and three waves, and its performance can be increased, in particular, by increasing the number of characteristics.[9]

4 Experimental Evidence

This section presents experimental evidence. The purpose of the preregistered experiment is threefold: (i) to test the predictions of our model with respect to survey reports, (ii) to study the performance of our estimator of τ, and (iii) to investigate the impact of self-knowledge on regression estimates. The idea of the experiment is to create a choice environment where the researcher observes subjects' reports (allowing us to estimate τ) while at the same time *knowing the true type* θ. This allows us to study whether the relationship between reports and types follows our model's predictions, and to investigate whether our estimator is successful in identifying subjects whose reports are relatively more informative. In addition, by using an experiment, we can *exogenously* vary the level of self-knowledge τ. This allows us to study whether the impact on responses follows our model's predictions and whether our estimator of τ is capable of detecting this induced variation. Such tests are difficult—if not impossible—with non-experimental data, where true types are unknown to the researcher and self-knowledge cannot be exogenously varied.

[9] In these simulations (and the construction of the estimator), we have assumed that the characteristics relevant for estimating $\hat{\tau}$ are independent. In practice, this assumption may be violated, implying that responses from items measuring those characteristics are less informative for the estimation of $\hat{\tau}$. As we show in various simulations in Appendix Section C.2, however, $\hat{\tau}$ remains highly informative about the true τ, even if characteristics are strongly correlated. To illustrate, for the case of 100 respondents, 50 characteristics, and 15 waves, the rank correlation between $\hat{\tau}$ and true τ is 0.90 if characteristics are drawn independently. If we instead assume correlations among characteristics of 0.2, 0.5, or 0.8, respective rank correlations are 0.90, 0.88, and 0.82, i.e., quite similar. As expected, if all characteristics are perfectly correlated, $\hat{\tau}$ is no longer informative for τ.

4.1 Experimental Design

To create a choice environment with known types θ and an exogenous variation in knowledge τ, the experiment exposed subjects to a simple, repeated, and incentivized estimation task. The setup mimics a panel data set where respondents are repeatedly asked to respond to a set of different questions.[10]

Types. The requirement that the researcher knows true types implies that we cannot work with individual characteristics such as personality traits, preferences, or IQ, simply because these cannot be known with certainty. In Section 5, we analyze a more standard survey environment where the researcher does not know the types. To implement types known to the researcher (θ_i), we use a classic psychometric dot estimation task. We presented subjects with a series of 60 screens, each showing an image that contained between 60 and 150 dots. For an example of a dot image, see Appendix Figure E.1. On each screen, the dots were randomly distributed across the image. For each subject, we created 15 dot images by independently drawing the number of dots from a normal distribution. The 15 images were repeated four times, thereby creating 60 images for each screen. Within each repetition, we randomized the order of the images. For each image, subjects report the number of dots by choosing one of eleven size categories (see Appendix Table F.1 for the categories and their respective likelihoods). This procedure implements a panel structure, i.e., for every subject i, we observe a total of 60 reports for $K = 15$ characteristics in $T = 4$ periods. We use this structure to estimate $\hat{\tau}$, our estimator for self-knowledge τ, as described in Section 3.1. Prior to seeing the images, subjects received detailed descriptions and visualizations of the expected dot distribution and corresponding dot images. For details on the instructions, see Appendix Section L.

Payoff Function. We incentivized subjects to estimate the number of dots for each image as precisely as possible. The payoff function, π, implements a quadratic loss function and corresponds exactly to Equation (1) in the model, with

$$\pi(r) = -\left(r - \theta\right)^2 ,$$

where θ indicates the true type (number of dots) and r a subject's report. For the payoff, one of the 60 screens was randomly selected. For the selected screen and respective report, subjects received 10€ minus the product of 0.10€ and the squared difference between the true type and the report. For example, if a subject was shown an image of category 1 (50 dots – 68 dots) and estimated a number of dots corresponding to category 8 (111 dots – 117 dots), the subject received $10€ - (1 - 8)^2 \times 0.10€ = 5.10€$. The endowment of 10€ rules out losses even if the difference between the true and the estimated type was maximal.

Signal Precision and Treatments. To exogenously vary the precision τ of the signal, we ran two treatments that only differed in terms of how long subjects saw each of the 60 screens and how visible the dots were. In the treatment *Long*, subjects saw each screen for 7.5 seconds, and the dots were in dark grey. In contrast, in treatment *Short*, subjects saw each screen only for 0.5 seconds, and

[10] We also ran an experiment that used a different estimation task and was conducted in a laboratory instead of an online setting. Our results also replicate in this setting, see Appendix Section G for details.

the dots were in light grey, making them less visible. Treatments were randomly assigned between-subject such that each subject participated in only one condition.

Procedural Details. In total, 308 subjects took part in the experiment, most of them being undergraduate students of various majors at the University of Bonn. As pre-registered, we exclude 10 subjects who gave the same estimate for every one of the 60 screens. We also exclude one subject who correctly estimated the category for all 60 screens, which we consider only possible with computer-assisted tools. This leaves 297 subjects for the main analysis, 153 in the treatment *Long* and 144 in the treatment *Short*. The experiment was conducted online, for which we used oTree as the experimental software (Chen, Schonger, and Wickens, 2016). Recruitment was organized using the software hroot (Bock, Baetge, and Nicklisch, 2014). Subjects received detailed information about the rules and the structure of the experiment and were required to correctly answer several control questions. For participation, subjects received a show-up fee of 4€.

4.2 Hypotheses and Results of the Experiment

Our experimental data is well-suited for testing several hypotheses related to our model. Specifically, we formulate and test three sets of hypotheses. The first set concerns predictions our model makes on the relationship between true types and responses. The second concerns the performance of our estimator $\hat{\tau}$ in recovering subjects' level of self-knowledge τ. The third set relates to whether splitting samples based on $\hat{\tau}$ improves regression estimates.

4.2.1 Relationship between Types and Responses

Our model assumes that subjects' optimal reports are a weighted sum of the population average $\bar{\theta}$ and the received signal x, see Equation (2). In the experiment, we induce a normal distribution with a mean of five, which leads to the following testable hypothesis:

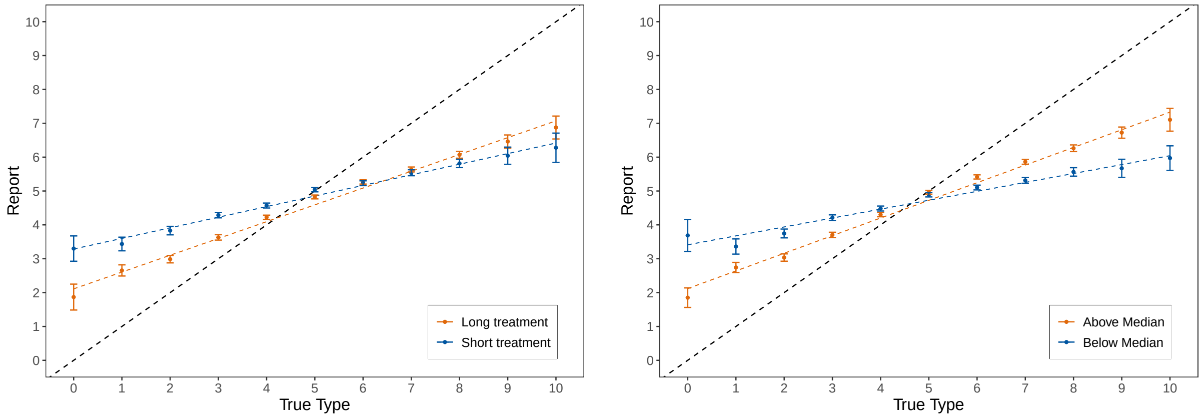
Hypothesis 1. *Average reports are linear in true types and biased towards the population average of the true types, i.e., towards five.*

The hypothesis can only be tested because, in our experiment, we know the true type. Graphically, we would expect average reports for different true types to lie on a straight line that is rotated clockwise around the population average. That is, we would expect an upward bias for low types and a downward bias for high types.

Importantly, as Equation (2) also shows, subjects' level of self-knowledge τ should influence the extent of the bias. In our model, subjects recognize (potentially imperfectly) and take into account their *individual-specific* level of τ. Accordingly, the lower a subject's level of self-knowledge τ, the stronger the bias in their reports toward the average value of the characteristic. Since we exogenously vary τ in the experiment, we can test this causal relationship empirically, as stated in the following hypothesis:

Hypothesis 2. *The bias towards the population average is stronger in the Short-treatment than in the Long-treatment.*

Figure 4: Relationship between reports and types in the experiment



(a) Reports vs. true types split by treatment

(b) Reports vs. true types split by estimated τ

Our third hypothesis concerns the predictive power of reports for true types. According to our model, τ influences the report's predictive power: the higher the level of τ in a given population, the stronger the predictive power of reports for true types. Using our exogenous variation, we can test this prediction empirically:

Hypothesis 3. *The predictive power of reports for types is stronger in the Long-treatment than in the Short-treatment.*

Testing Hypothesis 1. Figure 4 Panel A provides a visual test of Hypothesis 1. It plots, separately for the two treatments, the average report subjects provide given each true type. As shown in the figure, average reports increase fairly linearly in types. However, the increase is markedly less steep than the 45-degree line, and a linear regression of types on reports intersects the 45-degree line close to the population average of five. Hence, as predicted, average reports are biased towards the population average, with overestimation for types below the population average and underestimation for types above. Pooled across both treatments, the slope coefficient of regressing types on reports is 0.581, which is significantly smaller than one (see Column (1) of Table 2 for details).

Testing Hypothesis 2. Figure 4 Panel A also provides visual evidence for Hypothesis 2. The extent to which average reports are biased is markedly higher in the *Short* treatment compared to the *Long* treatment. Accordingly, the regression slope in the latter is steeper, which is quantified in Columns (2) and (3) of Table 2. Indeed, the interaction term between reports and treatment is significant ($p < 0.001$). Hence, we confirm our hypothesis that reports are more biased towards the population average when τ is low.

Testing Hypothesis 3. Lastly, we investigate the predictive power of reports for identifying types. As a measure of predictive power, we use the R^2 of regressing types on reports. Our hypothesis implies that the R^2 in the *Long* treatment should be higher than in the *Short* treatment. This pattern is confirmed by the data: while the R^2 in the *Short* treatment is 0.149, it more than doubles to 0.349 in the *Long* treatment. Hence, exogenously increasing τ increases the predictive power of reports for types.

Table 2: Relationship between reports and true types

	Dependent variable: True type				
Subjects	*all*	*by treatment*		*by $\hat{\tau}$*	
		Short	*Long*	*low*	*high*
	(1)	(2)	(3)	(4)	(5)
Report	0.581*** (0.023)	0.462*** (0.028)	0.691*** (0.032)	0.391*** (0.027)	0.758*** (0.025)
Constant	2.168*** (0.120)	2.697*** (0.151)	1.706*** (0.178)	3.088*** (0.146)	1.303*** (0.134)
Subjects	297	144	153	144	153
Observations	17,820	8,640	9,180	8,940	8,880
Report $\neq 1$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
R^2	0.243	0.149	0.349	0.113	0.404
ΔR^2		134%		258%	

Notes: The table reports OLS estimates. Standard errors (in parentheses) are clustered at the subject level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

4.2.2 Performance of the Self-knowledge Estimator

Next, we evaluate the performance of $\hat{\tau}$ – our estimator for τ. First, since our treatment manipulates self-knowledge, and our estimator is designed to capture self-knowledge, $\hat{\tau}$ should be able to predict subjects' treatment status. That is, we expect that we can blindfold ourselves regarding the treatment status and be able to tell only from the patterns in answers to which treatment a given subject was assigned. As the *Long*-treatment increases self-knowledge, we have as hypothesis:

Hypothesis 4. *Estimates $\hat{\tau}$ are larger for subjects in the Long-treatment than for those in the Short-treatment.*

The next hypothesis relates $\hat{\tau}$ to response patterns. If our estimator indeed captures subjects' self-knowledge, we should see that the bias in reports towards the average value of the characteristic varies in $\hat{\tau}$. Specifically, we have:

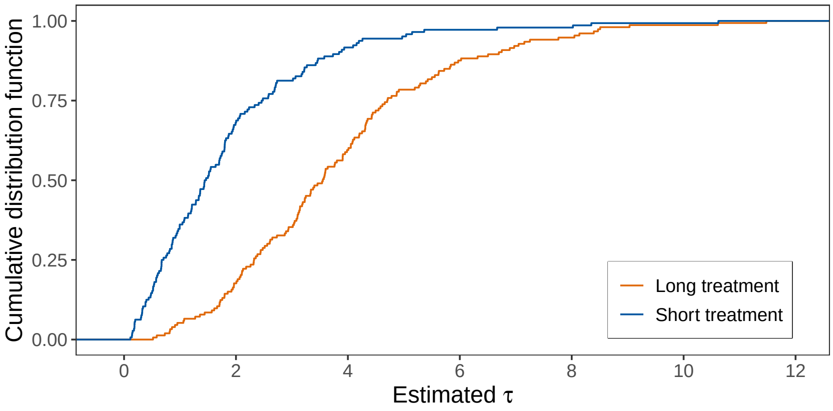
Hypothesis 5. *The bias towards the population average is stronger for high $\hat{\tau}$ subjects compared to low $\hat{\tau}$ subjects.*

Most importantly, we want to directly test whether our estimator is capable of identifying the informativeness of survey responses. That is, the responses of subjects for which we estimate high levels of $\hat{\tau}$ should have more predictive power for their respective true types than subjects with low $\hat{\tau}$:

Hypothesis 6. *The predictive power of reports for types is stronger for high $\hat{\tau}$ subjects compared to low $\hat{\tau}$ subjects.*

Testing Hypothesis 4. In support of this hypothesis, $\hat{\tau}$ is higher for subjects in the *Long*-treatment (average $\hat{\tau}$: 3.82, median: 3.55) compared to those in the *Short*-treatment (average $\hat{\tau}$: 1.86, median: 1.45), a significant difference ($p < 0.001$, t-test). In fact, the distribution of $\hat{\tau}$ in *Long* stochastically

Figure 5: Distribution of estimated τ in the experiment



dominates the distribution of $\hat{\tau}$ in *Short*. This is shown in Figure 5, which displays the cumulative distribution function of $\hat{\tau}$, separately for the *Short* and the *Long*-treatment, respectively. Put differently, our estimator predicts subjects' treatment status. This is also shown in a simple Probit regression where we regress an indicator for the *Long*-treatment on our estimates for τ ($p < 0.001$, two-sided).

Testing Hypothesis 5. Having established the sensitivity of our estimator, we now examine how $\hat{\tau}$ can help to understand bias between reports and types. Figure 4 Panel B provides visual evidence that the level of $\hat{\tau}$ can be used to assess the severity of the bias. In the figure, we split the sample into above and below median subjects based on their estimated level of self-knowledge, $\hat{\tau}$. We then plot for each sample separately their average response for each type. Average reports are more biased towards the population average among subjects with below median $\hat{\tau}$ compared to subjects with above median $\hat{\tau}$. This pattern closely resembles the pattern of Panel A, where subjects are split based on their treatment status. We quantify the difference in slopes in Columns (4) and (5) of Table 2, which display the respective regression slopes. They are significantly different from each other, as indicated by the significant interaction term between reports and an indicator of the median split ($p < 0.001$).

Testing Hypothesis 6. Lastly, we turn to analyzing whether $\hat{\tau}$ can identify the predictive power of reports for types. Using R^2 as a measure of predictive power, we find that the R^2 of regressing types on reports is 258% higher (from 0.113 to 0.404) when moving from below- to above-median $\hat{\tau}$ subjects.[11] Thus, high $\hat{\tau}$ subjects provide more informative reports about their true type than low $\hat{\tau}$ subjects. In addition, our results also show that using $\hat{\tau}$ leads to a larger increase in R^2 than moving from the *Short* to the *Long* treatment. This is remarkable, given that our estimator only uses the pattern of subjects' responses.

Using individual-level data, we can investigate the relationship between predictive power and $\hat{\tau}$ in more detail. Recall that each subject in the experiment made 60 estimation decisions. This means that we can run regressions of these 60 reports on the respective true states *separately for each individual*. The resulting individual-specific value of R^2 is informative about how well a subject is able to discriminate between different true states. Moreover, the individual slope parameter reveals how much weight is assigned to signals. The parameter is thus informative about the level of subjective

[11] Focusing on the top 20% subjects with the highest $\hat{\tau}$ further increases the R^2 to 0.516.

knowledge $\tilde{\tau}$. Several observations can be made. First, in individual-level regressions, the values of R^2 and the slope coefficients are strongly positively related, with a rank correlation of 0.73 ($p < 0.001$, two-sided, Pearson: 0.73). This positive correlation supports the central assumption of the model: agents who receive more precise signals (making their responses more predictive of their type, as measured by R^2) place more weight on those signals (as measured by the slope coefficient). Second, the individual-level values of R^2 allow us to further test the validity of our estimator, which does *not* use information about the true types. We find that the individual values of R^2 are strongly correlated with the values of $\hat{\tau}$: the rank correlation is 0.94 ($p < 0.001$, two-sided, Pearson: 0.87).

This relationship can be analyzed even more thoroughly. In light of our model, the R^2-values can be transformed into alternative estimates of τ according to the formula $\hat{\tau}_{\text{alt.}} = R^2/\left(1 - R^2\right)$. For the derivation, see Appendix I. The Pearson correlation between the alternative estimate and our main estimate $\hat{\tau}$ is 0.94 ($p < 0.001$, two-sided, Pearson: 0.93). This finding is not mechanic since the identification approaches behind the two estimators rest on entirely different information in the data: the R^2-based estimator uses the information about true states as input, while our self-knowledge estimator $\hat{\tau}$ only uses reports.

4.2.3 Impact of Sample Splitting on Regression Estimates

Lastly, we turn to testing Propositions 2 and 3. If our estimator $\hat{\tau}$ identifies τ, we can use it to improve regression estimates. Formally, when regressing an outcome variable on survey responses, we should observe that, in the presence of biased OLS regressions, subsetting based on $\hat{\tau}$ reduces the bias and increases R^2:

Hypothesis 7. *Restricting the sample to subjects with high $\hat{\tau}$ reduces bias in OLS regression estimates and increases the R^2 of regressions.*

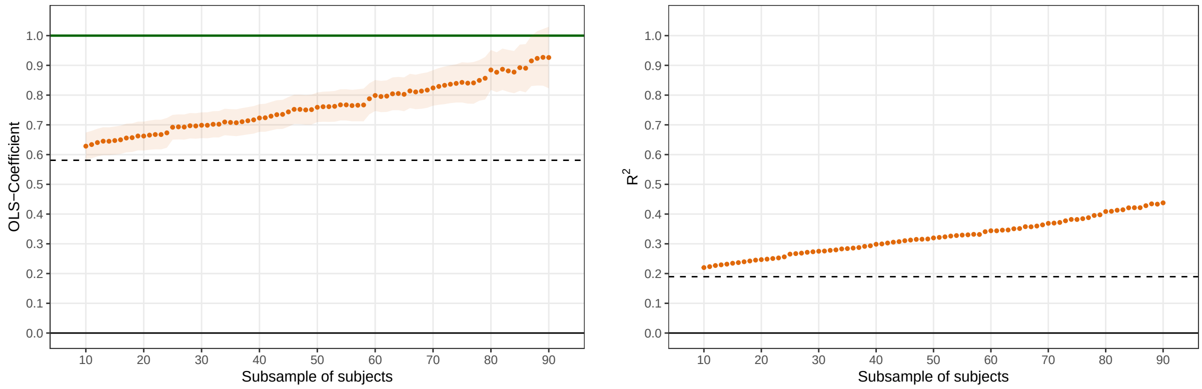
If confirmed by our data, the hypothesis provides evidence on the usefulness of our estimator for improving inference in regressions.

Empirical strategy. To investigate the relationship between our estimator of self-knowledge and bias in OLS regressions, we exploit the fact that we know the true type and simulate a linear relationship between types and an outcome variable. Against this benchmark, we can investigate how using subjects' responses instead of true types biases regression estimates, and, importantly, how we can use our estimate of self-knowledge to detect such biases.

Specifically, we simulate a new variable y_i that is a linear combination of the true type θ_{ik} (number of dots) plus a normal noise component. That is, we construct $y_{ik} = \beta_0 + \beta_1 \theta_{ik} + \varepsilon$, with $\varepsilon \sim N(0, 1)$ and set $\beta_0 = 0$, $\beta_1 = 1$. We now want to compare how closely we can estimate β when, instead of true types, we only observe subjects' reports of true types, which in the context of our experiment is the number of dots r_{ik} subjects report. The regression thus becomes $y_{ik} = b_0 + b_1 r_{ik}$. This setup thus replicates the context studied in Section 2.4. With our simulation based on experimental data, we can compare how b_1 relates to the objective benchmark β_1 depending on the degree of self-knowledge that we estimate from r_{ik}.

Results. We run $1,000$ simulations, in each of which we generate y_{ik} as described above and run an OLS regression using subjects' reports of the true type. Figure 6 displays our results. We start in

Figure 6: Relationship between reports and types in the experiment



(a) Simulating the influence of subsetting based on $\hat{\tau}$ on coefficients

(b) Simulating the influence of subsetting based on $\hat{\tau}$ on R^2

Notes: **Panel (a):** Each dot represents the average OLS-coefficient obtained from running 1,000 simulations, where in each we regress a simulated outcome variable based on the true number of dots on subjects' reported number of dots in the experiment. Each regression is run on a subsample where a percentage of subjects with the lowest estimated level of self-knowledge $\hat{\tau}$ are excluded. The x-axis denotes the respective percentage removed, i.e., the first dot denotes the sample where the 10% of subjects with the lowest $\hat{\tau}$ are removed. The solid line represents the true coefficient $\beta = 1$. The dashed line represents the full sample coefficient. Shaded areas indicate the average 95% confidence interval across all simulations. **Panel (b):** Each dot represents the average R^2 obtained from the regressions simulated in Panel (a). The dashed line represents the full sample R^2.

Panel (a) with the influence on regression coefficients. The dashed line represents the full sample coefficient obtained from averaging over all simulations, while the green line represents the true coefficient. As displayed, the OLS coefficient is biased away from the true coefficient.[12] Turning to the impact of subsetting based on $\hat{\tau}$, each dot in the figure represents the average OLS coefficient across all simulations of regressing y_{ik} on r_{ik}. We plot the coefficient for different subsamples where we subsequently remove more subjects with low $\hat{\tau}$. Displayed are subsamples in 1% increments. For instance, the first dot displays the coefficient when 10% of the subjects with the lowest $\hat{\tau}$ are removed from the sample. The second dot then displays the coefficient when 11% of subjects are discarded, and so on. As evident from the figure, removing subjects with low $\hat{\tau}$ brings the coefficient closer to the true coefficient of 1. Accordingly, subsetting based on $\hat{\tau}$ mitigates biases in regressions, and the mitigation is monotonic in the estimated level of τ, as predicted by Proposition 2. In Panel (b), we show that a similar effect occurs with respect to R^2: subsetting based on $\hat{\tau}$ increases the simulated regression's R^2. These results provide empirical evidence for the usefulness of our estimator in assessing the extent and direction of biases introduced by subjects' responses. In the next section, we will apply our $\hat{\tau}$ estimator to an actual survey environment.

5 Survey Evidence

An important aim of our exercise is to improve the informativeness of survey evidence for explaining and predicting behavior. Consider the case of economic preferences, for example. Many researchers are interested in explaining a particular behavioral outcome, such as investment or employment de-

[12] As for the source of the bias, in Appendix Section J, we provide some evidence that subjects have subjective self-knowledge, which according to Proposition 1 biases regression estimates.

cisions, with the help of self-reported survey measures of risk or social preferences.[13] In the presence of limited self-knowledge, however, the latter are more or less informative depending on how accurately individuals can assess their willingness to take risks or act prosocially in response to a given survey item. In this section, we demonstrate how researchers can assess survey respondents' levels of self-knowledge with the help of a simple, repeated version of the Big-5 inventory. Based on this "self-knowledge module", we investigate the degree to which our estimator of self-knowledge improves self-reported preference measures in predicting actual choices. More specifically, we ask whether individuals whom we classify as having relatively high levels of self-knowledge display a stronger association and a higher explained variance between self-reports and behavior, in comparison to those with comparatively low self-knowledge.

To provide such a test, we ran a large, pre-registered survey. The survey contains our self-knowledge module as well as a set of self-assessments and actual behaviors related to risk and altruism preferences. This set-up allows us to test our estimator's performance in improving survey responses. It also allows us to compare the performance of our estimator with that of other commonly used methods for improving the quality of survey responses.

5.1 Survey on self-assessment and behavior

Subjects participated in two subsequent waves, separated by one week. In the first wave, we asked subjects to self-assess their preferences with respect to risk and altruism and measure related behaviors with the help of incentivized and non-incentivized elicitations. The second wave repeats the self-assessments. Each session is supplemented by several questionnaires eliciting demographics and additional variables related to survey response behavior.[14] In the following, we explain these measures in detail.

Self-assessments. Our main variables of interest are self-assessed risk taking and altruism. For this purpose, we use the widely-used "general risk" question, which reads "How do you see yourself: are you generally a person who is willing to take risks or do you try to avoid taking risks?", using an 11-point Likert scale with higher values indicating a higher willingness to take risks (taken from Dohmen et al., 2011). For altruism, we use an item from Falk et al. (2018), which reads "How much would you be willing to give to a good cause without expecting anything in return?", again on an 11-point Likert scale with higher values representing a higher degree of altruism. We complement these two "general" self-assessments with five domain-specific assessments for risk taking (Dohmen et al., 2011), and five domain-specific assessments for altruism (Falk et al., 2018).

Each set of self-assessments (risk and altruism) was fielded two times in both waves, i.e., four times in total. Within each session, we first fielded the standard version of the respective item and then, later in the survey, a slightly rephrased version.[15] For instance, the rephrased general risk question reads "Which description fits you better: Do you tend to shy away from risks, or are you generally a risk-taker?", while the rephrased general altruism question reads "How willing are you to

[13] Among hundreds of examples, consider for instance that self-assessed risk and altruism have been related to sustainable investment decisions (Heeb et al., 2023), willingness to act against climate change (Andre et al., 2024), compliance with public health behaviors (Fang et al., 2022), labor market outcomes (Kosse and Tincani, 2020), job preferences (Non et al., 2022), self-employment (Caliendo, Fossen, and Kritikos, 2014), and social interactions (Falk et al., 2018).

[14] Complete instructions and details of the survey are reported in Appendix Section L.

[15] We use slightly rephrased versions to avoid triggering consistency concerns among subjects.

donate to a charitable cause without any personal benefit?". Both versions, standard and rephrased, are identical in waves one and two, respectively.

A module to assess self-knowledge. To estimate $\hat{\tau}$ according to equation (9) we need to observe repeated responses to a set of personality-related items. In many existing panel data sets, such data already exist. If such data are not available, however, it is straightforward to construct them. Here, we suggest using a repeated measure of the Big-Five inventory (neuroticism, extraversion, openness, conscientiousness, and agreeableness). Specifically, we use the 15-item scale (Schupp and Gerlitz, 2008) that is regularly fielded as part of the German Socio-Economic Panel (SOEP). Using this short Big-5 version as a module for estimating self-knowledge has the advantage that it needs relatively little survey time[16] and, as a byproduct, generates useful data on respondents' personality. The module consists of the original 15-item scale and one repetition, using a slightly rephrased version[17], i.e., it consists of 30 items in total. The complete module can be found in Appendix Table F.2. It was fielded in both waves.

Behavior under risk. To measure incentivized risk behavior, we elicit subjects' certainty equivalent for a lottery. Subjects face a series of 11 binary decisions between (i) a lottery with a 50% chance of paying 1€ and a 50% chance of paying nothing and (ii) a safe option that pays with certainty a particular monetary amount. Between decisions, we vary the amount received with certainty from 0€ (first decision) to 1€ (final decision). As pre-registered, we define the certainty equivalent as the first amount for which subjects switch from choosing the lottery to choosing the certain amount. Higher values thus indicate a higher willingness to take risks.

In addition, we measure self-reported risk behavior in various specific contexts, building on the four dimensions of risky behavior studied in Dohmen et al. (2011). In particular, we ask subjects whether (i) they currently smoke cigarettes, (ii) their household owned stocks in the past year, (iii) they are currently or were in the past year self-employed, and (iv) they currently do sports.

Altruistic behavior. To derive an incentivized measure of altruistic behavior, we run a simple dictator game with a charity as recipient. Subjects were endowed with 1€ and could allocate the money in increments of 0.10€ between themselves and a charity that helps children suffering from cancer. Higher values indicate a higher donation to the charity.

As for risk, we also assess altruistic behavior in four specific contexts. Here, we rely on the four dimensions of prosocial behavior that are elicited in the Gallup World Poll (see Falk et al., 2018). Specifically, we ask subjects whether, in the last month, they (i) donated money to a charity, (ii) did voluntary work in a non-profit, (iii) helped a stranger, or (iv) gave a gift to another person.

Measures to improve survey response behavior. To compare how our estimator performs in improving survey evidence relative to other commonly used methods, we draw from a wide range of previous work and include their proposed variables in our survey. We use self-reported and behavioral measures of attention and effort, different averaging strategies, as well as anchoring vignettes

[16] In our survey, the median completion time for the 15-item scale was 53 seconds, and 90% finished in less than 100 seconds.

[17] For example, instead of agreement to "I'm someone who has a vivid imagination and ideas," the rephrased version reads "I see myself as a person who is imaginative and has creativity." .

and strategies building on subjects' response time. Table 3 summarizes the methods and measures we use.

Sample and procedure. We partnered with the survey company Kantar, a commonly used provider of online survey participants. We recruited a quota-representative sample with respect to age (three bins), gender (two bins), and education (two bins). In total, 1001 subjects completed the first wave. Of those, 740 also completed the second wave. As pre-registered, we restrict our analyses to the latter sample. Of those, 48% are male, the average age is 53 years (SD = 15), and 37% have a school degree that enables admission to universities ("Abitur"). Subjects took a median time of 11 minutes to complete the first, and 7 minutes to complete the second wave. At the end of the second wave, for each subject, we randomly selected either the incentivized risk or altruism choice. If the former was chosen, we randomly selected one of the binary risky decisions. The selected decision was then implemented and paid. Subjects were informed about this procedure prior to making their decisions.

5.2 Results

In presenting our results, we first describe our estimate of $\hat{\tau}$ and show that our test environment is well suited to study the improvement of survey items, i.e., we show that our measures of self-assessed risk and altruism significantly correlate with risky and altruistic behavior, respectively. This sets the stage for our main analysis. Our criteria for assessing improvements in survey items are as follows: higher OLS coefficients of self-assessments when regressing behavior on self-assessments (Criterion 1), increased explained variance in terms of R^2 (Criterion 2), and higher test-retest stability of self-assessments (Criterion 3).

Estimating $\hat{\tau}$. We estimate $\hat{\tau}$ according to equation (9) using subjects' responses to the self-knowledge module (Big-Five inventory) fielded in waves 1 and 2. This allows us to assign a value for $\hat{\tau}$ for 99.2% of subjects in our sample. The remaining 0.8% (6 subjects) display no variation in their responses to the Big-Five questionnaire (neither across items nor over time), rendering it impossible to estimate $\hat{\tau}$ for them. The average $\hat{\tau}$ has a value of 24.88, with a median of 11.88 (see Appendix Figure E.2 for the distribution of $\hat{\tau}$). Note that our model assumes that τ is an individual-specific parameter, which is stable over time and across domains. Using our survey data, we test and confirm this notion. In particular, we find that when we estimate τ separately for the first wave and the second wave, both estimates are highly correlated ($p < 0.001$. Likewise, there is a significant positive relationship between values of τ that are estimated using our Big-Five module with estimates of τ when using the set of self-assessments concerning risk attitudes and altruism (see Appendix D for details). Lastly, supporting the notion that our estimator is related to the informativeness of survey responses more generally, we find that $\hat{\tau}$ is also significantly correlated with both self-reported and revealed measures of response quality (for details, see Appendix Table F.3).

Test environment: the association of self-assessments with behavior. To study improvements in survey items, we need a test environment that displays a significant association between self-reports and behavior, which is precisely what guided our selection of survey items and respective behaviors. Consistent with previous research, we find that self-assessed risk-taking and altruism indeed predict

Table 3: Summary of Measures and Estimators

Measure	**Measurement**	**Construction**
Self-knowledge estimator $\hat{\tau}$ Top 50%	We estimate $\hat{\tau}$ according to equation (9) using subjects' responses to the Big Five items.	Median-split using $\hat{\tau}$.
Self-knowledge estimator $\hat{\tau}$ Top 20%	We estimate $\hat{\tau}$ according to equation (9) using subjects' responses to the Big Five items.	Subsetting on top 20% $\hat{\tau}$ subjects.
Self-reported attention	Using the item from Meade and Craig (2012), we ask subjects at the end of the survey how much attention they paid to the study (5-point Likert scale).	Median-split using the self-reported attention measure.
Revealed attention	We use two standard attention checks (Berinsky et al., 2021). The first asks for subjects' favorite color with instructions to respond "Orange", the second asks about survey participation with instructions to select both "Frequent" and "Rarely."	Splitting the sample into whether subjects pass the first check (measure 1), the second (measure 2), or both (measure 3).
Self-reported effort	Using the item from Meade and Craig (2012), we ask subjects at the end of the survey how much effort they put towards the study (5-point Likert scale).	Median-split using the self-reported effort measure.
Revealed effort	We ask subjects about their opinion on Daylight saving time. We ask subjects to write at least 25 words as a response.	Splitting the sample into whether subjects write at least 25 words or not.
Self-reported reliability	Using the item from Dohmen and Jagelka (2023), we ask subjects at the end of the survey how reliable their responses are (11-point Likert scale).	Median-split using the self-reported reliability measure.
Response time	We measure completion time for both sessions, used as an indicator for careless responses (Curran, 2016).	Splitting the sample by excluding the top 10% fastest subjects (measure 1) or by excluding the fastest and slowest 25% (measure 2).
Average domain response	We average each self-assessment across all domain-specific self-assessments plus the respective general risk or altruistic self-assessment.	Using the average response across domains.
Average responses	We average each self-assessment across all four repetitions.	Using the average response across repetitions.
Demographic control	Controlling for age, gender, education, income, and happiness.	Multivariate regression using the control variables.
Anchoring vignettes	Using the method developed by King and Wand (2007), we field six anchoring vignettes, three for the general risk question and three for the general altruism question. The vignettes describe individuals as highly, medium, or little risk-seeking/altruistic and subjects rate each individual using the same scale as for the altruism/risk question.	We adjusted the risk and altruism self-assessments according to their vignette responses.
ORIV	Using the IV-strategy of Gillen, Snowberg, and Yariv (2019) with repeated self-assessments.	Using the repeated measures as an instrument in the regression.

behavior[18]: For all five measures of risky behavior, the OLS coefficient of the general risk question is positive and significant, ranging from 0.009 to 0.047. Similarly, for all five measures of altruistic behavior, the coefficient of the general altruism question is positive and significant, with values ranging from 0.020 to 0.067 (for details, see columns (1) – (5) of Appendix Table F.4).

We also derive a composite measure for the relation between self-assessment and the whole set of behavioral measures. To estimate this composite effect, we stack all five behavioral measures of each domain (risk, altruism) together. Then, we run a fixed-effects regression of behavior on assessments, controlling for each behavior with dummies and clustering standard errors on the subject level. In this specification, we find a coefficient of 0.030 for the general risk question and a coefficient of 0.041 for the general altruism question, both significant at any conventional level (see column (6) in Appendix Table F.4). Hence, these positive associations provide an appropriate test environment for our main analysis, to which we turn next.

Table 4: Improving the relationship between self-assessments and incentivized behavior using $\hat{\tau}$

	Dependent variable:					
	Lottery certainty equivalent			Dictator game giving		
	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median
	(1)	(2)	(3)	(4)	(5)	(6)
General risk question	0.009^{**}	0.017^{***}	0.003			
	(0.004)	(0.006)	(0.005)			
General altruism question				0.066^{***}	0.077^{***}	0.053^{***}
				(0.004)	(0.004)	(0.007)
Constant	0.229^{***}	0.214^{***}	0.243^{***}	0.142^{***}	0.079^{**}	0.215^{***}
	(0.017)	(0.021)	(0.027)	(0.028)	(0.033)	(0.046)
Observations	734	367	367	734	367	367
R^2	0.008	0.026	0.001	0.233	0.357	0.132

Notes: The table shows OLS estimates. In Column (1), the independent variable is the general risk question. Measured on an 11-point Likert scale, higher values indicate more self-assessed risk-seeking. The dependent variable is the elicited certainty equivalent of a lottery. Higher values indicate a higher certainty equivalent measured in money. Columns (2) and (3) build on this regression and display results of a sample split based on the median level of self-knowledge, estimated using the estimator $\hat{\tau}$. The independent variable of Column (4) is the general altruism question. Measured on an 11-point Likert scale, higher values indicate more self-assessed altruism. The dependent variable is the monetary amount given to charity in a dictator game. Columns (5) and (6) build on this regression and display results of a sample split based on the median level of estimated self-knowledge. Robust standard errors are displayed in parentheses are clustered at the subject level. Significance levels: $^{*}p < 0.1$,$^{**}p < 0.05$, $^{***}p < 0.01$.

Criterion 1: Improving OLS-regression coefficients using $\hat{\tau}$. Our model predicts that the positive relationship between self-assessments and behavior is biased in the presence of limited self-knowledge whenever subjects do not perfectly know their own level, as documented in Proposition 1. Importantly, as we show in Proposition 2, by subsetting based on $\hat{\tau}$, we can use $\hat{\tau}$ to assess the extent and direction of the bias. Accordingly, subsetting should improve the quality of survey items based on our first quality criterion.

We start testing this hypothesis by focusing on incentivized behavior. In Table 4, we regress behavior in the form of the elicited certainty equivalent on the general risk in columns (1) – (3) and dictator game giving on the altruism question in columns (4) – (6). To investigate the impact of limited self-knowledge, we perform a median split based on $\hat{\tau}$. We find that the coefficient of the above median sample is larger than for the below median sample in both instances. In fact, for

[18] We use the first wave standard measure of self-assessed risk-taking and altruism for the regression. Using the repeated measure or the measure elicited in the second wave yields similar results.

incentivized risk behavior, the association is no longer statistically significant among subjects with below-median $\hat{\tau}$ (column (3)).[19]

In Table 4, we study improvement in the coefficients of self-assessments using only the two incentivized behavioral measures and using a median split in self-knowledge. In Panel A1 and A2 of Figure 7, we expand our test strategy to the composite effect across all ten behavioral measures and take into account a more fine-grained range of splits.[20] Each dot in the figure represents the OLS coefficient of one stacked regression of risk behavior on the general risk question in Panel A1 and of altruistic behavior on the general altruism question in Panel A2. We plot the coefficient for subsamples where the x% of subjects with the lowest $\hat{\tau}$ are removed, i.e., we subsequently subset on subjects with higher $\hat{\tau}$. Displayed are subsamples in 1% increments. For instance, the first dot in Panel A1 displays the coefficient of regressing risk behavior on the general risk question when 10% of the subjects with the lowest $\hat{\tau}$ are removed. As evident from the figure, the relationship between self-assessments and behavior increases when focusing on subjects with higher $\hat{\tau}$. Thus, we can replicate the finding that subsetting based on $\hat{\tau}$ improves OLS coefficients that we document in our experiment in Figure 6.[21] Overall, the improvements are fairly monotonic, in line with Proposition 2.

Criterion 2: Improving explained variance using $\hat{\tau}$. Next, we turn to analyzing improvements in R^2 by splitting samples according to $\hat{\tau}$. According to Proposition 3, subsetting should increase the R^2 under some assumptions in the presence of bias. Starting with incentivized behavior in Table 4, we find an increase in R^2 for risk and altruism based on a median split of $\hat{\tau}$. Relative to baseline, the increase is $2,936\%$ for risk (columns (1) and (2)), and 171% for altruism (columns (4) and (5)).

Turning to all ten behavioral measures and more fine-grained splits, we repeat the regressions of Panel A1 and A2 of Figure 7, but now display in Panel B1 and B2 the respective R^2 of each regression. That is, each dot represents the R^2 of one stacked regression of risk (altruistic) behavior on the general risk (altruism) question. As before, dots display subsamples based on removing fractions of subjects with low $\hat{\tau}$. We again find a relatively monotonic relationship between subsamples of $\hat{\tau}$ and values of R^2. The more the sample focuses on subjects with high $\hat{\tau}$, the higher the R^2, in line with Proposition 3.

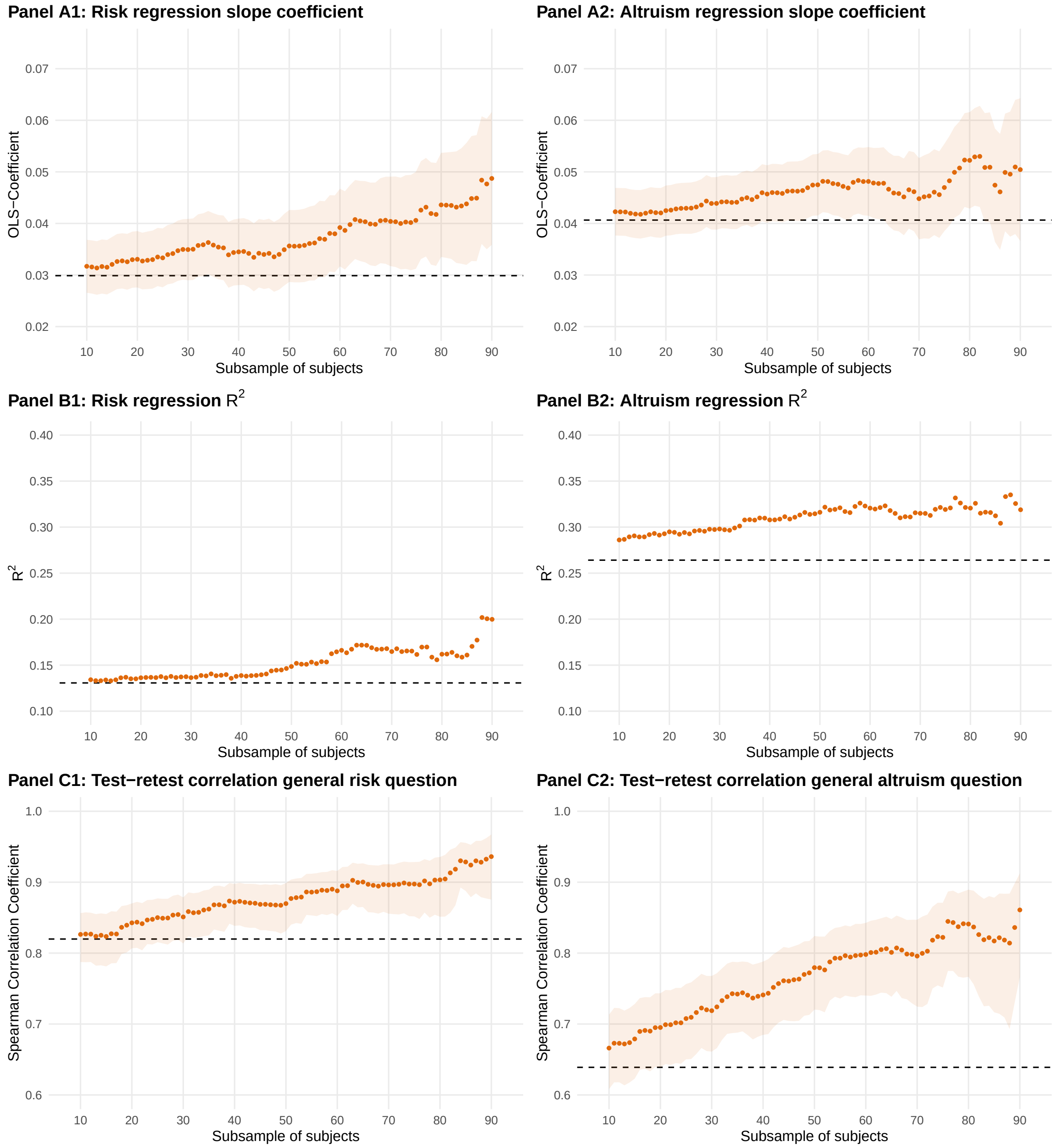
Criterion 3: Improving test-retest correlations using $\hat{\tau}$. A common criterion for the quality of a survey item is its test-retest stability. Our third analysis thus concerns the relationship between our self-knowledge estimator and the stability of self-assessments over time. According to our theory, sufficiently high levels of self-knowledge should be associated with a high test-retest stability in self-assessments, simply because signals about individuals' true types are more precise. We test this prediction in Panel C of Figure 7. The figure displays pairwise correlations in Panel C1 for the general risk question measured in the first wave of the survey and the same question measured in the second wave. Likewise, Panel C2 shows pairwise correlations for the general altruism item. As before, we display the correlations for different subsamples based on values of $\hat{\tau}$.

[19] To illustrate the value of using the ratio of between and within variance instead of either one individually, we can compare their respective percentage improvements. Performing a median split using the ratio, i.e., alongside our estimator $\hat{\tau}$, leads to improvements in coefficients of 85% for risk and 17% for altruism. Performing a median split using only the inverse within-variance leads to improvements of 27% and 8%, while using only the between-variance leads to improvements of 66% and 3%.

[20] See Appendix Table F.5 for the specification of Table 4 with the other behavioral variables as dependent variables.

[21] Another potentially interesting comparison is how subsamples with high levels of $\hat{\tau}$ compare to subsamples with low levels of $\hat{\tau}$. See Appendix Figure E.3 for the results.

Figure 7: The influence of subsetting based on $\hat{\tau}$ on coefficients, predictive power and test-retest correlations



Panel A1: Risk regression slope coefficient

Panel A2: Altruism regression slope coefficient

Panel B1: Risk regression R^2

Panel B2: Altruism regression R^2

Panel C1: Test−retest correlation general risk question

Panel C2: Test−retest correlation general altruism question

Notes: In **Panel A1** and **A2**, each dot represents the OLS-coefficient from a stacked regression. In Panel A1 (Panel A2), the general risk (altruism) question is the regression's independent variable, and risk (altruistic) behavior is the dependent variable. Each regression is run on a subsample where a percentage of subjects with the lowest estimated level of self-knowledge $\hat{\tau}$ are excluded. The x-axis denotes the respective percentage removed, i.e., the first dot denotes the sample where the 10% of subjects with the lowest $\hat{\tau}$ are removed. Shaded areas indicate 95% confidence intervals. **Panel B1** and **B2**: These panels display the R^2 values instead of the OLS coefficients from the respective regressions. **Panel C1** and **C2**: Each dot in Panel C1 represents Spearman's rank correlation coefficient between the general risk question in wave 1 and wave 2 of the survey. In Panel C2, each dot represents the wave 1 to wave 2 correlation of the general altruism question. Shaded areas indicate 95% confidence intervals.

We find that samples consisting of subjects with high $\hat{\tau}$ display much higher test-retest correlations. At the extreme, the 10% of subjects with the highest $\hat{\tau}$ show a test-retest correlation of 0.94 for the general risk question, and 0.86 for the general altruism question. Note that these results are not specific to the general risk and altruism questions. As we show in Appendix Figure E.4, we find similar effects for the domain-specific risk and altruism self-assessments: across self-assessments, focusing on subjects with high $\hat{\tau}$ increases the test-retest stability relative to the baseline.

5.3 Comparing Alternative Methods to Improve Survey Evidence

As described in Section 5.1, we made a comprehensive selection of alternative methods to improve self-assessments and included their measures in our survey. In this section, we are interested in comparing the performance of these methods with the performance of our $\hat{\tau}$ estimator.

Comparing methods using criterion 1. First, we assess the relative performance of the different methods in terms of their ability to increase the strength of the association between behavior and self-assessments (Criterion 1). In Table 5, we focus on three aspects of this relationship. In columns (2) and (3), we display the coefficient of the general risk question (column (2)) and the general altruism question (column (3)) in a stacked regression of composite risk and altruistic behavior.[22] This specification allows us to compare the average improvement of each method relative to the baseline association across the five behavioral variables for each domain.

Recall that we collected ten behavioral measures, five for risk and five for altruism. Regressing each individual behavior on either the general risk question or the general altruism question, we can check how often a particular method leads to a strictly higher coefficient relative to baseline. This measure is thus informative about the reliability of each method with respect to improving the association across different behavioral outcomes and contexts. Results are shown in columns (4) and (5), where the maximum number of improvement is five. Lastly, in columns (6) – (9), we report the quantitative range of improvements for these ten regressions by displaying the minimum and maximum improvement relative to baseline.

We find that the $\hat{\tau}$-subset estimator performs well relative to the other methods. The coefficient estimated on the sample with above median self-knowledge yields one of the highest average associations, while both coefficients estimated on the top 20% of subjects are the highest across all methods. Moreover, splitting the samples based on estimated self-knowledge is the only method that consistently increases the association across all ten behavioral outcomes (see columns (4) and (5)). Most of the other methods generally lead to increases for some variables and decreases for others. In particular, two commonly used methods, attention checks and excluding speeders, are sensitive to the specific behavioral variable used. Moreover, attention checks are sensitive to the specific check used, as using the first attention check decreases the association in eight out of the ten cases, while using the second method, or a combination of both, yields improvements in half of the cases.

Comparing methods using criterion 2. Turning to criterion 2, we investigate how each method affects the predictive power of a regression as measured in terms of R^2. Across methods and applications we find an improvement in R^2 in 55% of the 170 tests (17 methods × 5 variables × 2 domains) we run, and lower levels of R^2 for the remaining cases. In comparison to the alternative methods, subsetting on $\hat{\tau}$ works well. In both risk and altruism regressions, the R^2 from subsetting on $\hat{\tau}$ regressions are among the highest overall. In particular, subsetting on the top 20% $\hat{\tau}$ yields the highest

[22] For details on the individual regressions for each of the ten behaviors, see Appendix Table F.6 for risk and F.7 for altruism.

Table 5: Comparing methods to improve survey evidence

	Sample size	Stacked regression OLS coefficient		Number of improvements		Smallest improvement		Largest improvement	
		Risk	Altruism	Risk	Altruism	Risk	Altruism	Risk	Altruism
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Baseline	734	0.030	0.041						
		(0.002)	(0.002)						
Self-knowledge									
Above median self-knowledge	367	0.036	0.047	5	5	0.003	0.003	0.008	0.011
		(0.004)	(0.003)						
Top 20% self-knowledge	147	0.044	0.052	5	5	0.002	0.009	0.022	0.018
		(0.005)	(0.005)						
Attention									
Above median reported attention	508	0.029	0.039	2	2	-0.011	-0.008	0.009	0.004
		(0.003)	(0.003)						
Attention check 1 passed	581	0.027	0.038	1	1	-0.008	-0.008	0.003	0.005
		(0.003)	(0.003)						
Attention check 2 passed	390	0.031	0.036	3	2	-0.011	-0.020	0.010	0.009
		(0.004)	(0.003)						
Both attention checks passed	367	0.031	0.036	3	2	-0.009	-0.021	0.011	0.011
		(0.004)	(0.003)						
Effort									
Above median reported effort	383	0.033	0.040	3	1	-0.003	-0.004	0.007	0.004
		(0.003)	(0.003)						
Effort check passed	316	0.028	0.038	3	2	-0.017	-0.012	0.015	0.004
		(0.004)	(0.004)						
Reliability									
Above median reported reliability	439	0.034	0.040	4	2	-0.004	-0.008	0.015	0.006
		(0.003)	(0.003)						
Response times									
Excluding Top 10% speeders	661	0.028	0.039	1	1	-0.009	-0.003	0.006	0.002
		(0.003)	(0.002)						
Excluding slowest & fastest 25%	366	0.027	0.040	1	2	-0.016	-0.007	0.011	0.009
		(0.004)	(0.003)						
Averaging									
Averaging across domains	734	0.039	0.050	4	4	0.000	-0.004	0.018	0.025
		(0.003)	(0.003)						
Averaging responses (2 repetitions)	734	0.031	0.046	3	5	-0.001	0.003	0.002	0.009
		(0.003)	(0.002)						
Averaging responses (4 repetitions)	734	0.033	0.052	3	5	0.000	0.002	0.009	0.017
		(0.003)	(0.002)						
Further methods									
Adding demographic controls	734	0.021	0.039	0	1	-0.013	-0.003	-0.003	0.001
		(0.003)	(0.002)						
Adjustment w/ anchoring vignettes	734	0.021	0.024	1	0	-0.020	-0.028	0.006	-0.004
		(0.004)	(0.004)						
ORIV	734	0.032	0.051	3	5	-0.001	0.008	0.005	0.015
		(0.001)	(0.002)						

Notes: The table shows the impact of different methods on OLS coefficients. For details on the different methods, see Table 3 for details. Column (1) displays the resulting sample size from applying each method. Columns (2) and (3) display the OLS coefficient obtained from a stacked regression of the five individual risk behavioral variables (Column (2)) and the five altruistic behavior variables (Column (3)) on self-assessed risk and altruism, respectively. See Section 5.3 for details. Standard errors in parentheses are clustered at the subject level. Columns (4) and (5) display the number of times (out of five) that applying the respective method strictly increases the OLS coefficient when regressing behavior on self-assessments. Columns (6)-(9) display the smallest (6-7) and largest (8-9) improvements in OLS coefficients when applying the respective method.

R^2 in the stacked regression across all methods. Moreover, subsetting on $\hat{\tau}$ leads to improvements for all ten variables. For details, see Appendix Table F.8 and F.9.

Comparing methods using criterion 3. We also compare the ability of the different methods to improve test-retest correlations (criterion 3). Across methods and application (risk and altruism), the test-retest stability is higher than for baseline in 77% of our tests, and lower for the remaining tests. As with the other two criteria, subsetting on $\hat{\tau}$ performs well compared to the other methods. It leads to the second largest test-retest correlations for the general risk question, and the largest for the general altruism question compared to the other methods. For details, see Appendix Table F.10.

Interpretation. Taken together, we conclude that subsetting on $\hat{\tau}$ performs well on all three criteria relative to other commonly used methods for improving survey evidence. When interpreting these results, we note that some of the methods included in the comparison were not necessarily developed with the aim of improving OLS regressions or test-retest correlations. For instance, attention checks are regularly employed when subjects receive information, and where the research design requires them to be attentive to that specific information.[23] Similarly, anchoring vignettes are oftentimes used when comparing responses across cultures. Moreover, the methods differ widely in the demands they place on the researcher. Some measures, like response times, do not require the researcher to add any new items, while others require the addition of (multiple) items. Hence, our analysis is not meant to identify which one is the "best" overall method to improve survey evidence. Instead, it is meant to illustrate how different methods affect estimates. This may help researchers to select the method most appropriate for their individual purposes, given the trade-offs they face in designing surveys and interpreting survey responses.

In this sense, we regard our method as complementary to alternative methods. A key strength of our method is that it can be estimated on items unrelated to the main variables of interest. For instance, in our application, we use responses to personality items (the Big-Five) to improve the relationship between risk and altruism assessments and behavior. This also means that in some instances, researchers can use the method without fielding any additional items, but rely on existing measures. This is the case, for example, when researchers work with panel surveys (like the CPS, the SOEP, or the LISS panel) in which repeated measures of the Big-Five or other self-assessments are available. We provide one such example in the next section.

5.4 Robustness

5.4.1 Replicating the Survey Results Using Large Panel Data

We apply our estimator to data from the German Socio-economic Panel[24], a large, representative panel data set. This exercise explores whether the insights gained from our survey generalize to panel data that is frequently used by researchers.

To derive an estimate of tau for each SOEP respondent, we use the 2005, 2009, 2013, and 2017 waves of the SOEP, as they contain the same 15-item Big Five inventory that we fielded to the survey described in the previous section. We use the maximum number of waves available for a given respondent. This gives us $21,157$ respondents in total. For 47.4% of the respondents, we observe their responses in two waves, for 22.1% in three waves, and for 30.4% in all four waves.

[23] However, often subjects are then excluded if they fail the check. As our results show, this practice may have unintended consequences because associations between variables can become skewed in the "wrong" direction.

[24] Socio-Economic Panel (SOEP), data for years 1984–2017, version 34, SOEP, 2019, doi:10.5684/soep.v34.

Results. Due to data availability, in the SOEP we focus on the relationship between self-assessed risk and three measures of risk behavior: whether individuals own risky financial securities, whether they smoke, and whether they receive performance-based payments. As before, we regress behavior on self-assessments and compare coefficients when splitting the sample into individuals with above and below median level of estimated self-knowledge $\hat{\tau}$. For all three behavioral variables, we find a stronger relationship for above median individuals. Differences in coefficients between the above and below median sample range from 31% (smoking) to 115% (performance pay). In addition, we observe an increase in R^2 in all three regressions, ranging from 87% to 610%. For details, see Appendix Table F.11. Hence, we replicate the improvements with respect to criterion 1 and 2 of the previous section in a large, representative sample.

5.4.2 Empirical evidence on subjective scale use

One potentially important influence on self-assessments is subjective scale use. In our extension of the model (Section 2.3), we theoretically show that subjective scale use may influence reports but not the estimation of τ. To investigate this prediction empirically, we fielded a scale use module at the end of the survey's second wave.[25] The module consists of two questions, displayed in Figure 8. In the first, subjects see two differently colored circles, and are asked to assess how much darker one circle is relative to the other using the same 11-point Likert scale that is used for the self-assessments. In the second, subjects see two differently sized circles. Subjects are asked to assess how much larger they think one circle is relative to the other. Since the size and colors are the same across all subjects, we can study individual-specific response patterns in Likert scales to an objective signal and relate it to our estimate of τ and self-assessments. Specifically, we construct two variables: the direct response (0 to 10), and the absolute deviation from the scale midpoint (0 to 5).

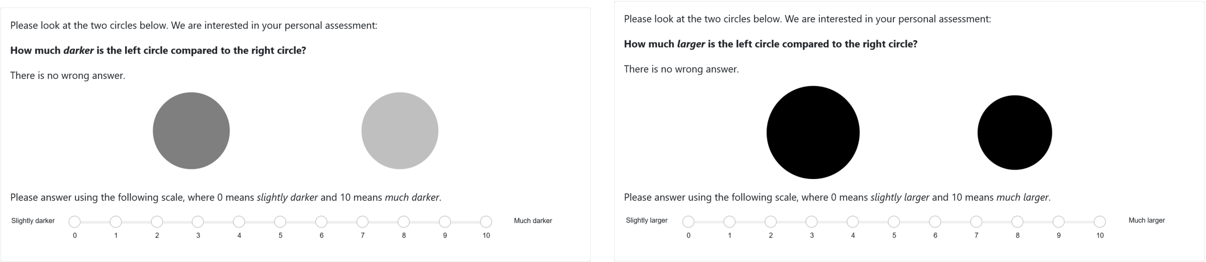
We find that subjects' responses are highly correlated across the module's two questions ($\rho = 0.59$ for direct responses, and $\rho = 0.41$ for midpoint deviations, both $p < 0.001$). Because there is no objective relationship between the size and color differences, this supports the presence of individual-specific response patterns. Indeed, when we average the two questions into two indices (one for direct responses and one for midpoint deviations), they predict responses in the general risk (direct response: $\rho = 0.21, p < 0.001$, midpoint deviation: $\rho = 0.08, p = 0.03$) and altruism questions (direct response: $\rho = 0.14, p < 0.001$, midpoint deviation: $\rho = 0.17, p < 0.001$). These results show that our subjective scale use module can be used to account or correct for individual scale use differences. Importantly, however, our estimate of τ, $\hat{\tau}$, is neither significantly correlated with the direct response index ($\rho = -0.02, p = 0.52$) nor the midpoint deviation index ($\rho = -0.05, p = 0.15$). Hence, we find empirical evidence for the prediction of our model: subjective scale use can be relevant for self-assessments, but it is uncorrelated with our estimator.

6 Conclusion

In this paper, we developed a theoretical framework of survey response behavior. We assume that respondents try to provide accurate answers but lack perfect self-knowledge, for example, because information acquisition is costly. In addition, survey responses may be affected by inaccurate beliefs

[25] This module was first included in the Bonn Family Panel (Kosse et al., 2020) in 2019.

Figure 8: Screenshots displaying the subjective scale use module



Please look at the two circles below. We are interested in your personal assessment:

How much *darker* is the left circle compared to the right circle?

There is no wrong answer.

Please answer using the following scale, where 0 means *slightly darker* and 10 means *much darker*.

(a) Size assessment

Please look at the two circles below. We are interested in your personal assessment:

How much *larger* is the left circle compared to the right circle?

There is no wrong answer.

Please answer using the following scale, where 0 means *slightly larger* and 10 means *much larger*.

(b) Color assessment

about one's self-knowledge, subjective scale use, trembling hand errors, as well as image or social desirability effects.

A key insight of the model is that we can extract individual differences in self-knowledge based on response patterns by using the ratio of the variance between characteristics and the variance for a given characteristic over time. This is important since we show that regression estimates will generally be biased in the presence of limited self-knowledge, and subsetting on subjects with high self-knowledge helps assess the direction and magnitude of the bias. Building on these findings, we suggest a consistent and unbiased estimator of self-knowledge, discuss its properties, and apply it to experimental data as well as survey data. We show that the estimator reliably identifies individual differences in the informativeness of answers in the experiment where we know the true types. Splitting the experimental sample shows that in the group estimated to have high self-knowledge, subjects' responses better predict true types, and regression estimates are less biased. We then show the usefulness of our estimator in a large survey, where it improves the coefficients and explanatory power of self-assessments for behavior. Moreover, our estimator performs as well or better than leading alternative methods that aim to improve response quality. These applications illustrate the value of distinguishing between respondents with high vs. low self-knowledge for improving survey evidence. They suggest further econometric implications for the study of measurement error and highlight the potential of integrating self-knowledge into regression frameworks.

The framework is kept deliberately simple but could be extended to allow for a richer and more realistic analysis of survey response behavior. For example, we assume that the outcome of inspecting one's individual characteristics is simply an (exogenous) signal about one's type. It would be interesting to explore cognitive (and emotional) processes involved in this introspection process in more detail, e.g., the role of limited memory and retrieval, how individuals select choice contexts to evaluate their characteristics, or how social comparison or life experience affects introspection. The framework also allows for integrating the role and meaning of response times, which could hold strong practical importance. For example, many binary choice experiments in neuroscience and psychology find that accuracy decreases as response time increases, in the sense that slower decisions are less likely to be correct (Luce, 1986; Ratcliff and McKoon, 2008; Swensson, 1972).[26] Another interesting extension of the model and its applications would consider people's actual priors about the distribution of characteristics in the population. These priors may be heterogeneous and group-specific. Eliciting and using actual priors may further improve inference from surveys. Finally, the

[26] Fudenberg, Strack, and Strzalecki (2018) and Alós-Ferrer, Fehr, and Netzer (2021) provide theoretical analyses of the relationship between response times and the accuracy of binary decisions.

model could be extended to analyze more closely whether and how a lack of self-knowledge impacts responses and inference when moving from survey items to incentive-compatible elicitation methods.

A better understanding of the survey response process may also inform the "optimal" design of research. Conditional on survey respondents' behavior, we can ask the question of how surveys or other elicitation methods should be designed to extract the maximum amount of information. Such a design perspective would consider research as a principal–agent relationship where agents participate in surveys, experiments, or related research contexts that are designed by researchers who optimize research paradigms conditional on agents' behavior. Such an approach could be used to investigate how to design survey items and response scales, when and how incentives should be given, or how to design specific modules meant to correct for expected biases.

Throughout the paper, we make several assumptions concerning the nature and stability of self-knowledge and personality traits in general, on which we would like to briefly comment on. First, we treat τ as an individual-specific characteristic that is informative across domains. This allows us to estimate and assign a unique estimate of self-knowledge (e.g., based on the Big-Five), and to use this estimate for applications in different domains (such as risk and altruism). Using our survey evidence, we show that the estimates of τ are in fact similar when estimated on different characteristics: there is a significant positive relationship between values of τ that are estimated using our Big-Five module with estimates of τ when using the set of self-assessments concerning risk attitudes and altruism (see Appendix D for details). Our interpretation is not that self-knowledge is identical across domains, however. Rather, we consider it a "latent factor" that carries informational value across domains. Importantly, for practical applications, no assumption about the absence of domain specificity is needed. A researcher interested in estimating a specific relation in a given domain may use an estimator for τ that is built on data that uses only information about that domain.

Second, the model assumes no learning in the process of survey response and stability of personality traits over longer periods of time. The assumption of "forgetting signals" about one's characteristics is made for simplicity. Extending the model to incorporate learning from previous signals would complicate the model and would not allow us to characterize a closed-form solution for τ or the estimator of τ. Whether this type of learning affects response behavior and the usefulness of our method is ultimately an empirical question. In this respect, we offer two empirical findings that provide suggestive evidence in favor of the model's assumption. In the experiment, we can study improvements over time by comparing their estimates with correct states. Our findings suggest that little learning takes place.[27] Turning to our survey evidence, recall that we ran the survey in two waves and can therefore estimate τ for each participant separately in wave 1 and wave 2, respectively. We find that both estimates are highly correlated at the individual level ($\rho = 0.62, p < 0.001$), suggesting little variance over time.

With respect to long-run stability of personality traits and τ, work in personality psychology suggests that important traits, such as the Big-5 are relatively stable over time (see, e.g., Caspi, Roberts, and Shiner, 2005; Cobb-Clark and Schurer, 2012). While this does not mean that personality is completely time-invariant, it does suggest that for reasonable time frames, it is a fair approximation to

[27] Specifically, recall that subjects in the experiment see distinct images, which are repeated 4 times. Within each repetition, the sequence of images is randomized. For the first set of images, the average absolute difference between estimated and correct categories is 1.34. For the second to fourth, differences are 1.39, 1.36, and 1.35, respectively. Accordingly, subjects are not getting better at estimating categories over time.

assume stability. In this respect, our approach and method should be viewed as meaningful for time spans where traits are relatively stable, a precondition that equally applies to the whole research program on the effects of preferences and personality on relevant socio-economic outcomes. Importantly, when using our short survey module to estimate τ, no assumption about the stability of traits above and beyond the duration of the survey is needed.

In conclusion, we have introduced a simple method that can be used without fielding new items in existing panels. For researchers conducting new surveys, our short module can be directly incorporated, or the underlying approach can be adapted to their own survey items. Furthermore, our study provides a comprehensive evaluation of alternative methodologies aimed at enhancing the quality of survey data, accompanied by empirical comparisons of their relative effectiveness. These findings offer researchers practical guidance for selecting the most suitable method for their specific research objectives.

References

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz (2011). "Personality Psychology and Economics." In *Handbook of the Economics of Education*. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, ed. vol. 4, 2008. Amsterdam, Netherlands: Elsevier B.V., 1–181. [1]

Alós-Ferrer, Carlos, Ernst Fehr, and Nick Netzer (2021). "Time Will Tell: Recovering Preferences When Choices Are Noisy." *Journal of Political Economy* 129 (6): 1828–77. [36]

Andre, Peter, Teodora Boneva, Felix Chopra, and Armin Falk (2024). "Misperceived Social Norms and Willingness to Act Against Climate Change." *Review of Economics and Statistics*, (6): 1–46. [25]

Beauchamp, Jonathan, Daniel J. Benjamin, David I. Laibson, and Christopher F. Chabris (2020). "Measuring and Controlling for the Compromise Effect When Estimating Risk Preference Parameters." *Experimental Economics* 23: 1069–99. [5]

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 (3): 203–37. [5]

Bénabou, Roland, Armin Falk, Luca Henkel, and Jean Tirole (2023). "Eliciting Moral Preferences under Image Concerns: Theory and Experiment." *Working Paper*, [4]

Benjamin, Daniel J, Kristen Cooper, Ori Heffetz, Miles S Kimball, and Jiannan Zhou (2023). "Adjusting for Scale-Use Heterogeneity in Self-Reported Well-Being." *Working Paper*, [46]

Berinsky, Adam J., Michele F. Margolis, Michael W. Sances, and Christopher Warshaw (2021). "Using Screeners to Measure Respondent Attention on Self-Administered Surveys: Which Items and How Many?" *Political Science Research and Methods* 9 (2): 430–37. [5, 28]

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch (2014). "hroot: Hamburg Registration and Organization Online Tool." *European Economic Review* 71: 117–20. [19, 69]

Bolsinova, Maria, Paul de Boeck, and Jesper Tijmstra (2017). "Modelling Conditional Dependence between Response Time and Accuracy." *Psychometrika* 82 (4): 1126–48. [5]

Bound, John, Charles Brown, and Nancy Mathiowetz (2001). "Measurement Error in Survey Data." In *Handbook of Econometrics*. James J. Heckman and Edward Leamer, ed. vol. 5, Amsterdam, Netherlands: Elsevier Science B.V. Chap. 59, 3705–843. [4]

Bradburn, Norman M., Seymour Sudman, and Brian Wansink (2004). *Asking Questions: The Definitive Guide to Questionnaire Design*. San Francisco: Jossey-Bass, 448 pp. [5]

Caliendo, Marco, Frank Fossen, and Alexander S. Kritikos (2014). "Personality Characteristics and the Decisions to Become and Stay Self-Employed." *Small Business Economics* 42 (4): 787–814. [25]

Camerer, Colin, and Dan Lovallo (1999). "Overconfidence and Excess Entry: An Experimental Approach." *American Economic Review* 89 (1): 306–18. [8]

Camerer, Colin F., and Robin M. Hogarth (1999). "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework." *Journal of Risk and Uncertainty* 19 (1-3): 7–42. [10]

Caplin, Andrew, Dániel Csaba, John Leahy, and Oded Nov (2020). "Rational Inattention, Competitive Supply, and Psychometrics." *Quarterly Journal of Economics* 135 (3): 1681–724. [4]

Caplin, Andrew, and Mark Dean (2015). "Revealed Preference, Rational Inattention, and Costly Information Acquisition." *American Economic Review* 105 (7): 2183–203. [4]

Caspi, Avshalom, Brent W. Roberts, and Rebecca L. Shiner (2005). "Personality Development: Stability and Change." *Annual Review of Psychology* 56 (1): 453–84. [37, 54]

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016). "oTree-An Open-Source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance* 9: 88–97. [19]

Chen, Yuanyuan, Shuaizhang Feng, James J. Heckman, and Tim Kautz (2020). "Sensitivity of Self-Reported Noncognitive Skills to Survey Administration Conditions." *Proceedings of the National Academy of Sciences* 117 (2): 931–35. [4]

Cobb-Clark, Deborah A., and Stefanie Schurer (2012). "The Stability of Big-Five Personality Traits." *Economics Letters* 115 (1): 11–15. [37, 54]

Curran, Paul G. (2016). "Methods for the Detection of Carelessly Invalid Responses in Survey Data." *Journal of Experimental Social Psychology* 66 (9): 4–19. [5, 28]

Cvitanić, Jakša, Dražen Prelec, Blake Riley, and Benjamin Tereick (2019). "Honesty Via Choice-Matching." *American Economic Review: Insights* 1 (2): 179–92. [4]

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): 522–50. [3, 25, 26]

Dohmen, Thomas, and Tomáš Jagelka (2023). "Individual-Specific Reliability of Self-Assessed Measures of Economic Preferences and Personality Traits." *Journal of Political Economy Microeconomics*, (forthcoming): [5, 28]

Dougherty, Christopher R. S. (2016). *Introduction to Econometrics*. Fifth edition. Oxford New York: Oxford University Press. [11]

Drerup, Tilman, Benjamin Enke, and Hans Martin von Gaudecker (2017). "The precision of subjective data and the explanatory power of economic models." *Journal of Econometrics* 200 (2): 378–89. [5]

Edwards, Michael C. (2009). "An Introduction to Item Response Theory Using the Need for Cognition Scale." *Social and Personality Psychology Compass* 3 (4): 507–29. [5]

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde (2018). "Global Evidence on Economic Preferences." *Quarterly Journal of Economics* 133 (4): 1645–92. [1, 3, 25, 26]

Falk, Armin, David Huffman, and Uwe Sunde (2006a). "Do I Have What It Takes? Equilibrium Search with Type Uncertainty and Non-Participation." IZA Discussion Paper 2531. Bonn, Germany: IZA Institute for the Study of Labor. [4]

Falk, Armin, David Huffman, and Uwe Sunde (2006b). "Self-Confidence and Search." IZA Discussion Paper 2525. Bonn, Germany: IZA Institite for the Study of Labor. [4]

Falk, Armin, and Florian Zimmermann (2013). "A Taste for Consistency and Survey Response Behavior." *CESifo Economic Studies* 59 (1): 181–93. [4]

Falk, Armin, and Florian Zimmermann (2017). "Consistency as a Signal of Skills." *Management Science* 63 (7): 2049–395. [4]

Fang, Ximeng, Timo Freyer, Chui-Yee Ho, Zihua Chen, and Lorenz Goette (2022). "Prosociality Predicts Individual Behavior and Collective Outcomes in the COVID-19 Pandemic." *Social Science & Medicine* 308 (9): 115192. [25]

Fischbacher, Urs (2007). “Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments.” *Experimental Economics* 10 (2): 171–78. [69]

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki (2018). “Speed, Accuracy, and the Optimal Timing of Choices.” *American Economic Review* 108 (12): 3651–84. [36]

Gerlitz, Jean-Yves, and Jürgen Schupp (2005). “Zur Erhebung Der Big-Five-Basierten Persönlichkeitsmerkmale Im SOEP. Dokumentation Der Instrumententwicklung BFI-S Auf Basis Des SOEP-Pretests 2005.” Research Notes 4. Berlin: Deutsches Institut für Wirtschaftsforschung (DIW), 1–44. [84]

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): 1826–63. [4, 28]

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): 411–82. [1]

Heeb, Florian, Julian F Kölbel, Falko Paetzold, and Stefan Zeisberger (2023). “Do Investors Care about Impact?” *Review of Financial Studies* 36 (5): Tarun Ramadorai, ed., 1737–87. [25]

Hyslop, Dean R., and Guido W. Imbens (2001). “Bias from Classical and Other Forms of Measurement Error.” *Journal of Business and Economic Statistics* 19 (4): 475–81. [4]

Kimball, Miles S., Claudia R. Sahm, and Matthew D. Shapiro (2008). “Imputing Risk Tolerance from Survey Responses.” *Journal of the American Statistical Association* 103 (483): 1028–38. [5]

King, Gary, and Jonathan Wand (2007). “Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes.” *Political Analysis* 15 (1): 46–66. [28]

Kosse, Fabian, Thomas Deckers, Pia Pinger, Hannah Schildberg-Hörisch, and Armin Falk (2020). “The Formation of Prosociality: Causal Evidence on the Role of Social Environment.” *Journal of Political Economy* 128 (2): 434–67. [35]

Kosse, Fabian, and Michela M. Tincani (2020). “Prosociality Predicts Labor Market Success around the World.” *Nature Communications* 11 (1): 5298. [25]

Kyllonen, Patrick, and Jiyun Zu (2016). “Use of Response Time for Measuring Cognitive Ability.” *Journal of Intelligence* 4 (4): 14. [5]

Luce, R. Duncan (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press. [36]

Malmendier, Ulrike, and Geoffrey Tate (2005). “CEO Overconfidence and Corporate Investment.” *Journal of Finance* 60 (6): 2661–700. [8]

Matějka, Filip, and Alisdair McKay (2015). “Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model.” *American Economic Review* 105 (1): 272–98. [4]

Meade, Adam W., and S. Bartholomew Craig (2012). “Identifying Careless Responses in Survey Data.” *Psychological Methods* 17 (3): 437–55. [5, 28]

Non, Arjan, Ingrid Rohde, Andries De Grip, and Thomas Dohmen (2022). “Mission of the Company, Prosocial Attitudes and Job Preferences: A Discrete Choice Experiment.” *Labour Economics* 74 (1): 102087. [25]

Prelec, Dražen (2004). “A Bayesian Truth Serum for Subjective Data.” *Science* 306 (5695): 462–66. [4, 10]

Ratcliff, Roger, and Gail McKoon (2008). “The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks.” *Neural Computation* 20 (4): 873–922. [36]

Schupp, J., and J.-Y. Gerlitz (2008). “Big Five Inventory-SOEP (BFI-S).” *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*, [26, 58]

Schwarz, Norbert (2007). “Cognitive Aspects of Survey Methodology.” *Applied Cognitive Psychology* 21: 277–87. [5]

Sims, Christopher (1998). “Stickiness.” *Carnegie-Rochester Conference Series on Public Policy* 49: 317–56. [4]

Sims, Christopher (2003). “Implications of Rational Inattention.” *Journal of Monetary Economics* 50 (3): 665–90. [1, 4]

Smith, Vernon L. (1976). “Experimental Economics: Induced Value Theory.” *American Economic Review* 66 (2): 274–79. [10]

Stantcheva, Stefanie (2023). "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15 (1): 205–34. [1]

Storey, Hannah (2016). "Animal Pictograms and Sleep Infographic." [69]

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz (1996). *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass, 304 pp. [5]

Swensson, Richard G. (1972). "The Elusive Tradeoff: Speed vs Accuracy in Visual Discrimination Tasks." *Perception & Psychophysics* 12 (1A): 16–32. [36]

A Proofs

Proof of Proposition 1. Define $\gamma = \tau/(1+\tau)$ and $\tilde{\gamma} = \tilde{\tau}/(1+\tilde{\tau})$ and recall that $\tilde{\tau}$ was the subjective level of self-knowledge. By assumption, $\gamma_i, \tilde{\gamma}_i$ are drawn across agents independently of their characteristic type θ_i. Denote by $\bar{r} = \frac{1}{N}\sum_{i=1}^{n} r_i$ the average response. First, as $\mathbb{E}\left[r\right] = \bar{\theta}$ we have that a.s. $\lim_{N\to\infty} \bar{r} = \bar{\theta}$. Denote by $\bar{y} = \beta_0 + \beta_1\bar{\theta}$ the expected outcome in the population. We have that

$$y_i - \bar{y} = \beta_1\left(\theta_{i1} - \bar{\theta}_1 + \epsilon_i\right) .$$

Recall that $r - \bar{\theta} = \tilde{\gamma}(x - \bar{\theta})$ and that $\mathbb{E}\left[(\theta - \bar{\theta})^2\right] = \sigma^2$. As ϵ is independent of everything else, we have

$$\mathbb{E}\left[(r-\bar{\theta})(y-\bar{y})\right] = \mathbb{E}\left[\tilde{\gamma}(x-\bar{\theta})\beta_1(\theta-\bar{\theta})\right] = \mathbb{E}\left[\tilde{\gamma}(\theta-\bar{\theta})\beta_1(\theta-\bar{\theta})\right] = \beta_1\mathbb{E}\left[\tilde{\gamma}\right]\sigma^2 .$$

Furthermore, we have that

$$\begin{aligned}\mathbb{E}\left[(r-\bar{\theta})^2\right] &= \mathbb{E}\left[\tilde{\gamma}^2(x-\bar{\theta})^2\right] = \mathbb{E}\left[\tilde{\gamma}^2\,\mathbb{E}\left[(x-\bar{\theta})^2 \mid \tilde{\tau}\right]\right]\\ &= \mathbb{E}\left[\tilde{\gamma}^2\mathbb{E}\left[(\sigma^2+\sigma^2/\tau)\right]\tilde{\tau}\right] = \sigma^2\mathbb{E}\left[\tilde{\gamma}^2\mathbb{E}\left[1/\gamma \mid \tilde{\tau}\right]\right]\\ &= \sigma^2\mathbb{E}\left[\mathbb{E}\left[\tilde{\gamma}^2/\gamma \mid \tilde{\tau}\right]\right] = \sigma^2\mathbb{E}\left[\tilde{\gamma}^2/\gamma\right]\\ &= \sigma^2\left(\mathbb{E}\left[\tilde{\gamma}\right]\times\mathbb{E}\left[\tilde{\gamma}/\gamma\right] + \text{cov}\left(\tilde{\gamma},,\tilde{\gamma}/\gamma\right)\right)\\ &= \sigma^2\,\mathbb{E}\left[\tilde{\gamma}\right]\left(\mathbb{E}\left[\tilde{\gamma}/\gamma\right] + \frac{\text{cov}\left(\tilde{\gamma},\tilde{\gamma}/\gamma\right)}{\mathbb{E}\left[\tilde{\gamma}\right]}\right) .\end{aligned}$$

By the law of large numbers, we have that a.s.

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left(r_{i1}-\bar{r}\right)\left(y_i-\bar{y}\right) &= \mathbb{E}\left[(r-\bar{\theta})(y-\bar{y})\right]\\ \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left(r_{i1}-\bar{r}\right)^2 &= \mathbb{E}\left[(r-\bar{\theta})^2\right] .\end{aligned}$$

The regression estimate $\hat{\beta}_1$ thus satisfied

$$\lim_{N\to\infty}\hat{\beta}_1 = \lim_{N\to\infty}\frac{\sum_{i=1}^{N}\left(r_i-\bar{r}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{N}\left(r_{i1}-\bar{r}\right)^2} = \beta_1\frac{1}{\mathbb{E}\left[\tilde{\gamma}/\gamma\right]+\frac{\text{cov}(\tilde{\gamma},\tilde{\gamma}/\gamma)}{\mathbb{E}[\tilde{\gamma}]}} .$$

Thus, if the agents' subjective level of self-knowledge are correct, i.e. $\tau_i = \tilde{\tau}_i$, then $\lim_{N\to\infty}\hat{\beta}_1 = \beta_1$. □

Proof of Proposition 2. Let G be the distribution of objective self-knowledge. We note that

$$\frac{\tilde{\gamma}}{\gamma} = \frac{\frac{\tau}{1/\alpha+\tau}}{\frac{\tau}{1+\tau}} = \frac{1+\tau}{1/\alpha+\tau} .$$

Using this we have that

$$\frac{\lim_{N\to\infty}\hat{\beta}_1}{\beta_1} = \frac{\mathbb{E}[\tilde{\gamma}\,|\,\tau \geq \underline{\tau}]}{\mathbb{E}[\tilde{\gamma}^2/\gamma\,|\,\tau \geq \underline{\tau}]} = \frac{\mathbb{E}\left[\tilde{\gamma} \times \mathbf{1}\{\tau \geq \underline{\tau}\}\right]}{\mathbb{E}\left[\tilde{\gamma}^2/\gamma \times \mathbf{1}\{\tau \geq \underline{\tau}\}\right]} = \frac{\int_{\underline{\tau}}^{\infty} \frac{\tau}{\tau+1/\alpha} dG(\tau)}{\int_{\underline{\tau}}^{\infty} \frac{\tau(1+\tau)}{(\tau+1/\alpha)^2} dG(\tau)}.$$

Taking the derivative of the logarithm of the above term with respect to $\underline{\tau}$ yields

$$\begin{aligned}\frac{\partial}{\partial \underline{\tau}} \log\left(\frac{\lim_{N\to\infty}\hat{\beta}_1}{\beta_1}\right) &= \frac{\frac{\underline{\tau}(1+\underline{\tau})}{(\underline{\tau}+1/\alpha)^2}}{\int_{\underline{\tau}}^{\infty} \frac{\tau(1+\tau)}{(\tau+1/\alpha)^2} dG(\tau)} - \frac{\frac{\underline{\tau}}{\underline{\tau}+1/\alpha}}{\int_{\underline{\tau}}^{\infty} \frac{\tau}{\tau+1/\alpha} dG(\tau)} \\ &= \frac{\frac{\underline{\tau}}{\underline{\tau}+1/\alpha} \times \frac{1+\underline{\tau}}{\underline{\tau}+1/\alpha}}{\int_{\underline{\tau}}^{\infty} \frac{\tau}{\tau+1/\alpha} \times \frac{1+\tau}{\tau+1/\alpha} dG(\tau)} - \frac{\frac{\underline{\tau}}{\underline{\tau}+1/\alpha}}{\int_{\underline{\tau}}^{\infty} \frac{\tau}{\tau+1/\alpha} dG(\tau)}.\end{aligned}$$

As for $\alpha \geq 1$ the term $\frac{1+\tau}{\tau+1/\alpha}$ is decreasing in τ we get that

$$\frac{\frac{\underline{\tau}}{\underline{\tau}+1/\alpha} \times \frac{1+\underline{\tau}}{\underline{\tau}+1/\alpha}}{\int_{\underline{\tau}}^{\infty} \frac{\tau}{\tau+1/\alpha} \times \frac{1+\tau}{\tau+1/\alpha} dG(\tau)} > \frac{\frac{\underline{\tau}}{\underline{\tau}+1/\alpha} \times \frac{1+\underline{\tau}}{\underline{\tau}+1/\alpha}}{\int_{\underline{\tau}}^{\infty} \frac{\tau}{\tau+1/\alpha} \times \frac{1+\underline{\tau}}{\underline{\tau}+1/\alpha} dG(\tau)} = \frac{\frac{\underline{\tau}}{\underline{\tau}+1/\alpha}}{\int_{\underline{\tau}}^{\infty} \frac{\tau}{\tau+1/\alpha} dG(\tau)}.$$

Hence, we have that

$$\frac{\partial}{\partial \underline{\tau}} \log\left(\frac{\lim_{N\to\infty}\hat{\beta}_1}{\beta_1}\right) > 0$$

and as

$$\frac{\lim_{N\to\infty}\hat{\beta}_1}{\beta_1} = \frac{\mathbb{E}\left[\tilde{\gamma}\right]}{\mathbb{E}\left[\tilde{\gamma} \times \frac{\tilde{\gamma}}{\gamma}\right]} = \frac{\mathbb{E}\left[\tilde{\gamma}\right]}{\mathbb{E}\left[\tilde{\gamma} \times \frac{1+\tau}{1/\alpha+\tau}\right]} \leq 1$$

we get that the bias is decreasing. The case $\alpha < 1$ follows from an analogous argument. □

Proof of Proposition 3. We have that a.s. the limit of the coefficient of determination when restricting to $\tau \geq \underline{\tau}$ is given by

$$\begin{aligned}\lim_{N\to\infty} R^2 &= \frac{\mathbb{E}\big[(r-\bar{\theta})(y-\bar{y})\,\big|\,\tau \geq \underline{\tau}\big]}{\sqrt{\mathbb{E}\big[(r-\bar{\theta})^2\,|\,\tau \geq \underline{\tau}\big]} \times \sqrt{\mathbb{E}[(y-\bar{y})^2\,|\,\tau \geq \underline{\tau}]}} = \frac{\beta_1 \mathbb{E}[\tilde{\gamma}\,|\,\tau \geq \underline{\tau}]\,\sigma^2}{\sqrt{\sigma^2 \mathbb{E}[\tilde{\gamma}^2/\gamma\,|\,\tau \geq \underline{\tau}]} \times \sqrt{\beta_1^2\sigma^2 + \sigma_\epsilon^2}} \\ &= \frac{\mathbb{E}[\tilde{\gamma}\,|\,\tau \geq \underline{\tau}]}{\sqrt{\mathbb{E}[\tilde{\gamma}^2/\gamma\,|\,\tau \geq \underline{\tau}]\,(1+\frac{\sigma_\epsilon^2}{\beta_1^2\sigma^2})}} = \sqrt{\frac{\mathbb{E}[\tilde{\gamma}\,|\,\tau \geq \underline{\tau}]}{\frac{\mathbb{E}[\tilde{\gamma}^2/\gamma\,|\,\tau\geq\underline{\tau}]}{\mathbb{E}[\tilde{\gamma}\,|\,\tau\geq\underline{\tau}]}}} \times \frac{1}{\sqrt{1+\frac{\sigma_\epsilon^2}{\beta_1^2\sigma^2}}}.\end{aligned}$$

Consider the case $\alpha \geq 1$ and recall that we have established in Proposition 2 that $\frac{\mathbb{E}[\tilde{\gamma}^2/\gamma\,|\,\tau\geq\underline{\tau}]}{\mathbb{E}[\tilde{\gamma}\,|\,\tau\geq\underline{\tau}]}$ is decreasing in $\underline{\tau}$. As $\tilde{\gamma} = \frac{\tau}{1/\alpha+\tau}$ is increasing in τ it follows that R^2 is increasing. □

Proof of Theorem 1. We will prove the result in the more general setting with subjective self-knowledge and scale use as introduced in Sections 2.2 and Appendix Section B.1, respectively. The case without subjective self-knowledge and scale use stated in the basic version of the model corresponds to the special case where $\tilde{\tau}_i = \tau_i$ and $\phi_i = 1$.

Throughout the proof, we fix $\tau_i, \tilde{\tau}_i > 0$ and $\phi_i \in (0, 1]$. The answer of agent i when asked for the

t^{th} time about the k^{th} characteristic is given by

$$r_{ikt} = (1-\phi_i)\, c + \phi_i \, \frac{\bar{\theta} + \tilde{\tau}_i \, x_{ikt}}{1+\tilde{\tau}_i} \, .$$

By assumption, there exist independent, standard normally distributed random variables $\epsilon_{ikt}, \eta_{ik}$ such that

$$\begin{aligned} x_{ikt} &= \theta_{ik} + \frac{\sigma}{\sqrt{\tau_i}} \epsilon_{ikt} \, , \\ \theta_{ik} &= \bar{\theta} + \sigma \, \eta_{ik} \, . \end{aligned}$$

Plugging into the equation for the agent's responses yields that

$$r_{ikt} = (1-\phi_i)\, c + \phi_i \left(\bar{\theta} + \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i} \sigma \left[\eta_{ik} + \frac{\epsilon_{ikt}}{\sqrt{\tau_i}} \right] \right) \, . \tag{12}$$

Denote agent i's average answer for question k by $\bar{r}_{ik} = \frac{1}{T}\sum_{t=1}^{T} r_{ikt}$, her average answer over all questions by $\bar{r}_i = \frac{1}{K}\sum_{k=1}^{K} \bar{r}_{ik}$, and similarly $\bar{x}_{ik} = \frac{1}{T}\sum_{t=1}^{T} x_{ikt}$, $\bar{\epsilon}_{ik} = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{ikt}$, $\bar{x}_i = \frac{1}{K}\sum_{k=1}^{K} \bar{x}_{ik}$, $\bar{\epsilon}_i = \frac{1}{K}\sum_{k=1}^{K} \bar{\epsilon}_{ik}$, and $\bar{\eta}_i = \frac{1}{K}\sum_{k=1}^{K} \bar{\eta}_{ik}$. We have that

$$\frac{r_{ikt} - \bar{r}_{ik}}{\phi_i} = \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}(x_{ikt} - \bar{x}_{ik}) = \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\frac{\sigma}{\sqrt{\tau_i}}(\epsilon_{ikt} - \bar{\epsilon}_{ik}) \, . \tag{13}$$

Similarly, we get that

$$\begin{aligned} \frac{\bar{r}_{ik} - \bar{r}_i}{\phi_i} &= \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}(\bar{x}_{ik} - \bar{x}_i) = \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\left(\left(\theta_{ik} + \frac{\sigma}{\sqrt{\tau_i}}\bar{\epsilon}_{ik}\right) - \left(\bar{\theta}_i + \frac{\sigma}{\sqrt{\tau_i}}\bar{\epsilon}_i\right) \right) \\ &= \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\left(\left(\theta_{ik} - \bar{\theta}_i\right) + \frac{\sigma}{\sqrt{\tau_i}}\left(\bar{\epsilon}_{ik} - \bar{\epsilon}_i\right) \right) \\ &= \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\left(\sigma\left(\eta_{ik} - \bar{\eta}_i\right) + \frac{\sigma}{\sqrt{\tau_i}}\left(\bar{\epsilon}_{ik} - \bar{\epsilon}_i\right) \right) \, . \end{aligned} \tag{14}$$

We first show that

$$A := \frac{(1+\tilde{\tau}_i)^2}{\tilde{\tau}_i^2 \sigma^2} \tau_i \sum_{k=1}^{K}\sum_{t=1}^{T} \left(\frac{r_{ikt} - \bar{r}_{ik}}{\phi_i} \right)^2$$

is χ^2 distributed with $K(T-1)$ degrees of freedom. It follows from Equation (13) that

$$A = \sum_{k=1}^{K}\sum_{t=1}^{T} \left(\epsilon_{ikt} - \bar{\epsilon}_{ik}\right)^2 \, .$$

We have that $A_k := \sum_{t=1}^{T} \left(\epsilon_{ikt} - \bar{\epsilon}_{ik}\right)^2$ is χ^2 distributed with $T-1$ degrees of freedom as it equals the sum of the squared distance of i.i.d. normals from the mean. As $A_k, A_{k'}$ are independent for $k' \neq k$ and $A = \sum_{k=1}^{K} A_k$, it follows that A is χ^2 distributed with $\sum_{k=1}^{K}(T-1) = K(T-1)$ degrees of freedom.

We next argue that

$$B := \frac{(1+\tilde{\tau}_i)^2}{\tilde{\tau}_i^2\sigma^2}\frac{1}{1+\frac{1}{T\tau_i}}\sum_{k=1}^{K}\left(\frac{\bar{r}_{ik}-\bar{r}_i}{\phi_i}\right)^2$$

is χ^2 distributed with $K-1$ degrees of freedom. It follows from Equation (14) that

$$B = \sum_{k=1}^{K}\left(\lambda_{ik}-\bar{\lambda}_i\right)^2$$

where $\lambda_{ik} = \frac{1}{\sqrt{1+\frac{1}{T\tau_i}}}(\eta_{ik}+\frac{1}{\sqrt{\tau_i}}\bar{\epsilon}_{ik})$. As

$$\text{var}(\lambda_{ik}) = \frac{\text{var}(\eta_{ik})+\frac{1}{\tau_i}\text{var}(\bar{\epsilon}_{ik})}{1+\frac{1}{T\tau_i}} = \frac{1+\frac{1}{\tau_i}\text{var}\left(\frac{1}{T}\sum_{t=1}^{t}\epsilon_{ikt}\right)}{1+\frac{1}{T\tau_i}} = 1\,,$$

the random variables $(\lambda_{ik})_{k\in\{1,\ldots,K\}}$ are i.i.d. standard normal random variables. Again, as $\lambda_{ik}, \lambda_{ik'}$ are independent for $k \neq k'$, it follows that B is χ^2 distributed with $K-1$ degrees of freedom.

Next, recall that for the Normal distribution, the sample variance $\frac{1}{T-1}\sum_{t=1}^{T}\left(\epsilon_{ikt}-\bar{\epsilon}_{ik}\right)^2$ is independent of the sample mean $\bar{\epsilon}_{ik}$. As η is independent of ϵ it follows that $\sum_{t=1}^{T}\left(\epsilon_{ikt}-\bar{\epsilon}_{ik}\right)^2$ and $\lambda_{ik} = \frac{1}{\sqrt{1+\frac{1}{T\tau_i}}}(\eta_{ik}+\frac{1}{\sqrt{\tau_i}}\bar{\epsilon}_{ik})$ are independent. This implies that A and B are independent. As A and B are independently χ^2 distributed it follows that

$$F_i := \frac{\frac{1}{K-1}B}{\frac{1}{K(T-1)}A}$$

follows an F-distribution with parameters $K-1$ and $K(T-1)$.[28] Recall that in Equation (9), we defined $\hat{\tau}_i$.

$$\hat{\tau}_i = \frac{\frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{r}_{ik}-\bar{r}_i\right)^2}{\frac{1}{K(T-1)-2}\sum_{k=1}^{K}\sum_{t=1}^{T}\left(r_{ikt}-\bar{r}_{ik}\right)^2} - \frac{1}{T}$$

Plugging in the definition of A and B yields that

$$\begin{aligned}
\hat{\tau}_i + \frac{1}{T} &= \frac{K(T-1)-2}{K(T-1)}\frac{\frac{1}{K-1}\sum_{k=1}^{K}\left(\frac{\bar{r}_{ik}-\bar{r}_i}{\phi_i}\right)^2}{\frac{1}{K(T-1)}\sum_{k=1}^{K}\sum_{t=1}^{T}\left(\frac{r_{ikt}-\bar{r}_{ik}}{\phi_i}\right)^2} \\
&= \frac{K(T-1)-2}{K(T-1)}\frac{\frac{1}{K-1}B\,\frac{\tilde{\tau}_i^2\sigma^2}{(1+\tilde{\tau}_i)^2}\left(1+\frac{1}{T\tau_i}\right)}{\frac{1}{K(T-1)}A\,\frac{\tilde{\tau}_i^2\sigma^2}{(1+\tilde{\tau}_i)^2}\frac{1}{\tau_i}} \\
&= \frac{K(T-1)-2}{K(T-1)} \times \tau_i\left(1+\frac{1}{T\tau_i}\right) \times \frac{\frac{1}{K-1}B}{\frac{1}{K(T-1)}A} \\
&= \frac{K(T-1)-2}{K(T-1)} \times \left(\tau_i+\frac{1}{T}\right) \times F_i\,.
\end{aligned}$$

This establishes the first part of the theorem, i.e., Equation (10). Part 2 of the Theorem follows as

[28] See https://en.wikipedia.org/wiki/F-distribution#Characterization (accessed on June 17, 2021).

$\mathbb{E}[F_i] = \frac{K(T-1)}{K(T-1)-2}$.[29] Part 3 follows as

$$\text{var}(F_i) = \mathbb{E}[F_i]^2 \, \frac{2((K-1)+K(T-1)-2)}{(K-1)(K(T-1)-4)} \, .$$

To prove Part 4, observe that Equation (11) is decreasing in T, and thus an upper bound is given by setting $T = 2$.

$$\begin{aligned} \sqrt{\mathbb{E}[(\hat{\tau}_i - \tau_i)^2 \,|\, \tau_i]} &\leq \left(\tau_i + \frac{1}{2}\right) \sqrt{\frac{2((K-1)+K-2)}{(K-1)(K-4)}} = \left(\tau_i + \frac{1}{2}\right) \sqrt{\frac{4K-6}{(K-1)(K-4)}} \\ &\leq \left(\tau_i + \frac{1}{2}\right) \sqrt{\frac{4}{K-4}} = (2\tau_i + 1) \, \frac{1}{\sqrt{K-4}} \, . \end{aligned}$$

This establishes the result. □

B Model Extensions

In this section, we first extend our basic model to allow for strategic motives that bias agents' responses in a particular direction. Specifically, we incorporate parameterized forms of social desirability effects and subjective scale use. We then investigate the role of trembling hand errors, i.e., errors caused by normally distributed noise.

B.1 Subjective Scale Use

Empirical research typically assumes that individuals who want to express the same level of agreement or disagreement with respect to a particular survey item will respond in the exact same way. For example, two respondents intending to express the exact same willingness to take risks on a Likert scale would be expected to choose the exact same answer category. However, if response scales are subjectively interpreted, responses may differ (Benjamin et al., 2023). Hence, the mapping from an intended response to some scale may depend on individual-specific notions of how to express a given level of agreement or disagreement. We suggest a simple way to model this kind of subjective scale use and show that it affects responses in general but not the estimation approach for τ suggested by Equation (6). In Section 5.4.2, we also provide a simple survey module on subjective scale use that enables correcting for scale use at the individual level.

To formalize subjective scale use, assume that an agent has arrived at her *intended* report and now needs to map it to an *actual* report r on an answering scale. This mapping may be individual-specific in the sense that some agents may use more "extreme" answers while others use more "moderate" answers to express the same information. For a given intended response, therefore, two agents may come up with different actual responses. We assume that the agent's response is scaled away from some point $c \in \mathbb{R}$, e.g., the center of the scale, by a factor $\phi \in (0, 1]$. The report and its expected value (corresponding to Equations (B.1) and (3), respectively) are then given by

$$r = (1-\phi)\, c + \phi \left(\frac{\bar{\theta} + \tau\, x}{1+\tau} \right) \qquad \text{and} \qquad \mathbb{E}[r \,|\, \theta] = (1-\phi)\, c + \phi \left(\frac{\bar{\theta} + \tau\, \theta}{1+\tau} \right) .$$

[29] See https://en.wikipedia.org/wiki/F-distribution (accessed on June 17, 2021).

Depending on ϕ, actual responses may thus be pushed towards the center of the scale, rendering the interpretation of responses more difficult. This holds in particular if ϕ is systematically correlated with underlying types (such as preferences) or group characteristics under study (such as gender or socioeconomic status).

The between-variance (corresponding to Equation (4)) becomes

$$\begin{aligned}\sigma^2_{\text{between}} &= \text{var}(\mathbb{E}[R \mid \theta]) = \text{var}\left((1-\phi)\,c + \phi\,\frac{\bar{\theta}+\tau\,\theta}{1+\tau}\right)\\ &= \phi^2\left(\frac{\tau}{1+\tau}\right)^2 \text{var}(\theta) = \phi^2\left(\frac{\tau}{1+\tau}\right)^2 \sigma^2\,,\end{aligned}$$

and the within-variance (corresponding to Equation (5)) becomes

$$\begin{aligned}\sigma^2_{\text{within}} &= \text{var}(r \mid \theta) = \text{var}\left((1-\phi)\,c + \phi\,\frac{\bar{\theta}+\tau\,x}{1+\tau}\,\middle|\,\theta\right)\\ &= \phi^2\left(\frac{\tau}{1+\tau}\right)^2 \text{var}(x \mid \theta) = \phi^2\,\frac{\tau}{(1+\tau)^2}\,\sigma^2\,.\end{aligned}$$

We see that both variances increase quadratically in the scale use parameter ϕ. However, for the ratio of the two, the effect of scale use cancels out, and it still holds that the ratio equals τ.

$$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{\phi^2\left(\frac{\tau}{1+\tau}\right)^2\sigma^2}{\phi^2\,\frac{\tau}{(1+\tau)^2}\,\sigma^2} = \tau$$

B.2 Social Desirability Effects

In some situations, respondents might not want to truthfully report their type but rather provide an answer that is deemed socially desirable. These contexts are likely to arise if the interview situation is not anonymous (audience effects) and/or if items are image-relevant to the respondent. For example, it is plausible that a respondent feels more comfortable reporting that she is an honest rather than a dishonest person. Such concerns can be integrated into our framework by adding a desirable answer $d \in \mathbb{R}$. Accordingly, the respondents' objective is to minimize the weighted sum of the squared distances to their type and the desirable answer. The utility function is thus

$$u_{\theta,d}(r) = -\,(1-\psi)\,(r-\theta)^2 - \psi\,(r-d)^2\ ,$$

where $\psi \in [0,1]$ measures the intensity of the preference to report d. The optimal report of a respondent equals the weighted sum of the best guess of her type θ and the desirable answer

$$r = (1-\psi)\left(\frac{\bar{\theta}+\tau x}{1+\tau}\right) + \psi\,d\,.$$

The respondent thus acts as if subject to subjective scale use, as introduced in Section B.1. The main difference between subjective scale use and desirability arises in the context of multiple agents and characteristics: while the scale use parameters (ϕ, c) are naturally agent-specific, the desirability parameters (ψ, d) are naturally specific to the characteristic.

B.3 Trembling Hand Errors

Lastly, instead of being biased by strategic motives, responses may be influenced by random noise that is independent of respondents' level of signal precision τ. For example, respondents may unintentionally misposition a slider, misread the response of a Likert scale item, or accidentally select the wrong category of a drop-down menu. Formally, we allow for noise in the observed response $\tilde{r} = r + \epsilon_r$ by setting it equal to the intended response r plus an independent Normal shock ϵ_r with variance σ_r^2. A straightforward computation shows that the within and between variances are then given by

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

Thus, the ratio of variance equals

$$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{\left(\frac{\tau}{1+\tau}\right)^2 \sigma^2}{\frac{\tau}{(1+\tau)^2}\,\sigma^2 + \sigma_r^2} = \tau \, \frac{1}{1 + \sigma_r^2 \frac{(1+\tau)^2}{\tau}} . \tag{15}$$

As a consequence, under this type of noise, the ratio of variance will underestimate the true level of τ.

C Robustness of the Estimator

C.1 Characteristics with Different Averages and Variances

The estimator introduced in Section 3.1 assumes that the population means and variances of types are identical for all K characteristics being used. Empirically, however, this is usually not the case (at least not exactly). For this reason, we next describe a generalization of the estimator derived in Section 3.1 to the case where the population mean $\bar{\theta}_k$ and variance σ_k^2 of each characteristic k are potentially different. We make no assumption about the distribution of these population means and variances, but maintain the assumption that the agent's prior belief equals the distribution of characteristics in the population and that characteristics are independent.

Fix an infinite sequence of levels of perceived and objective self-knowledge of the respondents, $\tau_1, \tau_2, \ldots$ and $\tilde{\tau}_1, \tilde{\tau}_2, \ldots$, respectively. We denote by

$$C := \frac{1}{I} \sum_{i=1}^{I} \left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2 \left(1 + \frac{1}{T\tau_i}\right)$$

and note that C is a non-negative constant independent of any specific characteristic. Throughout, we assume that each agent's self-knowledge τ_i is bounded from below by $\underline{\tau}$, which implies that C is bounded by $C \leq 1 + \frac{1}{T\underline{\tau}}$. There exist i.i.d. standard normally distributed random variables $(\epsilon_{ikt})_{ikt}$

and $(\eta_{ik})_{ik}$ such that

$$
\begin{aligned}
x_{ikt} &= \theta_{ik} + \frac{\sigma_k}{\sqrt{\tau_i}}\epsilon_{ikt}\,, \\
\theta_{ik} &= \bar{\theta}_k + \sigma_k \eta_{ik}\,.
\end{aligned}
$$

We get that the agent's response when asked for the t^{th} time about characteristic k is then given by

$$
r_{ikt} = \frac{\bar{\theta}_k + \tilde{\tau}_i\, x_{ikt}}{1+\tilde{\tau}_i} = \bar{\theta}_k + \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}(x_{ikt} - \bar{\theta}_k) = \bar{\theta}_k + \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\sigma_k\left(\eta_{ik} + \frac{1}{\sqrt{\tau}_i}\epsilon_{ikt}\right)\,.
$$

We define the average response by agent i to question about characteristic k as $\bar{r}_{ik} = \frac{1}{T}\sum_{t=1}^{T} r_{ikt}$ and as $\bar{r}_k = \frac{1}{I}\sum_{i=1}^{I}\bar{r}_{ik}$ the average response to question k.

Lemma 1. *The average response to question k is normally distributed with mean $\bar{\theta}_k$ and variance*

$$
\mathit{var}(\bar{r}_k) = \frac{\sigma_k^2}{I}C\,.
$$

Furthermore, $\lim_{I\to\infty}\bar{r}_k = \bar{\theta}_k$ almost surely.

Proof. As η and ϵ are normally distributed with mean zero, it follows that $\bar{r}_k$ is normally distributed and has mean $\bar{\theta}_k$. We are thus left to compute the variance of $\bar{r}_k$. We define $\bar{\epsilon}_{ik} = \frac{1}{T}\sum_{t=1}^{T}\epsilon_{ikt}$ as the average signal shock of agent i for characteristic k. As η_{ik} and $\bar{\epsilon}_{ik}$ are independent across agents, we have that

$$
\begin{aligned}
\text{var}(\bar{r}_k) &= \frac{1}{I^2}\sum_{i=1}^{I}\text{var}(\bar{r}_{ik}) = \frac{1}{I^2}\sum_{i=1}^{I}\text{var}\left(\bar{\theta}_k + \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\sigma_k\left(\eta_{ik} + \frac{1}{\sqrt{\tau}_i}\bar{\epsilon}_{ik}\right)\right) \\
&= \frac{\sigma_k^2}{I^2}\sum_{i=1}^{I}\left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2\text{var}\left(\eta_{ik} + \frac{1}{\sqrt{\tau}_i}\bar{\epsilon}_{ik}\right) \\
&= \frac{\sigma_k^2}{I^2}\sum_{i=1}^{I}\left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2\left(1 + \frac{\text{var}(\bar{\epsilon}_{ik})}{\tau_i}\right) = \frac{\sigma_k^2}{I^2}\sum_{i=1}^{I}\left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2\left(1 + \frac{\frac{1}{T^2}\sum_{t=1}^{T}\text{var}(\epsilon_{ikt})}{\tau_i}\right) \\
&= \frac{\sigma_k^2}{I^2}\sum_{i=1}^{I}\left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2\left(1 + \frac{1}{T\tau_i}\right)\,.
\end{aligned}
$$

The almost sure convergence follows from Kolmogorov's strong law of large numbers for independently but not identically distributed random variables. □

Similarly, we define the variance in responses to question k as

$$
s_k^2 = \frac{1}{I-1}\sum_{i=1}^{I}(\bar{r}_{ik} - \bar{r}_k)^2\,.
$$

Lemma 2. *We have that the expected sample variance converges almost surely*

$$
\lim_{I\to\infty} s_k^2 = \sigma_k^2 C\,.
$$

Proof. As $\lim_{I\to\infty} \bar{r}_k = \bar{\theta}_k$ a.s., the sample variance a.s. satisfies

$$\begin{aligned}
\lim_{I\to\infty} s_k^2 &= \lim_{I\to\infty} \frac{1}{I-1} \sum_{i=1}^{I} \left[(\bar{r}_{ik} - \bar{\theta}_k)^2 + (\bar{\theta}_k - \bar{r}_k)^2 + 2(\bar{r}_{ik} - \bar{\theta}_k)(\bar{\theta}_k - \bar{r}_k)\right] \\
&= \lim_{I\to\infty} \frac{1}{I-1} \sum_{i=1}^{I} \left[(\bar{r}_{ik} - \bar{\theta}_k)^2 + (\bar{\theta}_k - \bar{r}_k)^2\right] \\
&= \lim_{I\to\infty} \frac{I}{I-1} \left[(\bar{\theta}_k - \bar{r}_k)^2 + \frac{1}{I} \sum_{i=1}^{I} (\bar{r}_{ik} - \bar{\theta}_k)^2\right] .
\end{aligned}$$

As $I/(I-1)$ converges to 1 and $(\bar{\theta}_k - \bar{r}_k)^2$ converges to zero almost surely, we get that almost surely

$$\lim_{I\to\infty} s_k^2 = \lim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} (\bar{r}_{ik} - \bar{\theta}_k)^2 .$$

Note that $\bar{r}_{ik} - \bar{\theta}_k$ is independently normally distributed with mean zero and variance

$$\sigma_k^2 \left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2 \left(1 + \frac{1}{T\tau_i}\right) .$$

Thus, we get that

$$\mathbb{E}[(\bar{r}_{ik} - \bar{\theta}_k)^2] = \sigma_k^2 \left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2 \left(1 + \frac{1}{T\tau_i}\right)$$

and

$$\operatorname{var}\big((\bar{r}_{ik} - \bar{\theta}_k)^2\big) = 2\sigma_k^4 \left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^4 \left(1 + \frac{1}{T\tau_i}\right)^2 \leq 2\sigma_k^4 \left(1 + \frac{1}{T\underline{\tau}}\right)^2 .$$

As the variance of $\left(\bar{r}_{ik} - \bar{\theta}_k\right)^2$ is bounded, we can apply Kolmogorov's strong law of large numbers and get that

$$\lim_{I\to\infty} s_k^2 = \lim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} (\bar{r}_{ik} - \bar{\theta}_k)^2 = \lim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} \sigma_k^2 \left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2 \left(1 + \frac{1}{T\tau_i}\right) = \sigma_k^2 C . \qquad \square$$

We define the normalized response n_{ikt} as the difference between agent i's response and the average response, divided by the standard deviation of agents' average responses for the given characteristic k, i.e.

$$n_{ikt} = \frac{r_{ikt} - \bar{r}_k}{s_k} .$$

Together Lemma 1 and 2 imply the following result.

Lemma 3. *The normalized response times $\sqrt{C}$ almost surely converge in the number of agents to*

$$\lim_{I\to\infty} \sqrt{C}\, n_{ikt} = \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i} \left(\eta_{ik} + \frac{1}{\sqrt{\tau}_i} \epsilon_{ikt}\right) \tag{16}$$

We observe that the above asymptotic distribution for $I \to \infty$ of the normalized responses multiplied by $\sqrt{C}$ does not depend on the means and variances of characteristics. Moreover, the comparison of Equations (16) and (12) shows that the normalized responses are distributed exactly as if all

Table C.1: Accuracy of estimates with different means and variances

	(1)	(2)	(3)	(4)	(5)
I (respondents)	100	10,000	100	100	100
K (characteristics)	15	15	50	15	50
T (waves)	3	3	3	10	10
Correlation τ and $\hat{\tau}$	0.68	0.68	0.87	0.76	0.91
Rank correlation τ and $\hat{\tau}$	0.76	0.77	0.90	0.82	0.93
Median split correct	79%	80%	88%	83%	90%

Notes: The table replicates Table 1 with the only difference being that the mean characteristics are drawn from a normal distribution.

means $\bar{\theta}_k$ were zero, and the variances σ_k^2 of characteristics all took the value of $1/C$. We define the population-based estimator as

$$\hat{\tau}_i^{POP} = \frac{\frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{n}_{ik} - \bar{n}_i\right)^2}{\frac{1}{K(T-1)-2}\sum_{k=1}^{K}\sum_{t=1}^{T}\left(n_{ikt} - \bar{n}_{ik}\right)^2} - \frac{1}{T}\,. \tag{17}$$

The proof given for the theorem now yields the following result:

Proposition 4. *For every K, T that satisfy $K(T-1) > 4$.*

1. *The estimator $\hat{\tau}_i^{POP}$ satisfies almost surely*

$$\lim_{I\to\infty} \hat{\tau}_i^{POP} = \left(\tau_i + \frac{1}{T}\right)\frac{K(T-1)-2}{K(T-1)}F_i - \frac{1}{T} \tag{18}$$

for some random variable F_i that is F distributed with $K-1, K(T-1)$ degrees of freedom for every fixed vector of parameters $\tau_i, \sigma, \bar{\theta}$.

2. *$\hat{\tau}_i^{POP}$ is a consistent estimator for τ_i^{POP}, i.e., $\lim_{I\to\infty}\mathbb{E}\left[\hat{\tau}_i^{POP} \mid \tau_i\right] = \tau_i$ almost surely.*

3. *The standard error of the estimator $\hat{\tau}_i^{POP}$ in large populations is given by*

$$\lim_{I\to\infty}\sqrt{\mathbb{E}\left[(\hat{\tau}_i^{POP} - \tau_i)^2 \mid \tau_i\right]} = \left(\tau_i + \frac{1}{T}\right)\sqrt{\frac{2((K-1)+K(T-1)-2)}{(K-1)(K(T-1)-4)}}\,. \tag{19}$$

4. *$\hat{\tau}_i^{POP}$ converges to τ_i at the rate $1/\sqrt{K}$ in the number of attributes, and for all $K > 4$ satisfies the following upper bound independent of the number of repeated observations T*

$$\lim_{I\to\infty}\sqrt{\mathbb{E}\left[(\hat{\tau}_i^{POP} - \tau_i)^2 \mid \tau_i\right]} \leq \frac{2\tau_i + 1}{\sqrt{K-4}}\,.$$

The properties of the population-based estimator are now asymptotic and do not necessarily hold in small samples. However, the only dimension of the sample size relevant for convergence is the number I of respondents. While, in most applications, the number of characteristics and waves (K and T, respectively) will probably be limited, the number of respondents is usually fairly large. Therefore, the asymptotic properties might be a realistic approximation of the actual behavior of the population-based estimator in many relevant contexts, as we illustrate with simulation results next.

In Table C.1, we replicate Table 1, with the only difference being the assumption put on the means of the characteristics. The means $\bar{\theta}$ are independently drawn from a normal distribution with a mean of 5 and a standard deviation of 1. The standard deviations of characteristics, θ, are drawn from a log-normal distribution with the parameters $-1/2$ and 1, such that the expected standard deviation still equals one. A comparison of the results shows that the performance is almost identical to the case with equal means. This even holds for the cases where the simulated number of respondents is just 100 – a sample size that most studies exceed.

C.2 Correlated Characteristics

Table C.2: Accuracy of estimates with correlated characteristics

	No corr.		Low corr.		Moderate corr.		High corr.		Perfect corr.		Random corr.	
	$Cov_{\theta_k,\theta_{j\neq k}} = 0$		$Cov_{\theta_k,\theta_{j\neq k}} = 0.20$		$Cov_{\theta_k,\theta_{j\neq k}} = 0.50$		$Cov_{\theta_k,\theta_{j\neq k}} = 0.80$		$Cov_{\theta_k,\theta_{j\neq k}} = 1$		$Cov_{\theta_k,\theta_{j\neq k}} \sim U[0,1]$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
I (respondents)	100	100	100	100	100	100	100	100	100	100	100	100
K (characteristics)	15	50	15	50	15	50	15	50	15	50	15	50
T (waves)	3	3	3	3	3	3	3	3	3	3	3	3
Cronbach's alpha among θ_k	0.01	0.01	0.70	0.89	0.88	0.96	0.95	0.98	0.97	0.99	0.87	0.95
Correlation τ and $\hat{\tau}$	0.68	0.87	0.68	0.87	0.65	0.85	0.55	0.79	0.01	0.02	0.60	0.80
Rank correlation τ and $\hat{\tau}$	0.76	0.90	0.76	0.90	0.72	0.88	0.61	0.82	0.01	0.02	0.68	0.85
Median split correct	80%	88%	80%	88%	77%	87%	73%	83%	51%	51%	76%	85%

Notes: The table shows the results of simulating response behavior according to the model for different degrees of correlations between characteristics θ_k. For each resulting dataset, we estimate $\hat{\tau}$ and compare its correlation with the true τ. In the last row, we furthermore look at how many respondents are correctly classified as above or below median τ when performing a median split according to $\hat{\tau}$.

Our self-knowledge estimator, developed in Section 3.1, assumes that characteristics are independent, or equivalently, that signals about underlying characteristics are not correlated across questions. This may often be the case on the trait level, i.e., the Big Five are constructed to be independent of each other. However, traits are often measured using related items. For instance, the Big Five inventory we use measures each of the five traits using three individual survey items, which are correlated with each other. This does not impede the logic behind our estimator: subjects with high self-knowledge should give similar answers over time to the same questions, and they should give different answers to questions about different traits. However, the informational value of each item is no longer the same across items, which may decrease the efficiency of our estimator. To examine how inter-item correlations affect estimator performance, we build on the simulation setup of Section 3.2. That is, we again draw agents' levels of self-knowledge τ_i from a uniform distribution with support $[0.1, 5]$, set the true average value of characteristics $\bar{\theta}$ to 5 and the true population variance σ^2 to 1. However, instead of imposing independence between characteristics, we introduce correlations and report the results in Table C.2. As a benchmark, in Columns (1) and (2), we reproduce the results of Table 1, where independence was imposed. Starting with Column (3), we gradually increase the degree of correlation among characteristics. Specifically, we draw characteristics from a multivariate normal distribution with off-diagonal covariances set to 0.20 in Columns (3)-(4), 0.50 in Columns (5)-(6), 0.80 in Columns (7)-(8), and 1.00 (perfect collinearity) in Columns (9)-(10). Lastly, in Columns (11) and (12), we randomly draw each off-diagonal covariance entry from a uniform distribution[30]

As a summary and benchmark measure of item intercorrelation, we also report Cronbach's alpha,

[30] We discard entries which would lead the resulting matrix no longer to be positive semi-definite.

estimated among characteristics of the first wave and averaged across simulations. Cronbach's alpha is often used to assess the reliability of scales, with rules of thumb generally categorizing scales with an alpha below 0.6 as poor, between 0.6 and 0.7 as acceptable, between 0.7 and 0.9 as good, and above 0.9 as excellent.

We find that our estimate performs well even in cases of high correlation among characteristics. For instance, the rank correlation of τ with $\hat{\tau}$ drops by just 0.15 points from 0.76 to 0.61 when moving from the no correlation to the high correlation case using 15 characteristics. In this case, Cronbach's alpha is 0.95, indicating a highly internally consistent scale. Only in the case of perfect collinearity (Columns (9) and (10)) does the identification break down, as expected. When comparing odd with even columns, we also observe how increasing the number of characteristics from 15 to 50 reduces the impact of correlated characteristics. For instance, in the high correlation case, the rank correlation of τ with $\hat{\tau}$ drops by merely 0.08 points.

While in the previous Table, we either imposed a constant correlation or randomly drew covariances, in Table C.3, we instead incorporate empirically observed correlations. Specifically, we impose that characteristics are correlated in the same way as the answers to the 15 Big Five questions in the 2009 wave of the SOEP (see Section 5.4.1). We then replicate all columns of Table 1 that use 15 characteristics using this new simulation.

Table C.3: Accuracy of estimates assuming SOEP correlated characteristics

	(1)	(2)	(3)
I (respondents)	100	10,000	100
K (characteristics)	15	15	15
T (waves)	3	3	10
Correlation τ and $\hat{\tau}$	0.65	0.64	0.72
Rank correlation τ and $\hat{\tau}$	0.74	0.74	0.80
Median split correct	78%	78%	81%

Notes: The table replicates Table 1 with the only difference being that characteristics are no longer independently drawn but correlated according to the realized correlation among Big-Five items in the SOEP.

The results are reported in Table C.3, whose columns are identically constructed as Columns 1, 2, and 4 in Table 1. The main result is that the fraction of respondents who are correctly classified as having below- or above-median self-knowledge decreases only by about two percentage points, i.e., the informativeness of the median-splits remains.

In summary, our simulation results suggest that our estimator extracts sufficient information from responses, even in cases of highly correlated characteristics.

D Stability of the Estimator across Domains and Time

In our model, we treat τ as an individual-specific parameter, which is assumed to be stable over time and across domains. We use our survey (Section 5) to test this assumption.

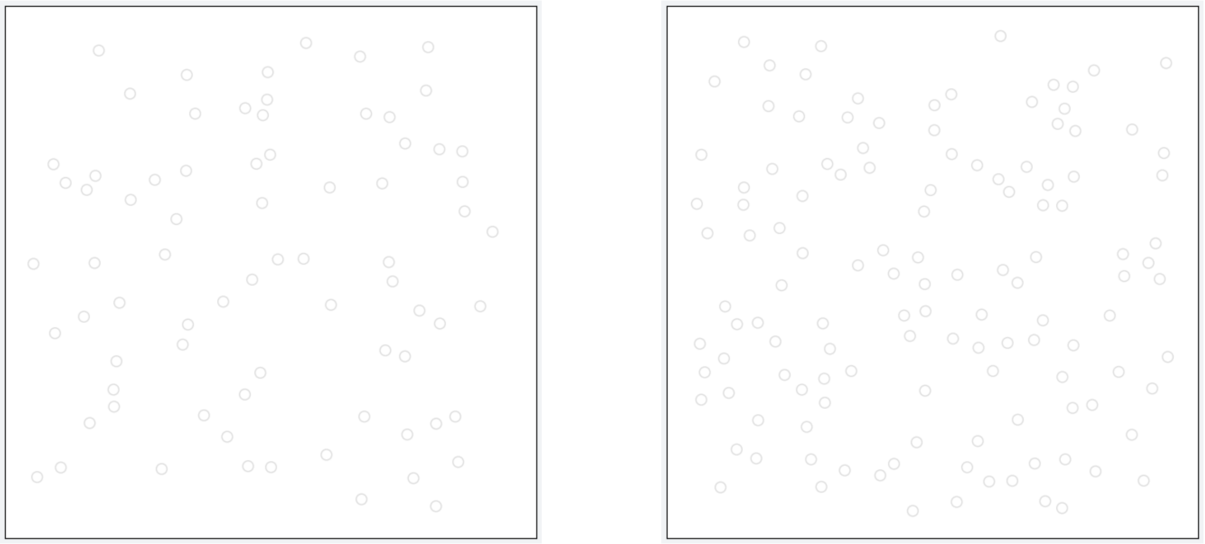
Stability over time. To test for stability over time, we estimate τ separately for the first wave and the second wave. That is, we use the two within-session repetitions to generate a session-one $\hat{\tau}$ and a session-two $\hat{\tau}$. Correlating the two separate measures, we find a strong correlation of $\rho = 0.62$ ($p < 0.001$) at the individual level, indicating a high degree of stability over time for the one-week time lag of the survey. In comparison, this correlation is similar to the correlations of conscientiousness ($\rho = 0.67$, $p < 0.001$) and neuroticism ($\rho = 0.65$, $p < 0.001$) between the two waves. These measures are commonly considered stable character traits (see, e.g., Caspi, Roberts, and Shiner, 2005; Cobb-Clark and Schurer, 2012).

Stability across domains. To investigate stability across domains, we compare our estimate of τ if estimated using the Big-Five responses to estimates of τ obtained using the domain-specific self-assessments of risk and altruism. As these self-assessments each contain 5 items that are repeated twice per wave, we have a similar panel set as with our Big-Five module. We find significant correlations of $\rho = 0.44$ ($p < 0.001$) when we compare τ estimated from the Big-Five with τ estimated from the altruism items, and $\rho = 0.40$ ($p < 0.001$) when we compare the Big-Five estimate with τ estimated from the risk items.

Taken together, in the context of our standard survey application, $\hat{\tau}$ appears to be stable, both over time and across domains.

E Additional Figures

Figure E.1: Example images

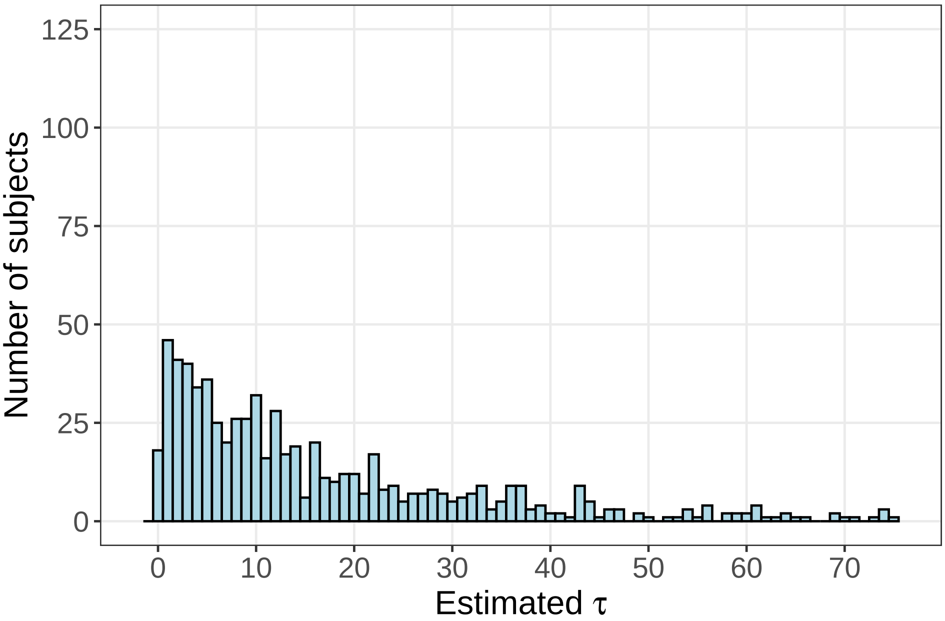


(a) Size within Category 1

(b) Size within Category 9

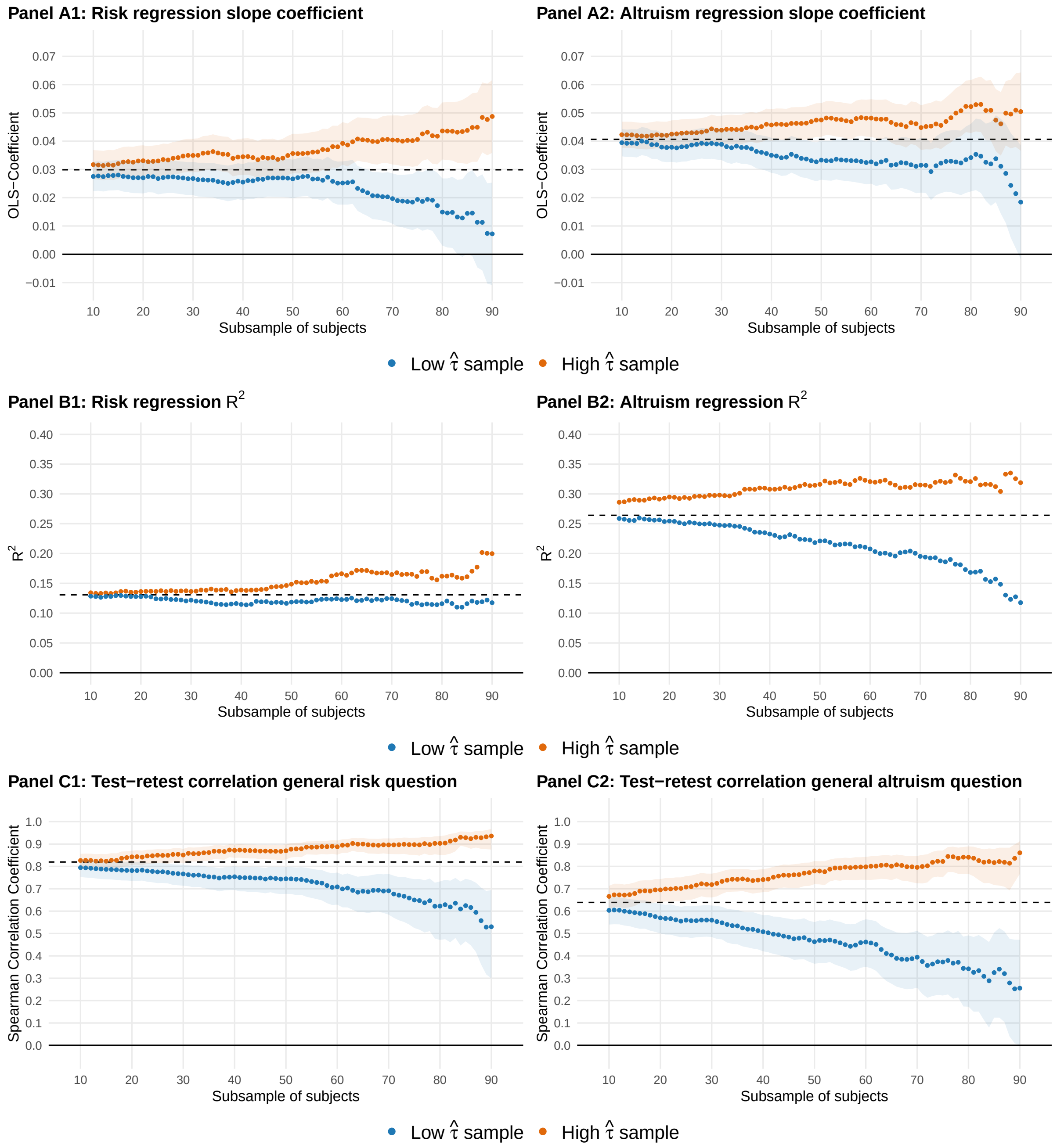
Note: The panel on the left shows an image belonging to Category 1, containing 75 dots. The panel on the right shows an image of Category 9, containing 125 dots.

Figure E.2: Distribution of $\hat{\tau}$ in the survey



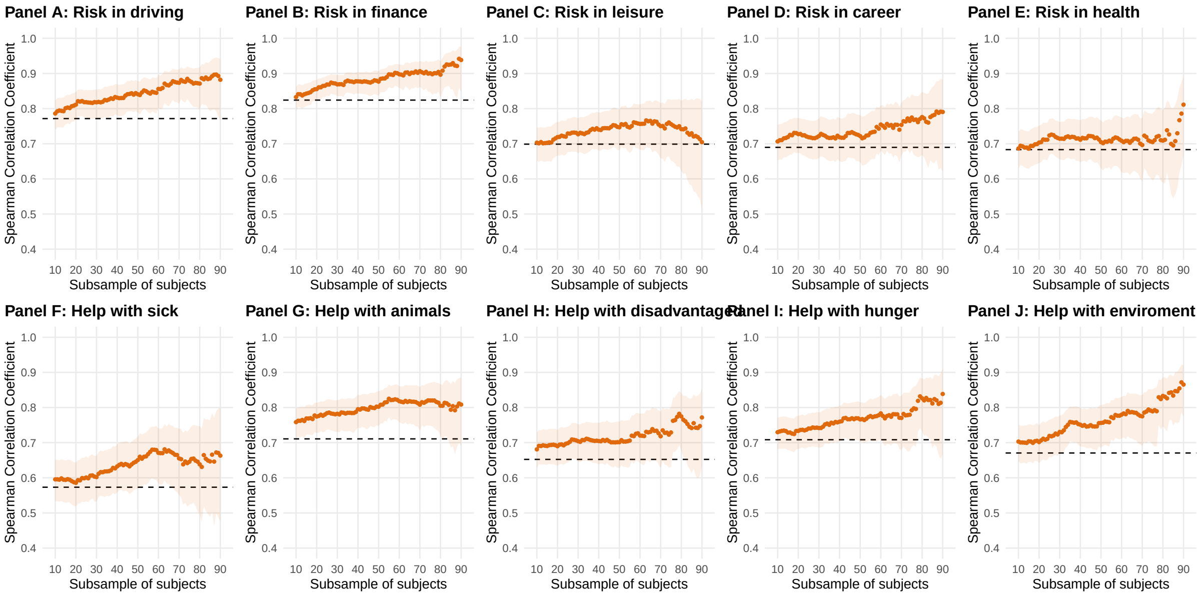
Notes: The figure shows as histogram the distribution of the estimated τ ($\hat{\tau}$) for the online survey (Section 5. Displayed are the estimates for 694 subjects out of the survey sample of 740 subjects. 6 Subjects are excluded because they have non-finite $\hat{\tau}$ and 40 additional subjects are excluded because they have values of $\hat{\tau}$ larger than 75. The binwidth is 1.

Figure E.3: The influence of subsetting based on $\hat{\tau}$ on coefficients, predictive power and test-retest correlations



Notes: In **Panel A1** and **A2**, each dot represents the OLS-coefficient from a stacked regression. In Panel A1 (Panel A2), the general risk (altruism) question is the regression's independent variable, and risk (altruistic) behavior is the dependent variable. Each regression is run on a subsample based on removing x% of subjects based on the estimated level of self-knowledge $\hat{\tau}$. Orange dots represent removing x% of subjects with low $\hat{\tau}$ (subsetting on higher levels of self-knowledge), while blue dots represent removing x% of subjects with high $\hat{\tau}$ (subsetting on lower levels of self-knowledge). Shaded areas indicate 95% confidence intervals. **Panel B1** and **B2**: These panels display the R^2 values instead of the OLS coefficients from the respective regressions. **Panel C1** and **C2**: Each dot in Panel C1 represents Spearman's rank correlation coefficient between the general risk question in wave 1 and wave 2 of the survey. In Panel C2, each dot represents the wave 1 to wave 2 correlation of the general altruism question. Shaded areas indicate 95% confidence intervals.

Figure E.4: Test-retest correlations of domain-specific self-assessments



Notes: Each dot displays Spearman's rank correlation coefficient between the respective domain-specific self-assessment in wave 1 and wave 2 of the survey. For example, Panel A displays the correlation of self-assessed risk-taking while driving between wave 1 and wave 2. Each correlation is obtained from a subsample where a percentage of subjects with the lowest estimated level of self-knowledge $\hat{\tau}$ are excluded. Dots represent removing x% of subjects with low $\hat{\tau}$ (subsetting on higher levels of self-knowledge). Shaded areas indicate 95% confidence.

Table F.1: Choice categories

0	1	2	3	4	5	6	7	8	9	10
50–68	69–75	76–82	83–89	90–96	97–103	104–110	111–117	118–124	125–131	132–150
0.6%	1.9%	5.6%	12.0%	18.9%	22.0%	18.9%	12.0%	5.6%	1.9%	0.6%

Note: top row: categories; middle row: number of dots in image; bottom row: respective likelihoods.

F Additional Tables

Table F.2: Module to measure self-knowledge

Original version I am someone who...	**Rephrased version** I see myself as a person who ...	**Coding**	**Dimension**
is sometimes a bit rough towards others.	is partly a bit rough.	-	Agreeableness
can forgive.	forgives others.	+	Agreeableness
is considerate and friendly towards others.	is generally friendly and considerate to others.	+	Agreeableness
works thoroughly.	places value on thoroughness in work.	+	Conscientiousness
is rather lazy.	has a tendency towards laziness.	-	Conscientiousness
completes tasks effectively and efficiently.	approaches tasks efficiently and effectively	+	Conscientiousness
is communicative, talkative.	enjoys communicating and talking with others.	+	Extraversion
is reserved.	tends to hold back.	-	Extraversion
can open up, is sociable.	enjoys sociability and opens up.	+	Extraversion
often worries.	is often worried.	+	Neuroticism
gets nervous easily.	tends to be nervous.	+	Neuroticism
is relaxed, can handle stress well.	remains relaxed even under stress.	-	Neuroticism
is original, brings in new ideas.	stands out with new ideas and originality.	+	Openness
values artistic experiences.	is interested in art.	+	Openness
has a vivid imagination and ideas.	is imaginative and has creativity.	+	Openness

Notes: Displayed is the module used to estimate self-knowledge in our survey (see Section 5). It is based on the 15-item Big Five questionnaire of (Schupp and Gerlitz, 2008), displayed in column "Original version"). Each item is repeated in a rephrased version (column "Rephrased version").

Table F.3: The correlation between different measures of the quality of survey response behavior

	Reported attention	Attention check 1 passed	Attention check 2 passed	Both attention checks passed	Reported effort	Effort check passed	Reported reliability	Not speeding (not top 10%)	Not slowest or fastest 25%
Estimated self-knowledge $\hat{\tau}$	0.25***	0.20***	0.22***	0.22***	0.19***	0.16***	0.27***	0.19***	0.09
Reported attention		0.25***	0.20***	0.18***	0.45***	0.16***	0.51***	0.24***	0.19***
Attention check 1 passed			0.39***	0.51***	0.19***	0.25***	0.14***	0.21***	0.12*
Attention check 2 passed				0.94***	0.14***	0.26***	0.17***	0.18***	0.08
Both attention checks passed					0.13**	0.27***	0.16***	0.17***	0.08
Reported effort						0.09	0.46***	0.15***	0.13**
Effort check passed							0.09	0.20***	0.09
Reported reliability								0.17***	0.12*
Not speeding (not top 10%)									0.33***
Not slowest or fastest 25%									

Notes: The table shows correlations between different measures of the quality of survey response behavior. For details on the measures, see Section 5. Significance levels (Bonferroni adjusted): *p$<$0.1, **p$<$0.05 and ***p$<$0.01.

Table F.4: The relationship between self-assessments and behavior

Panel A: Risk behavior						
Dependent variable:	Lottery certainty equivalent	Smoking	Investment in stocks	Self-employed	Does sports	Risk stacked regression
	(1)	(2)	(3)	(4)	(5)	(6)
General risk question	0.009**	0.020***	0.045***	0.029***	0.047***	0.030***
	(0.004)	(0.006)	(0.006)	(0.005)	(0.006)	(0.002)
Constant	0.229***	0.306***	0.104***	0.007	0.344***	0.163***
	(0.017)	(0.030)	(0.023)	(0.018)	(0.030)	(0.018)
Observations	734	734	734	734	734	3,670
R^2	0.008	0.013	0.079	0.062	0.068	0.131
Unconditional average	0.263	0.383	0.278	0.117	0.523	
Panel B: Altruistic behavior						
Dependent variable:	Dictator game giving	Donated money	Volunteered time	Helped stranger	Sent gift	Altruistic stacked regression
	(1)	(2)	(3)	(4)	(5)	(6)
General altruism question	0.066***	0.045***	0.018***	0.038***	0.036***	0.041***
	(0.004)	(0.005)	(0.004)	(0.006)	(0.006)	(0.002)
Constant	0.142***	-0.037	0.026	0.244***	0.509***	-0.009
	(0.028)	(0.028)	(0.026)	(0.041)	(0.043)	(0.019)
Observations	734	734	734	734	734	3,670
R^2	0.233	0.084	0.021	0.044	0.053	0.264
Unconditional average	0.559	0.247	0.138	0.482	0.738	

Notes: The table shows OLS estimates. **Panel A**: The independent variable is the general risk question. Measured on an 11-point Likert scale, higher values indicate more self-assessed risk-seeking. The dependent variable varies between columns. In Column (1), it is the elicited certainty equivalent of a lottery. Higher values indicate a higher certainty equivalent measured in money. In columns (2) to (5), the dependent variable reflects different activities. Each is an indicator variable equal to one if the subject engages in the respective activity and zero otherwise. Column (6) presents the coefficients obtained from stacking the regressions of columns (1) to (5). **Panel B**: The independent variable in each case is the general altruism question. Measured on an 11-point Likert scale, higher values indicate more self-assessed altruism. The dependent variable varies between columns. In Column (1), it is the monetary amount given to charity in a dictator game. In columns (2) to (5), the dependent variable reflects different activities. Each is an indicator variable equal to one if the subject engages in the respective activity and zero otherwise. Column (6) presents the coefficients obtained from stacking the regressions of columns (1) to (5). Standard errors in parentheses are clustered at the subject level. Significance levels: $^{*}\,p < 0.1,^{**}\,p < 0.05,\,^{***}\,p < 0.01.$

Table F.5: Improving the relationship between self-assessments and incentivized behavior using $\hat{\tau}$ for all behaviors

Panel A: Risk behavior															
Dependent variable:	Lottery certainty equivalent			Smoking			Investment in stocks			Self-employed			Does sports		
	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
General risk question	0.009**	0.017***	0.003	0.020***	0.024**	0.017*	0.045***	0.052***	0.043***	0.029***	0.032***	0.027***	0.047***	0.054***	0.043***
	(0.004)	(0.006)	(0.005)	(0.006)	(0.009)	(0.009)	(0.006)	(0.008)	(0.008)	(0.005)	(0.007)	(0.007)	(0.006)	(0.008)	(0.009)
Constant	0.229***	0.214***	0.243***	0.306***	0.302***	0.307***	0.104***	0.116***	0.079**	0.007	0.005	0.007	0.344***	0.349***	0.327***
	(0.017)	(0.021)	(0.027)	(0.030)	(0.041)	(0.046)	(0.023)	(0.032)	(0.034)	(0.018)	(0.023)	(0.028)	(0.030)	(0.039)	(0.046)
Observations	734	367	367	734	367	367	734	367	367	734	367	367	734	367	367
R^2	0.008	0.026	0.001	0.013	0.018	0.010	0.079	0.095	0.074	0.062	0.074	0.053	0.068	0.088	0.058

Panel B: Altruistic behavior															
Dependent variable:	Dictator game giving			Donated money			Volunteered time			Helped stranger			Sent gift		
	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
General altruism question	0.066***	0.077***	0.053***	0.045***	0.052***	0.038***	0.018***	0.024***	0.011	0.038***	0.041***	0.036***	0.036***	0.043***	0.029***
	(0.004)	(0.004)	(0.007)	(0.005)	(0.006)	(0.008)	(0.004)	(0.006)	(0.006)	(0.006)	(0.008)	(0.009)	(0.006)	(0.008)	(0.009)
Constant	0.142***	0.079**	0.215***	−0.037	−0.113***	0.037	0.026	−0.038	0.090**	0.244***	0.190***	0.292***	0.509***	0.465***	0.558***
	(0.028)	(0.033)	(0.046)	(0.028)	(0.033)	(0.045)	(0.026)	(0.031)	(0.041)	(0.041)	(0.055)	(0.061)	(0.043)	(0.060)	(0.062)
Observations	734	367	367	734	367	367	734	367	367	734	367	367	734	367	367
R^2	0.233	0.357	0.132	0.084	0.130	0.052	0.021	0.047	0.006	0.044	0.056	0.037	0.053	0.080	0.031

Notes: The table shows OLS estimates. The dependent and independent variables are the same as in Table F.4. Added are sample splits based on the median level of self-knowledge, estimated using the estimator $\hat{\tau}$. Standard errors (in parentheses) are clustered at the subject level. Significance levels: *p<0.1, **p<0.05 and ***p<0.01.

Table F.6: Comparing methods to improve the relation between risk self-assessment and behavior

Dependent variable:	Lottery certainty equivalent	Smoking	Investment in stocks	Self-employed	Does sports	Stacked regression
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline	0.009	0.020	0.045	0.029	0.047	0.030
	(0.004)	(0.006)	(0.006)	(0.005)	(0.006)	(0.002)
Self-knowledge						
Above median self-knowledge	0.017	0.024	0.052	0.032	0.054	0.036
	(0.006)	(0.009)	(0.008)	(0.007)	(0.008)	(0.004)
Below median self-knowledge	0.003	0.017	0.043	0.027	0.043	0.027
	(0.005)	(0.009)	(0.008)	(0.007)	(0.009)	(0.003)
Top 20% self-knowledge	0.011	0.029	0.06	0.051	0.068	0.044
	(0.009)	(0.014)	(0.012)	(0.012)	(0.009)	(0.005)
Bottom 80% self-knowledge	0.009	0.018	0.042	0.024	0.041	0.027
	(0.004)	(0.007)	(0.007)	(0.005)	(0.007)	(0.003)
Attention						
Above median reported attention	0.018	0.016	0.043	0.018	0.053	0.029
	(0.005)	(0.008)	(0.007)	(0.006)	(0.007)	(0.003)
Below median reported attention	-0.009	0.033	0.045	0.042	0.043	0.031
	(0.007)	(0.012)	(0.011)	(0.01)	(0.011)	(0.005)
Attention check 1 passed	0.007	0.012	0.045	0.021	0.050	0.027
	(0.004)	(0.008)	(0.007)	(0.005)	(0.007)	(0.003)
Attention check 1 failed	0.015	0.038	0.048	0.044	0.035	0.036
	(0.009)	(0.013)	(0.012)	(0.011)	(0.013)	(0.005)
Attention check 2 passed	0.016	0.018	0.055	0.018	0.049	0.031
	(0.006)	(0.009)	(0.008)	(0.007)	(0.009)	(0.004)
Attention check 2 failed	0.005	0.018	0.038	0.037	0.046	0.029
	(0.005)	(0.009)	(0.008)	(0.007)	(0.008)	(0.004)
Both attention checks passed	0.015	0.013	0.056	0.020	0.052	0.031
	(0.006)	(0.01)	(0.009)	(0.007)	(0.009)	(0.004)
At least one attention check failed	0.006	0.021	0.038	0.034	0.044	0.029
	(0.005)	(0.009)	(0.008)	(0.007)	(0.008)	(0.003)
Effort						
Above median reported effort	0.016	0.027	0.044	0.026	0.053	0.033
	(0.006)	(0.009)	(0.008)	(0.007)	(0.008)	(0.003)
Below median reported effort	0.001	0.015	0.042	0.032	0.044	0.027
	(0.005)	(0.01)	(0.009)	(0.008)	(0.01)	(0.004)
Effort check passed	0.011	0.008	0.06	0.012	0.05	0.028
	(0.007)	(0.011)	(0.009)	(0.007)	(0.01)	(0.004)
Effort check failed	0.009	0.026	0.038	0.037	0.044	0.031
	(0.005)	(0.008)	(0.007)	(0.007)	(0.008)	(0.003)
Reliability						
Above median reported reliability	0.015	0.035	0.047	0.025	0.048	0.034
	(0.005)	(0.008)	(0.007)	(0.006)	(0.007)	(0.003)
Below median reported reliability	0.001	0.001	0.041	0.032	0.046	0.024
	(0.006)	(0.012)	(0.01)	(0.009)	(0.011)	(0.004)
Response times						
Excluding Top 10% speeders	0.015	0.015	0.043	0.020	0.045	0.028
	(0.004)	(0.007)	(0.006)	(0.005)	(0.007)	(0.003)
Top 10% speeders	-0.021	0.055	0.058	0.058	0.056	0.041
	(0.009)	(0.018)	(0.021)	(0.021)	(0.02)	(0.008)
Excluding slowest & fastest 25%	0.02	0.016	0.043	0.013	0.044	0.027
	(0.006)	(0.01)	(0.009)	(0.007)	(0.009)	(0.004)
Slowest & fastest 25%	-0.001	0.026	0.047	0.040	0.050	0.032
	(0.005)	(0.009)	(0.008)	(0.007)	(0.008)	(0.003)
Averaging						
Averaging across domains	0.009	0.023	0.06	0.038	0.065	0.039
	(0.005)	(0.008)	(0.006)	(0.006)	(0.007)	(0.003)
Averaging responses (2 repetitions)	0.008	0.020	0.047	0.030	0.049	0.031
	(0.004)	(0.007)	(0.006)	(0.005)	(0.006)	(0.003)
Averaging responses (4 repetitions)	0.009	0.020	0.048	0.031	0.056	0.033
	(0.004)	(0.007)	(0.006)	(0.005)	(0.006)	(0.003)
Further methods						
Adding demographic controls	0.001	0.017	0.032	0.020	0.034	0.021
	(0.005)	(0.007)	(0.006)	(0.005)	(0.007)	(0.003)
Adjustment w/ anchoring vignettes	0.015	0.012	0.037	0.009	0.032	0.021
	(0.007)	(0.011)	(0.011)	(0.009)	(0.011)	(0.004)
ORIV	0.008	0.020	0.050	0.032	0.052	0.032
	(0.002)	(0.004)	(0.003)	(0.003)	(0.003)	(0.001)

Notes: The table shows OLS estimates. The independent variable in each case is the general risk question. Measured on an 11-point Likert scale, higher values indicate more self-assessed risk-seeking. The dependent variable varies between columns. In Column (1), it is the elicited certainty equivalent of a lottery. Higher values indicate a higher certainty equivalent measured in money. In columns (2) to (5), the dependent variable reflects different activities. Each is an indicator variable equal to one if the subject engages in the respective activity and zero otherwise. Column (6) presents the coefficients obtained from stacking the regressions of columns (1) to (5). Each row presents estimates from separate regressions using different subsamples or methodological adjustments. See Table 3 for details. Standard errors in parentheses are clustered at the subject level.

Table F.7: Comparing methods to improve the relation between altruism self-assessment and behavior

Dependent variable:	Dictator game giving	Donated money	Volunteered time	Helped stranger	Sent gift	Stacked regression
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline	0.066	0.045	0.018	0.038	0.036	0.041
	(0.004)	(0.005)	(0.004)	(0.006)	(0.006)	(0.002)
Self-knowledge						
Above median self-knowledge	0.077	0.052	0.024	0.041	0.043	0.047
	(0.004)	(0.006)	(0.006)	(0.008)	(0.008)	(0.003)
Below median self-knowledge	0.053	0.038	0.011	0.036	0.029	0.033
	(0.007)	(0.008)	(0.006)	(0.009)	(0.009)	(0.004)
Top 20% self-knowledge	0.076	0.063	0.029	0.048	0.045	0.052
	(0.008)	(0.01)	(0.01)	(0.012)	(0.013)	(0.005)
Bottom 80% self-knowledge	0.064	0.040	0.014	0.036	0.034	0.038
	(0.005)	(0.006)	(0.005)	(0.007)	(0.007)	(0.003)
Attention						
Above median reported attention	0.069	0.040	0.010	0.042	0.032	0.039
	(0.004)	(0.005)	(0.004)	(0.007)	(0.007)	(0.003)
Below median reported attention	0.058	0.062	0.041	0.026	0.048	0.047
	(0.009)	(0.011)	(0.01)	(0.013)	(0.012)	(0.005)
Attention check 1 passed	0.071	0.042	0.013	0.033	0.028	0.038
	(0.004)	(0.005)	(0.004)	(0.007)	(0.007)	(0.003)
Attention check 1 failed	0.05	0.056	0.036	0.055	0.066	0.053
	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)	(0.005)
Attention check 2 passed	0.075	0.041	0.010	0.018	0.038	0.036
	(0.005)	(0.006)	(0.005)	(0.009)	(0.008)	(0.003)
Attention check 2 failed	0.057	0.053	0.031	0.06	0.033	0.047
	(0.006)	(0.008)	(0.007)	(0.008)	(0.009)	(0.003)
Both attention checks passed	0.077	0.042	0.009	0.017	0.037	0.036
	(0.005)	(0.007)	(0.005)	(0.01)	(0.009)	(0.003)
At least one attention check failed	0.057	0.051	0.029	0.057	0.035	0.046
	(0.006)	(0.007)	(0.007)	(0.008)	(0.008)	(0.003)
Effort						
Above median reported effort	0.070	0.045	0.014	0.038	0.032	0.040
	(0.005)	(0.006)	(0.005)	(0.008)	(0.008)	(0.003)
Below median reported effort	0.059	0.045	0.023	0.037	0.044	0.042
	(0.007)	(0.008)	(0.007)	(0.01)	(0.009)	(0.004)
Effort check passed	0.070	0.049	0.013	0.026	0.033	0.038
	(0.006)	(0.008)	(0.006)	(0.01)	(0.009)	(0.004)
Effort check failed	0.064	0.043	0.022	0.046	0.039	0.043
	(0.005)	(0.006)	(0.006)	(0.008)	(0.008)	(0.003)
Reliability						
Above median reported reliability	0.070	0.043	0.017	0.044	0.028	0.040
	(0.005)	(0.006)	(0.005)	(0.007)	(0.007)	(0.003)
Below median reported reliability	0.058	0.051	0.024	0.030	0.051	0.043
	(0.007)	(0.009)	(0.008)	(0.011)	(0.01)	(0.004)
Response times						
Excluding Top 10% speeders	0.068	0.044	0.016	0.035	0.034	0.039
	(0.004)	(0.005)	(0.004)	(0.006)	(0.006)	(0.002)
Top 10% speeders	0.051	0.050	0.040	0.070	0.060	0.054
	(0.015)	(0.022)	(0.021)	(0.02)	(0.022)	(0.009)
Excluding slowest & fastest 25%	0.075	0.046	0.014	0.037	0.029	0.040
	(0.005)	(0.006)	(0.005)	(0.008)	(0.008)	(0.003)
Slowest & fastest 25%	0.056	0.043	0.022	0.038	0.047	0.041
	(0.007)	(0.008)	(0.007)	(0.009)	(0.009)	(0.004)
Averaging						
Averaging across domains	0.062	0.049	0.024	0.063	0.053	0.050
	(0.006)	(0.006)	(0.005)	(0.007)	(0.007)	(0.003)
Averaging responses (2 repetitions)	0.069	0.054	0.024	0.044	0.041	0.046
	(0.004)	(0.005)	(0.004)	(0.006)	(0.006)	(0.002)
Averaging responses (4 repetitions)	0.068	0.062	0.029	0.053	0.046	0.052
	(0.004)	(0.005)	(0.004)	(0.007)	(0.007)	(0.002)
Further methods						
Adding demographic controls	0.067	0.043	0.015	0.037	0.033	0.039
	(0.004)	(0.005)	(0.004)	(0.006)	(0.006)	(0.002)
Adjustment w/ anchoring vignettes	0.062	0.024	-0.008	0.010	0.031	0.024
	(0.007)	(0.008)	(0.007)	(0.009)	(0.008)	(0.004)
ORIV	0.074	0.060	0.027	0.048	0.044	0.051
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.002)

Notes: The table shows OLS estimates. The independent variable in each case is the general altruism question. Measured on an 11-point Likert scale, higher values indicate more self-assessed altruism. The dependent variable varies between columns. In Column (1), it is the monetary amount given to charity in a dictator game. In columns (2) to (5), the dependent variable reflects different activities. Each is an indicator variable equal to one if the subject engages in the respective activity and zero otherwise. Column (6) presents the coefficients obtained from stacking the regressions of columns (1) to (5). Each row presents estimates from separate regressions using different subsamples or methodological adjustments. See Table 3 for details. Standard errors in parentheses are clustered at the subject level.

Table F.8: Comparing methods to improve the R^2 when predicting risk behavior with self-assessments

Dependent variable:	Lottery certainty equivalent	Smoking	Investment in stocks	Self-employed	Does sports	Stacked regression	Stacked regression w/ FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Baseline	0.008	0.013	0.079	0.062	0.068	0.036	0.147
Self-knowledge							
Above median self-knowledge	0.026	0.018	0.095	0.074	0.088	0.048	0.164
Below median self-knowledge	0.001	0.010	0.074	0.053	0.058	0.028	0.135
Top 20% self-knowledge	0.013	0.030	0.157	0.172	0.160	0.084	0.176
Bottom 80% self-knowledge	0.008	0.011	0.065	0.045	0.051	0.028	0.144
Attention							
Above median reported attention	0.028	0.007	0.071	0.031	0.081	0.033	0.161
Below median reported attention	0.007	0.034	0.068	0.087	0.055	0.034	0.118
Attention check 1 passed	0.005	0.004	0.074	0.038	0.074	0.028	0.143
Attention check 1 failed	0.019	0.051	0.093	0.104	0.044	0.052	0.155
Attention check 2 passed	0.021	0.010	0.104	0.027	0.067	0.036	0.147
Attention check 2 failed	0.002	0.011	0.061	0.096	0.071	0.034	0.151
Both attention checks passed	0.018	0.005	0.107	0.033	0.073	0.035	0.145
At least one attention check failed	0.004	0.015	0.063	0.082	0.064	0.034	0.153
Effort							
Above median reported effort	0.025	0.025	0.087	0.058	0.090	0.046	0.173
Below median reported effort	0.000	0.007	0.055	0.066	0.056	0.026	0.126
Effort check passed	0.010	0.002	0.114	0.014	0.067	0.028	0.131
Effort check failed	0.009	0.024	0.064	0.097	0.066	0.040	0.163
Reliability							
Above median reported reliability	0.020	0.041	0.096	0.062	0.077	0.049	0.169
Below median reported reliability	0.000	0.000	0.051	0.054	0.055	0.019	0.124
Response times							
Excluding Top 10% speeders	0.019	0.007	0.071	0.034	0.059	0.029	0.143
Top 10% speeders	0.050	0.087	0.103	0.107	0.099	0.053	0.185
Excluding slowest & fastest 25%	0.035	0.007	0.067	0.014	0.053	0.027	0.150
Slowest & fastest 25%	0.000	0.024	0.087	0.111	0.082	0.043	0.144
Averaging							
Averaging across domains	0.006	0.013	0.103	0.081	0.098	0.045	0.156
Averaging responses (2 repetitions)	0.006	0.012	0.083	0.065	0.072	0.036	0.147
Averaging responses (4 repetitions)	0.006	0.012	0.080	0.068	0.089	0.039	0.150
Further methods							
Adding demographic controls	0.057	0.055	0.223	0.118	0.117	0.058	0.169
Adjustment w/ anchoring vignettes	0.007	0.002	0.017	0.002	0.010	0.006	0.117
ORIV	0.005	0.011	0.074	0.058	0.064	0.032	0.143

Notes: The table shows R^2 values obtained from OLS regressions. The independent variable in each regression is the general risk question. Measured on an 11-point Likert scale, higher values indicate more self-assessed risk-seeking. The dependent variable varies between columns. In Column (1), it is the elicited certainty equivalent of a lottery. In columns (2) to (5), the dependent variable reflects different activities. Each is an indicator variable equal to one if the subject engages in the respective activity and zero otherwise. Column (6) presents the coefficients obtained from stacking the regressions of columns (1) to (5). In Column (7), fixed effects for the individual activities are added. Each row presents estimates from separate regressions using different subsamples or methodological adjustments. See Table 3 for details. Standard errors in parentheses are clustered at the subject level.

Table F.9: Comparing methods to improve the R^2 when predicting altruistic behavior with self-assessments

Dependent variable:	Dictator game giving	Donated money	Volunteered time	Helped stranger	Sent gift	Stacked regression	Stacked regression w/FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Baseline	0.233	0.084	0.021	0.044	0.053	0.057	0.264
Self-knowledge							
Above median self-knowledge	0.357	0.130	0.047	0.056	0.080	0.084	0.316
Below median self-knowledge	0.132	0.052	0.006	0.037	0.031	0.035	0.221
Top 20% self-knowledge	0.350	0.201	0.064	0.082	0.089	0.107	0.321
Bottom 80% self-knowledge	0.207	0.065	0.013	0.039	0.046	0.047	0.255
Attention							
Above median reported attention	0.270	0.084	0.011	0.059	0.046	0.056	0.315
Below median reported attention	0.151	0.107	0.059	0.018	0.073	0.062	0.180
Attention check 1 passed	0.263	0.081	0.015	0.034	0.032	0.048	0.294
Attention check 1 failed	0.138	0.111	0.055	0.097	0.166	0.098	0.192
Attention check 2 passed	0.288	0.073	0.009	0.010	0.060	0.043	0.312
Attention check 2 failed	0.173	0.110	0.047	0.115	0.043	0.077	0.225
Both attention checks passed	0.295	0.076	0.009	0.008	0.053	0.042	0.320
At least one attention check failed	0.176	0.104	0.043	0.105	0.050	0.075	0.224
Effort							
Above median reported effort	0.297	0.105	0.020	0.054	0.046	0.064	0.295
Below median reported effort	0.159	0.066	0.024	0.034	0.065	0.049	0.234
Effort check passed	0.262	0.099	0.016	0.021	0.044	0.049	0.303
Effort check failed	0.212	0.076	0.026	0.068	0.059	0.063	0.239
Reliability							
Above median reported reliability	0.268	0.089	0.024	0.065	0.037	0.061	0.298
Below median reported reliability	0.168	0.087	0.025	0.024	0.085	0.054	0.223
Response times							
Excluding Top 10% speeders	0.242	0.087	0.019	0.039	0.049	0.054	0.279
Top 10% speeders	0.145	0.068	0.045	0.127	0.114	0.083	0.172
Excluding slowest & fastest 25%	0.325	0.105	0.017	0.048	0.037	0.062	0.287
Slowest & fastest 25%	0.150	0.064	0.024	0.040	0.076	0.051	0.242
Averaging							
Averaging across domains	0.134	0.065	0.024	0.079	0.074	0.056	0.263
Averaging responses (2 repetitions)	0.240	0.116	0.035	0.057	0.062	0.070	0.277
Averaging responses (4 repetitions)	0.205	0.133	0.046	0.073	0.072	0.077	0.284
Further methods							
Adding demographic controls	0.269	0.126	0.080	0.092	0.107	0.073	0.280
Adjustment w/ anchoring vignettes	0.107	0.012	0.002	0.002	0.020	0.010	0.217
ORIV	0.182	0.085	0.026	0.042	0.046	0.052	0.259

Notes: The table shows R^2 values obtained from OLS regressions. The independent variable in each regression is the general altruism question. Measured on an 11-point Likert scale, higher values indicate more self-assessed altruism. The dependent variable varies between columns. In Column (1), it is the monetary amount given to charity in a dictator game. In columns (2) to (5), the dependent variable reflects different activities. Each is an indicator variable equal to one if the subject engages in the respective activity and zero otherwise. Column (6) presents the coefficients obtained from stacking the regressions of columns (1) to (5). In Column (7), fixed effects for the individual activities are added. Each row presents estimates from separate regressions using different subsamples or methodological adjustments. See Table 3 for details. Standard errors in parentheses are clustered at the subject level.

Table F.10: Comparing methods to improve the test-retest correlations of self-assessments

	Test-retest correlation	
	General risk question	General altruism question
	(1)	(2)
Baseline	0.82	0.64
Self-knowledge		
Above median self-knowledge	0.87	0.78
Below median self-knowledge	0.74	0.46
Top 20% self-knowledge	0.90	0.84
Bottom 80% self-knowledge	0.78	0.57
Attention		
Above median reported attention	0.83	0.66
Below median reported attention	0.73	0.57
Attention check 1 passed	0.81	0.63
Attention check 1 failed	0.79	0.66
Attention check 2 passed	0.84	0.61
Attention check 2 failed	0.79	0.65
Both attention checks passed	0.84	0.61
At least one attention check failed	0.79	0.65
Effort		
Above median reported effort	0.83	0.66
Below median reported effort	0.78	0.61
Effort check passed	0.82	0.68
Effort check failed	0.81	0.61
Reliability		
Above median reported reliability	0.82	0.64
Below median reported reliability	0.76	0.63
Response times		
Excluding Top 10% speeders	0.81	0.64
Top 10% speeders	0.76	0.64
Excluding slowest & fastest 25%	0.82	0.66
Slowest & fastest 25%	0.81	0.61
Averaging		
Averaging responses (2 repetitions)	0.87	0.72
Further methods		
ORIV	0.93	0.77

Notes: The table reports test-retest correlation coefficients of the general risk question (Column (1)) and the general altruism question (Column ((2)). The test-retest correlation is computed using Spearman's rank correlation coefficient between the question in wave 1 and wave 2 of the survey. Each row presents estimates from separate regressions using different subsamples or methodological adjustments. See Table 3 for details.

Table F.11: Improving the relationship between self-assessments and incentivized behavior in the SOEP

Dependent variable:	Investment in stocks			Performance pay			Smoking		
	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median	Full sample	$\hat{\tau} \geq$ median	$\hat{\tau} <$ median
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Risk question	0.070***	0.086***	0.052***	0.013***	0.017***	0.008***	0.020***	0.023***	0.018***
	(0.003)	(0.004)	(0.004)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	9095	4547	4548	5758	2879	2879	15162	7581	7581
R^2	0.083	0.114	0.052	0.009	0.014	0.004	0.012	0.017	0.009

Notes: The table reports OLS estimates, with binary dependent variables taking the values zero and one. Regressions are based only on respondents who are 18 years or older, and those for performance pay include only respondents up to the age of 66 who work full-time and receive wages or salaries. *Investment in stocks* are, in the SOEP, a residual category of securities without a fixed interest rate, like stocks or options ("other securities"). Since the relevant question was asked on the household level in 2010, the units of observation in the respective regressions are averaged on the household level in that year. *Performance pay* indicates that an employee receives payments from profit-sharing, premiums, or bonuses, asked 2009. Smoking refers to 2010. The independent variable refers to the respective domain-specific question asked in the SOEP. The contexts are *financial matters* for holding risky financial securities, *career* for performance pay, and *health* for smoking. Robust standard errors in parentheses. Significance levels: $^{*}\,p < 0.1$,$^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table G.1: Choice categories

0	1	2	3	4	5	6	7	8	9	10
<1.56	1.56–1.60	1.61–1.65	1.66–1.70	1.71–1.75	1.76–1.80	1.81–1.85	1.86–1.90	1.91–1.95	1.96–2.00	>2.00
0.1%	0.8%	3.8%	11.1%	21.1%	26.1%	21.1%	11.1%	3.8%	0.8%	0.1%

Note: top row: categories; middle row: height (in meters); bottom row: respective likelihoods.

G Additional experimental evidence

In the following, we describe the design and results of an additional experiment. The experiment was run before the experiment described in the main text (Section 4). Compared to the main experiment, this experiment was conducted in the lab and not online and used a different estimation paradigm. Otherwise, the core aspects of the experiment are the same, so that the hypotheses described in Section 4.2 are also applicable.

G.1 Design

Types. We presented subjects with a series of abstract figures. Specifically, subjects saw a total of 60 screens, each showing a stylized male figure of varying size (see Figure G.1). On each screen, the figure was randomly located at one of four different parts of the screen, i.e., either the upper left, upper right, lower left, or lower right part of the screen. The sizes of the figures were drawn from a normal distribution that closely matches the actual height distribution of men in Germany (based on data from the Socio-economic Panel, SOEP). In particular, sizes were grouped into eleven categories (in meters) with likelihoods as shown in Table G.1.

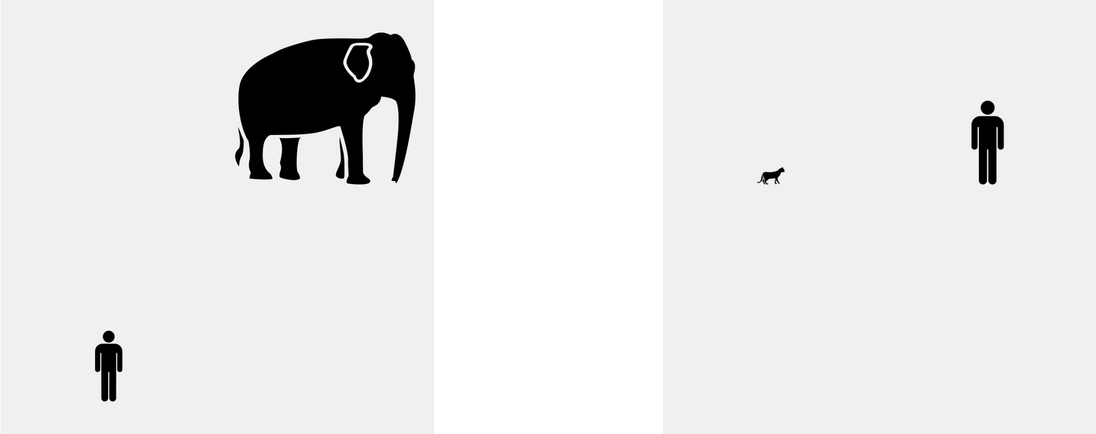
For example, Category 3 represents male persons of heights between 1.66 and 1.70 meters, occurring with a likelihood of 11.1%. Subjects received a handout showing this distribution and the corresponding figures underneath. Subjects were informed that 15 distinct sizes were independently drawn from the eleven categories and shown four times. Specifically, subjects saw four blocks, each comprising these 15 distinct sizes. This procedure hence implements a panel structure, i.e., for each subject i, we observe a total of 60 reports for $K = 15$ characteristics in $T = 4$ periods. The location of the male figures was randomly determined for each screen.

To facilitate the estimation task and vary the presentation style of the screens, subjects also saw a "reference category" next to the male figure, i.e., either an elephant or a cat (see Figure G.1). Subjects were informed that—unlike the male figures—the size of the two animals was always exactly the same. The height of the elephant was 3.50 meters, and it was 0.40 meters for the cat. Conditional on the randomly determined location of the male figures, the location and type of the reference category (elephant or cat) were also randomly drawn for each screen.

Payoff Function. We implemented the same payoff function as in the main experiment.

Signal Precision and Treatments. To exogenously vary the precision τ of the signal, we ran two between-subject treatments that only differed in terms of how long subjects saw each of the 60 screens. In the treatment *Long*, subjects saw each screen for 7.5 seconds, in contrast to treatment *Short*, where they saw each screen only for 0.5 seconds.

Figure G.1: Example screens

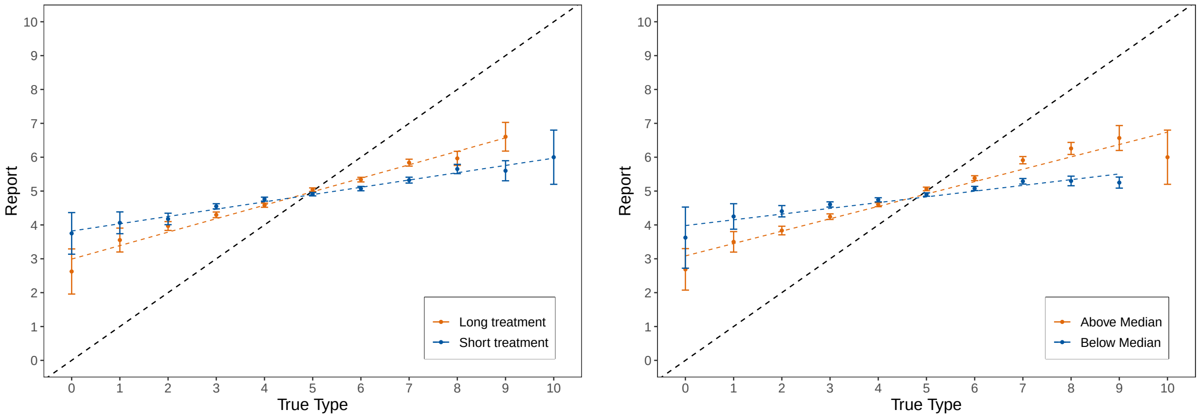


(a) Size within Category 2

(b) Size within Category 8

Note: The panel on the left shows a male figure with a height of 1.63m along with the elephant, which is 3.50m tall. The male figure on the right corresponds to 1.93m, and the cat has a height of 40cm. Animal pictograms adapted with permission from Storey (2016).

Figure G.2: Relationship between reports and types in the experiment



(a) Reports vs. true types split by treatment

(b) Reports vs. true types split by estimated τ

Procedural Details. In total, 199 subjects—mostly undergraduate university students from all majors—took part in the experiment, 101 subjects in the treatment *Long* and 98 in the treatment *Short*. We used z-Tree as the experimental software (Fischbacher, 2007). Subjects were recruited using the software hroot (Bock, Baetge, and Nicklisch, 2014). At the beginning of an experimental session, participants received detailed information about the rules and the structure of the experiment. In all treatments, the experiment only started after all participants had correctly answered several control questions. The experiments were run at the BonnEconLab as a laboratory experiment and subjects received a show-up fee of €5 for participation.

G.2 Results

When presenting the results, we closely follow the structure of the main experiment.

Results on the Relationship between Types and Responses. We start with Hypothesis 1, for which Figure G.2 Panel A provides the visual test. We find linearity in types and slope coefficients

Table G.2: Relationship between reports and true types

	Dependent variable: True type				
Subjects	*all*	*by treatment*		*by $\hat{\tau}$*	
		Short	*Long*	*low*	*high*
	(1)	(2)	(3)	(4)	(5)
Report	0.478^{***}	0.416^{***}	0.518^{***}	0.304^{***}	0.600^{***}
	(0.025)	(0.035)	(0.034)	(0.025)	(0.032)
Constant	2.571^{***}	2.917^{***}	2.336^{***}	3.506^{***}	1.887^{***}
	(0.131)	(0.181)	(0.178)	(0.132)	(0.169)
Subjects	188	89	99	89	99
Observations	11,280	5,340	5,940	5,640	5,640
Report $\neq 1$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
R^2	0.139	0.085	0.192	0.049	0.243
ΔR^2		127%		394%	

Notes: The table reports OLS estimates. Standard errors (in parentheses) are clustered at the subject level. $^{*}\, p < 0.1$, $^{**}\, p < 0.05$, $^{***}\, p < 0.01$.

that are less steep than the 45-degree line. Across both treatments, the slope coefficient is 0.478, indicating a bias towards the population average (see Column (1) of Table G.2 for details). Next, we turn to Hypothesis 2. As displayed in Figure G.2 Panel A, we find that the slope of types on reports is steeper in the *Long* treatment compared to the *Short* treatment. This is quantified in Columns (2) and (3) of Table G.2. In addition, we find again a significant interaction between reports and treatment ($p = 0.037$). Turning to the R^2 and Hypothesis 3, we find an increase of 127% in R^2 when going from the *Short* to the *Long* treatment. Accordingly, we find evidence for all three hypotheses.

Results on the Performance of the Self-knowledge Estimator. For Hypothesis 4, we find that the average $\hat{\tau}$ in the *Long*-treatment is 0.94 (median 0.24), while it is 0.16 (median 0.08) in the *Short*-treatment, a significant difference ($p < 0.001$, t-test). As displayed in Figure G.3, we again find that the distribution of $\hat{\tau}$ in *Long* stochastically dominates the distribution of $\hat{\tau}$ in *Short*. With respect to Hypothesis 5, we display visual results in Figure G.2 Panel B. As in the main experiment, we find that reports are more biased towards the population average among subjects with below median $\hat{\tau}$ compared to subjects with above median $\hat{\tau}$. The difference in slopes is quantified in Columns (4) and (5) of Table G.2, which display the respective regression slopes. Again, the interaction term between reports and an indicator of the median split is significantly different from zero ($p < 0.001$). Lastly, we turn to Hypothesis 6. As displayed in Table G.2, the R^2 increases by 349% when moving from below to above median $\hat{\tau}$ subjects. On the individual level, we find a correlation of 0.83 ($p < 0.001$, two-sided) between the individual-level R^2 and the slope coefficients. Moreover, the rank correlation between the individual-level R^2 and $\hat{\tau}$ is 0.83 ($p < 0.001$, two-sided, Pearson: 0.79). Lastly, the correlation between the alternative estimate that uses information about true types and our estimator is 0.83 ($p < 0.001$, two-sided, Pearson: 0.98). Hence, we find evidence for all three hypotheses related to the performance of $\hat{\tau}$.

Figure G.3: Distribution of estimated τ in the experiment

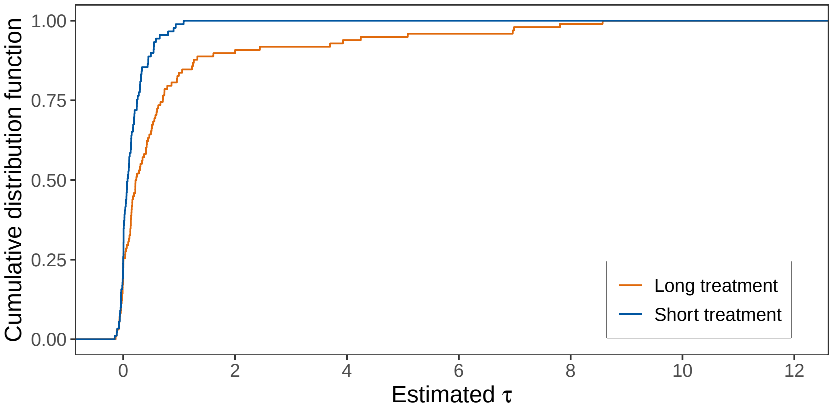
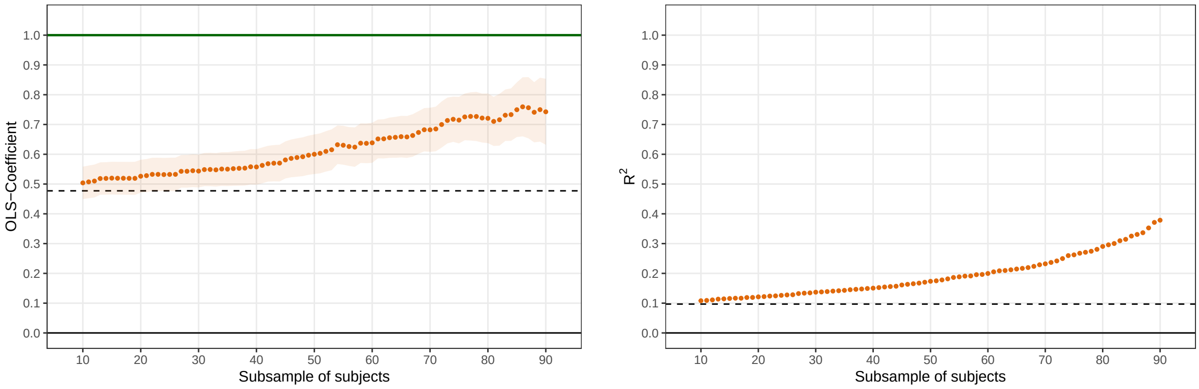


Figure G.4: Relationship between reports and types in the experiment



(a) Simulating the influence of subsetting based on $\hat{\tau}$ on coefficients

(b) Simulating the influence of subsetting based on $\hat{\tau}$ on R^2

Notes: **Panel (a):** Each dot represents the average OLS-coefficient obtained from running 1,000 simulations, where in each we regress a simulated outcome variable based on the true number of dots on subjects' reported number of dots in the experiment. Each regression is run on a subsample where a percentage of subjects with the lowest estimated level of self-knowledge $\hat{\tau}$ are excluded. The x-axis denotes the respective percentage removed, i.e., the first dot denotes the sample where the 10% of subjects with the lowest $\hat{\tau}$ are removed. The solid line represents the true coefficient $\beta = 1$. The dashed line represents the full sample coefficient. Shaded areas indicate the average 95% confidence interval across all simulations. **Panel (b):** Each dot represents the average R^2 obtained from the regressions simulated in Panel (a). The dashed line represents the full sample R^2.

Results on the Impact of Sample Splitting on Regression Estimates. Lastly, we run the same simulation as in the main experiment using the results of the additional experiment. Figure G.4 displays the results. As in the main experiment, we find that the OLS coefficient is biased away from the true coefficient, and that subsetting based on $\hat{\tau}$ brings the coefficient closer to the true value. Moreover, subsetting increases the regression's R^2. Hence, we also replicate the evidence that supports Hypothesis 7.

H Self-reports as the Dependent Variable

In Section 2.4, we analyzed regression estimates when survey reports were the independent variable in linear regressions. Here, we provide the complementary analysis when reports are the dependent variable. An example could be analyzing how survey responses such as self-assessed risk aversion differ based on gender. Formally, assume that we want to estimate the following equation:

$$\theta_i = \beta_0 + \beta_1\, y_i + \epsilon_i \,,$$

where y_i is the respective realization of the independent variable and ϵ_i is an i.i.d. error term with an expected value of zero that is independent of y_i and the signals that subjects receive. Crucially, as in the case of self-reports as the independent variable, θ_i is not observable and instead replaced with the response r_i. We again use the notation involving subjective self-knowledge (see Section 2.2). The asymptotic result of the standard OLS estimator is derived below.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\widehat{\text{cov}(r_i, y_i)}}{\widehat{\text{var}(y_i)}} \xrightarrow{p} \frac{\text{cov}(r_i, y_i)}{\text{var}(y_i)} = \frac{\mathbb{E}[(r_i - \bar{r})\,(y_i - \bar{y})]}{\mathbb{E}\Big[(y_i - \bar{y})^2\Big]} = \frac{\mathbb{E}\Big[\frac{\tilde{\tau}(x_i - \bar{\theta})}{1+\tilde{\tau}}\,(y_i - \bar{y})\Big]}{\mathbb{E}\Big[(y_i - \bar{y})^2\Big]} \\
&= \frac{\tilde{\tau}}{1+\tilde{\tau}}\,\frac{\mathbb{E}\big[\big(x_i - \theta_i + \theta_i - \bar{\theta}\big)\,(y_i - \bar{y})\big]}{\mathbb{E}\Big[(y_i - \bar{y})^2\Big]} = \frac{\tilde{\tau}}{1+\tilde{\tau}}\,\frac{\mathbb{E}[(x_i - \theta_i + \beta_1\,(y_i - \bar{y}) + \epsilon_i)\,(y_i - \bar{y})]}{\mathbb{E}\Big[(y_i - \bar{y})^2\Big]} \\
&= \frac{\tilde{\tau}}{1+\tilde{\tau}}\,\frac{\mathbb{E}[\beta_1\,(y_i - \bar{y})\,(y_i - \bar{y})]}{\mathbb{E}\Big[(y_i - \bar{y})^2\Big]} = \frac{\tilde{\tau}}{1+\tilde{\tau}}\,\beta_1
\end{aligned}$$

$$\hat{\beta}_0 \xrightarrow{p} \bar{\theta} - \beta_1\,\bar{y} = \beta_0 + \beta_1\,\bar{y} - \frac{\tilde{\tau}}{1+\tilde{\tau}}\,\beta_1\,\bar{y} = \beta_0 + \left(1 - \frac{\tilde{\tau}}{1+\tilde{\tau}}\right)\beta_1\,\bar{y}$$

Thus, regression coefficients will be biased in the presence of limited self-knowledge, irrespective of whether subjects know their own level of self-knowledge or not. Moreover, as long as a decrease in τ is accompanied by a decrease in $\tilde{\tau}$, the overall effect on the absolute value of the slope parameter β_1 is strictly negative. That is, a lack of self-knowledge leads to an attenuation effect on regression estimates.

I An Estimator for Self-knowledge Based on Known True Types

Suppose we know that τ is constant in the relevant population, or, alternatively, that all answers were given by the same individual. Suppose also that we know the true types, and we use them as the independent variable, i.e., $y_i = \theta_i$ for all i. It follows that $\beta_0 = 0$, $\beta_1 = 1$, and $\bar{y} = \bar{\theta}$. For predicted answers, it follows that

$$\hat{r}_i \xrightarrow{p} \bar{\theta} + \frac{\tilde{\tau}}{1+\tilde{\tau}}\,\big(\theta_i - \bar{\theta}\big)\ .$$

For the model fit, it holds that

$$
\begin{aligned}
R^2 &= 1 - \frac{\sum_{i=1}^{I}\left[(r_i - \hat{r}_i)^2\right]}{\sum_{i=1}^{I}\left[\left(r_i - \frac{1}{I}\sum_{i=1}^{I} r_i\right)^2\right]} \xrightarrow{p} 1 - \frac{\mathbb{E}\left[\left(r_i - \bar{\theta} - \frac{\tilde{\tau}}{1+\tilde{\tau}}\left(\theta_i - \bar{\theta}\right)\right)^2\right]}{\mathbb{E}\left[(r_i - \bar{r})^2\right]} \\
&= 1 - \frac{\mathbb{E}\left[\left(\bar{\theta} + \frac{\tilde{\tau}}{1+\tilde{\tau}}\left(x_i - \bar{\theta}\right) - \bar{\theta} - \frac{\tilde{\tau}}{1+\tilde{\tau}}\left(\theta_i - \bar{\theta}\right)\right)^2\right]}{\mathbb{E}\left[\left(\bar{\theta} + \frac{\tilde{\tau}}{1+\tilde{\tau}}\left(x_i - \bar{\theta}\right) - \bar{\theta}\right)^2\right]} \\
&= 1 - \frac{\left(\frac{\tilde{\tau}}{1+\tilde{\tau}}\right)^2 \mathbb{E}\left[(x_i - \theta_i)^2\right]}{\left(\frac{\tilde{\tau}}{1+\tilde{\tau}}\right)^2 \mathbb{E}\left[(x_i - \bar{\theta})^2\right]} = 1 - \frac{\frac{\sigma^2}{\tau}}{\sigma^2 + \frac{\sigma^2}{\tau}} = \frac{\tau}{1+\tau}\,.
\end{aligned}
$$

Rearranging yields that ${}^{R^2}/{}_{1-R^2}$ is a consistent estimator for τ.

J Testing Updating Behavior in the Experiment

In our model, individuals update based on Bayes' rule. Since we observe the actual number of dots (true type) in our experiment, we can construct the expected response that we would observe if subjects adhered to Bayesian updating. We can then compare this benchmark with the response that subjects actually give to test for systematic deviations from Bayesian updating.

Specifically, using $\hat{\tau}$ as an estimator of τ, we can use equation (3) to construct the expected report subjects should provide under Bayesian updating in the presence of subjective self-knowledge for an image with θ_{ik} dots:

$$
\mathbb{E}[r_{ik}|\theta_{ik}] = \frac{\bar{\theta} + \hat{\tau}\theta_{ik}}{1+\hat{\tau}} \tag{20}
$$

We then compare the expected report with the response that subjects actually provide for the image with θ_{ik} dots.

Results. Two interesting empirical patterns emerge. First, on average, subjects' reports are quite close to the Bayesian benchmark, with an average difference of 0.16 and a median difference of -0.01. That is, subjects are not systematically biased away from Bayesian updating. Second, subjects' reports have a substantially higher variance than what would be expected under Bayesian updating. While the realized variance is 2.56, the expected variance is only 1.71. One reason for the excessive volatility in responses may be subjects overestimating their level of self-knowledge (Section 2.2). If subjects overestimate their level, we should observe that regression estimates are biased towards zero following Proposition 1. Indeed, in our simulation in Section 2.4 we observe an attenuation effect.

K Research transparency

The experiment covered in Section 4 and the survey covered in Section 5 were preregistered at aspredicted.org (https://aspredicted.org/57B_R97 and https://aspredicted.org/R5L_L5W). The preregistrations include details on the experimental design, the planned sample size, exclusion criteria, hypotheses, and the main analyses. In the following, we describe the mapping between preregistration and the paper.

Experiment. For the experiment, we preregistered a target sample size of 300. In the actual experiment, 308 subjects took part. As preregistered, we exclude 10 subjects who gave the same estimate for all 60 images of the experiment. We further excluded one subject who was likely using computer-assisted tools to help with the dot estimation because the subject gave the correct estimate in every image (non-preregistered). The preregistration contained four hypotheses. In the paper (Section 4.2), we formulate seven hypotheses. With the exception of the last hypotheses, the hypotheses contained in the paper are reformulations and extended versions of the preregistered hypotheses. The four hypotheses contained in the preregistration are described in the paper, and their preregistered empirical tests are in Section 4.2. Accordingly, the empirical test of hypothesis 7 and thus the analyses of Sections 4.2.3 were not preregistered.

Survey. For the survey, we preregistered a sample size of 1,000 for the first session, and expected between 750 and 850 of subjects to complete both sessions. In the actual survey, 1,001 subjects completed the first, and 740 the second session. As preregistered, we exclude those 261 subjects who did not complete both sessions. We exclude 6 further subjects because we cannot estimate $\hat{\tau}$ for them. The preregistered comparison of methods is displayed in Table 5. The additionally preregistered analyses of correlations between our self-knowledge estimator and other methods are reported in Table F.3, while the scale use analysis mentioned in the preregistration is reported in Section 5.4.2.

L Experimental Instructions

L.1 Instructions Main Experiment Section 4

The instructions have been translated from German. Horizontal lines are used to separate screens.

Welcome

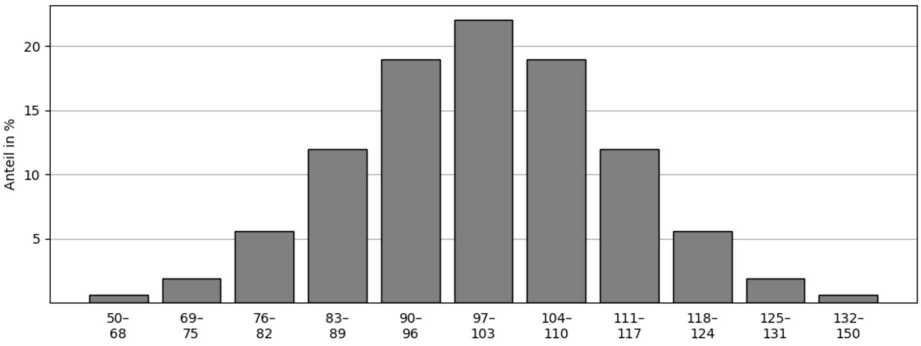
Welcome and thank you for participating in today's study!
For your participation, you will receive a flat fee of 4€, which is going to be paid out in cash at the end of the study. During the study, you will respond to estimation tasks. Depending on the quality of your answers, you can additionally earn up to 10€. You will receive all payments via bank transfer.
On the following screens, everything will be explained in detail. Please do not use phones or other aids during the study.
Please now click on "Continue" to proceed.

Your Task

Generally, your task in this study is to estimate the number of dots displayed to you in images. The dots are light gray, making them difficult to see. The more accurately you estimate the number of dots, the more money you can earn. Later, you will see a series of images with varying numbers of dots. Here is an example of such an image:
[Picture of a dot figure]
Each image can contain between 50 and 150 dots. The number of dots is divided into a total of 11 numerical groups:

Numerical group:	1	2	3	4	5	6	7	8	9	10	11
Number of dots:	50–68	69–75	76–82	83–89	90–96	97–103	104–110	111–117	118–124	125–131	132–150

The number of dots in the images is randomly generated. The following applies: Images with very few or very many dots are relatively rare. Images with around 100 dots are the most common. The further the number of dots in an image deviates from 100, the rarer the image becomes.
Below, you can see the frequency distribution of the individual numerical groups, as well as additional examples of images:



As you can see, the following applies:

- The most common number of dots is 97–103, occurring in 22% of all cases.
- The second most common ranges are 90–96 or 104–110 dots (each 18.9%).
- The third most common ranges are 83–89 or 111–117 dots (each 12%).

Much rarer are particularly high or particularly low numbers of dots:

- 76–82 or 118–124 dots occur in only 5.6% of cases.
- 69–75 or 125–131 dots occur in just 1.9% of cases each.
- Very rarely, 50–68 or 132–150 dots appear (each 0.6%).

It is important that you understand the relative frequencies of the number of dots, as the images you will later see are drawn from this distribution.

Procedure

You will see a total of 60 images, one after the other. For this, we will randomly select 15 different numbers of dots from the distribution we jsut showed you. Each of the selected numbers of dots will be displayed to you a total of 4 times, with the positions of the dots on the images potentially differing each time.

You will first see a countdown in seconds. Once the countdown ends, an image will be shown to you for [0.5/7.5] seconds. Afterward, the following question will be asked:

How many dots did you see?

You can provide your answer on the following scale:

below average						above average				
○	○	○	○	○	○	○	○	○	○	○
50–68	69–75	76–82	83–89	90–96	97–103	104–110	111–117	118–124	125–131	132–150

Your Payoff

For each shown picture, there is exactly one correct answer (one numerical group). If, for example, the number of dots shown is 98, the correct answer would be “97–103”. You must select one answer option for each image. At the end of today’s study, one of the displayed images will be randomly selected for you. Your payment, in addition to the fixed 4€, will depend on the answer you gave for this image.

If you chose the correct answer option, you will receive an additional 10€. The further your answer is from the correct option (how many steps to the left or right you would have needed to click), the more money will be deducted from the 10€. The deviation (steps to the left or right) will be squared and multiplied by 0.10€. For example, if the deviation is two steps, $2^2 \times 0.10 = 0.40$€ will be deducted. You would then receive $10.00 - -0.40 = 9.60$€ as an additional payment. The maximum deviation is ten steps (if, for example, you answered “132–150” but the correct answer was “50–68”). In that case, the entire 10€ would be deducted.

The less deviation there is between your chosen answer option and the correct answer option, the more money you will receive. The following table provides an overview of the possible deductions and the resulting additional payments.

Deviation (steps)	0	1	2	3	4	5	6	7	8	9	10
Deduction (€)	0.00	0.10	0.40	0.90	1.60	2.50	3.60	4.90	6.40	8.10	10.00
Additional payment (€)	10.00	9.90	9.60	9.10	8.40	7.50	6.40	5.10	3.60	1.90	0.00

Control Questions

Please respond to a few questions regarding your comprehension. Click the button to display a summary of the information from the previous pages.
Question 1

- In each case, which of the two is more likely: The picture shows . . . dots
  - ○ 97–103 132–150 ○
  - ○ 104–110 97–103 ○
  - ○ 97–103 90–96 ○
  - ○ 83–89 104–110 ○

Question 2

- How much money would be deducted from the additional 10€?
  - The correct answer would be “97–103”. You answered “132–150” (deviation of 5 steps).
  - The correct answer would be “132–150”. You answered “50–68” (deviation of 10 steps).
  - The correct answer would be “97–103”. You answered “104–110” (deviation of 1 step).
  - The correct answer would be “111–117”. You answered “97–103” (deviation of 2 steps).

You answered all the comprehension questions on the previous page correctly.
There is one more comprehension question. Click the button to display a summary of the information from the previous pages.
Question 3

- Lets assume you missed the image with the dots. You still have to provide an estimate. What is the best response in this case?

Trial Run

You have answered all the comprehension questions correctly.
Before you see the 60 images and estimate the number of dots in each, a practice round will take place. You will see 3 images and, after each one, indicate how many dots were shown in the image. Unlike later, you are afterward informed about the correct answer.
This trial run is unrelated to the final payout and is meant to introduce you to the task. The pictures will be displayed for [0.5/7.5] seconds, exactly as in later rounds.
When you are ready, click “Begin”.

Practice task 1/3

[Countdown]

[Picture]

How many dots did you see?
You can provide your answer on the following scale:

below average						above average				
○	○	○	○	○	○	○	○	○	○	○
50–68	69–75	76–82	83–89	90–96	97–103	104–110	111–117	118–124	125–131	132–150

[Picture]
Correct answer: [e.g., 50–68]
Your answer: [e.g., 132–150]

Two more practice rounds.

Beginning of the Main Part

Thank you for completing the trial rounds.
You can now begin with the main part of the study. At the end of the study, one of your following responses will be chosen and determine how much additional money you earn.

Task [n]/60

---

60 rounds like the practice rounds but without feedback.

Further Questions

Thank you for completing the main part.
Please now also respond to a few more additional questions.
How difficult did you feel was the task? [very easy – very difficult; 11-point scale]
How sure were you about your responses? [very unsure – very sure; 11-point scale]
How much can be trust your responses? [not at all – fully; 11-point scale]

Personal Details

How old are you?
What is your gender? Female male diverse
Are you a student? If yes, what is your major?
Have you ever been enrolled in economics as major?
How much many (in €) do you have available to you per month after deducting all costs?

Thank you for your participation in this study!
You will receive a flat payment of 4€.
In addition, answer no. [n] was chosen to determine your additional payoff. Due to the deviation of your answer from the correct answer you will additionally receive [X1] euros and [Y1] cents.
Hence, in total you receive [X2] euros and [Y2] cents. You will receive the money within three weeks.
This study is now finished. You can close this tab.

L.2 Instructions Additional Experiment Appendix Section G

The instructions have been translated from German. Horizontal lines are used to separate screens.

Welcome

Welcome and thank you for participating in today's study!

For your participation, you will receive a flat fee of €5, which is going to be paid out in cash at the end of the study. During the study, you will respond to estimation tasks. Depending on the quality of your answers, you can additionally earn up to €10. On the following screens, everything will be explained in detail.

During the study, communication with other participants is not allowed and the curtain of your cubicle has to remain closed. Your cellphone has to be switched off and no aids are permitted. On the computer, only use the designated functions and use the mouse and keyboard to make inputs. If you should have any questions, please stick your hand out of the cubicle. One of the experimenters is then going to approach you.

Please now click on "Continue" to proceed.

Your Task

Generally, your task in this experiment is to estimate the height of stylized depictions of men. The more precisely you estimate, the more money you can earn. For that, you will, later on, see a series of pictures with men of different heights.

More precisely, the men are going to be depicted as "stick figures." An example is shown below.

[Picture of a male stick figure]

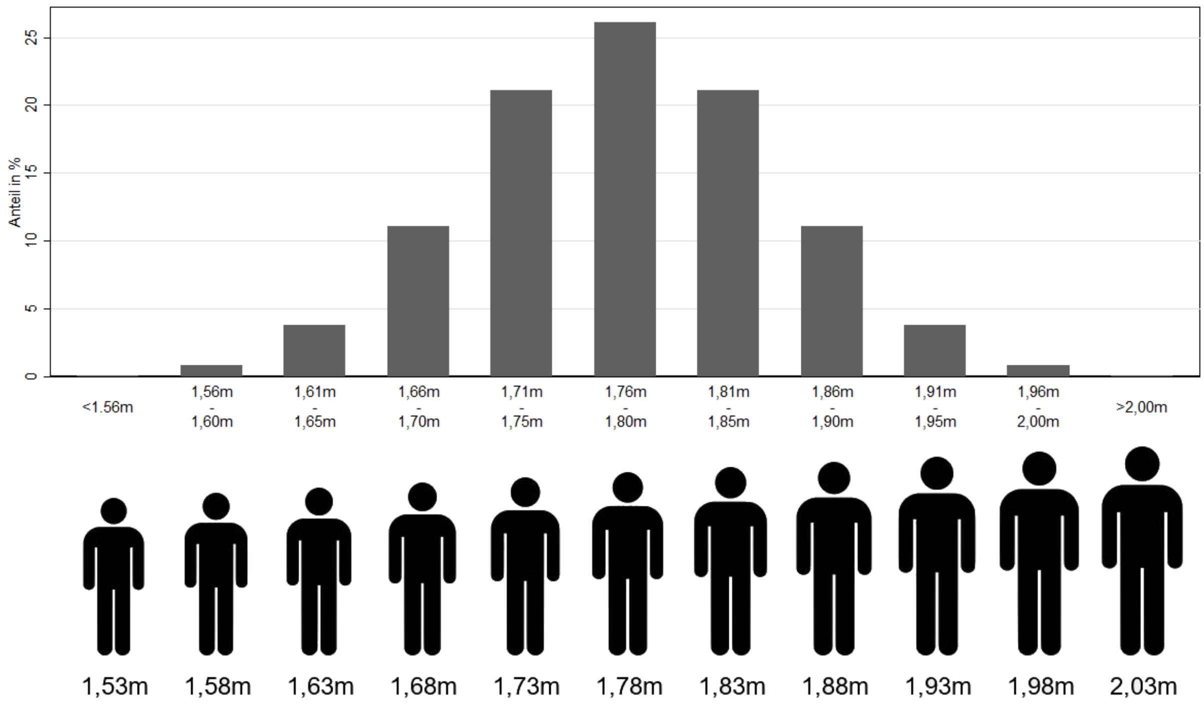
The men are split into eleven categories, depending on their body heights:

at most 1.55m	1.56m–1.60m
1.61m–1.65m	1.66m–1.70m
1.71m–1.75m	1.76m–1.80m
1.81m–1.85m	1.86m–1.90m
1.91m–1.95m	1.96m–2.00m
at least 2.01m	

Body Heights

As you know, very short and very tall men are found rather infrequently. Most common are men of around 1.78m. Exactly the same holds for the pictures that you are going to see later on. The pictures are informed by the actual height distribution among men in Germany. For that, the data from a large, representative sample of more than 20,000 people in Germany were used. The frequency of observing men of a given height is depicted in the image below.

For your orientation, we have also printed this image for you. It is lying on your desk.



Body Heights

On the image (on your desk) you see eleven different body heights. For each body height, it is said how often it is observed in the German population. Most common are men of a body height of 1.76m–1.80m, with 26.1%. The second most common are men with body heights of 1.71m–1.75m or 1.81m–1.85m (21.1% each). The third most common is a height of 1.66m–1.70m or 1.86m–1.90m (11.1% each). Considerably less common are very tall and very short men. Heights of 1.61m–1.65m or 1.91m–1.95m occur in only 3.8% of observations, heights of 1.56m–1.60m or 1.96m–2.00m each in only 0.8% of cases. Very uncommon are heights under 1.56m and above 2.00m (0.1% each).
It is important that you understand the relative frequencies of heights since the pictures that will be shown later are drawn from the displayed distribution. Thus, it is considerably more likely that you will see a man with a body height of 1.75m or 1.81m than a man with a body height of 1.58m or 2.03m.
To make the estimation of the body heights easier for you, every picture that will be displayed is accompanied by either a cat or an elephant. The cat has a height of 40cm, and the elephant is 3.50m tall (each at its highest points). In the picture below, you see an average man with a height of 1.78m next to the cat and the elephant, respectively.
[two example images here, as described]

Procedure

You will be shown a series of 60 pictures. For this purpose, we will randomly draw 15 different heights from the distribution in the population. Every drawn height will be shown to you four times in total. The accompanying animal and the position on the screen may change.

You will first be shown a countdown in seconds. After the countdown has finished, you will be shown a picture for [0.5/7.5] seconds. Afterward, the following question will be asked:
How tall was the displayed person?
You can provide your answer on the following scale:

The height of the displayed person was …

below average						above average				
…–	1.56m–	1.61m–	1.66m–	1.71m–	1.76m–	1.81m–	1.86m–	1.91m–	1.96m–	2.01m–
1.55m	1.60m	1.65m	1.70m	1.75m	1.80m	1.85m	1.90m	1.95m	2.00m	…
○	○	○	○	○	○	○	○	○	○	○

Your Payoff

For each shown picture, there is exactly one correct answer (an interval). For example, if the height of the shown man should be 1.78m, then this would be the answer "1.76m–1.80m." You always have to select exactly one answer. **At the end of today's study, one of the shown pictures will randomly be selected for you.** Your answer for this picture then determines the payoff that you receive on top of the €5 flat fee.

If you have chosen exactly the correct option, you will additionally receive €10. The further away you were from the correct answer (how much further to the left or right you should have clicked), the more is deducted from the €10. For this, the deviation (steps to the left or right) is squared and multiplied by 10 cents. The maximal deviation is ten steps (e.g., if you have answered "2.01m–…" but "…–1.55m" would have been correct). In this case, the entire €10 would be deducted.

You receive more money, the fewer steps are between your selected answer and the correct answer. The table gives you an overview of the possible deductions and the resulting additional payments. A printed version of this table is also available at your desk.

Deviation (steps)	0	1	2	3	4	5	6	7	8	9	10
Deduction (€)	0.00	0.10	0.40	0.90	1.60	2.50	3.60	4.90	6.40	8.10	10.00
Additional payment (€)	10.00	9.90	9.60	9.10	8.40	7.50	6.40	5.10	3.60	1.90	0.00

Control Questions

Please respond to a few questions regarding your comprehension. Feel free to use the printout at your desk as an aid.

- In each case, which of the two is more likely: the picture depicts a man with a height of …
  - 1.76m–1.80m [correct]
    2.01m–…

  - 1.81m–1.85m
    1.76m–1.80m [correct]
  - 1.76m–1.80m [correct]
    1.71m–1.75m
  - 1.66m–1.70m
    1.81m–1.85m [correct]

- How much money would be deducted from the additional €10?
  - Correct would be “1.76m–1.80m.” You responded “2.01m–. . . .” [€2.50]
  - Correct would be “2.01m–. . . .” You responded “. . . –1.55m.” [€10.00]
  - Correct would be “1.76m-1.80m.” You responded “1.81–1.85m.” [€0.10]
  - Correct would be “1.86m-1.90m”. You responded “1.76m–1.80m.” [€0.40]
- Suppose you have missed the picture of the man, but you nonetheless must give an estimate. What is the best answer? [1.76m–1.80m]

Thank you for your responses! Please wait.

Trial Run

Before you see the 60 pictures and estimate the heights, there will first be a trial run. You will see ten pictures and subsequently have to estimate the height of the respective man you saw. Unlike later, you are afterward informed about the correct answer.
This trial run is unrelated to the final payout and is meant to introduce you to the task. The pictures will be displayed for [0.5/7.5] seconds, exactly as in later rounds.
When you are ready, click on “Begin”.

Practice task [n]/10

[Countdown]

[Picture]

How tall was the shown person?

The height of the displayed person was . . .

below average						above average				
. . . – 1.55m	1.56m– 1.60m	1.61m– 1.65m	1.66m– 1.70m	1.71m– 1.75m	1.76m– 1.80m	1.81m– 1.85m	1.86m– 1.90m	1.91m– 1.95m	1.96m– 2.00m	2.01m– . . .
○	○	○	○	○	○	○	○	○	○	○

Nine more practice rounds.

Correct answer: [e.g., 1.71m–1.75m]
Your answer: [e.g., 1.81m–1.85m]

Thank you for your responses! Please wait.

Beginning of the Main Part

Thank you for completing the trial rounds.
You can now begin with the main part of the study. At the end of the study, one of your following responses will be chosen and determine how much additional money you earn.

Task [n]/60

60 rounds like the practice rounds but without feedback.

Further Questions

Thank you for completing the main part.
Please now also respond to a few more additional questions.
How difficult did you feel was the task? [very easy – very difficult; seven-point scale]
How sure were you about your responses? [very unsure – very sure; seven-point scale]

Further Questions

Big Five questionnaire (BFI-S; Gerlitz and Schupp (2005))
Scale-use module
Bayesian updating question

Personal Details

Your gender: female male diverse
Your age (in years):
Your body height (in cm):

Do you have any final comments?

Thank you for your participation in this study!

You will receive a flat fee of €5.

In addition, answer no. [n] was chosen to determine your additional payoff. Due to the deviation of your answer from the correct answer you will additionally receive [X] euros and [Y] cents.

We will soon begin with the payouts. Please wait at your seat and keep the curtain of your cubicle closed until your cabin number is called. Then, please enter the adjoining room and remember to take the card on which your cabin number is printed with you and return it.

L.3 Instructions Survey Section 5

The instructions have been translated from German. Horizontal lines are used to separate screens. The order of items for the modules (e.g., Big Five) are randomized between subjects.

L.3.1 Wave 1

[Demographics]

How old are you (in years)?
[Integer field]

What is your gender?
[Male; Female; Diverse]

What is your highest general education qualification, or which category is closest to it?
[Currently a student or no educational qualification; Secondary school diploma (Hauptschulabschluss); Intermediate school diploma (Realschulabschluss); Higher education entrance qualification (Abitur) or vocational baccalaureate (Fachhochschulreife)]

What is your highest further education qualification, or which category is closest to it?
[No vocational qualification or higher education degree yet; Completed school-based/dual/civil service vocational training (schulische/duale/beamtliche Berufsausbildung); Completed university degree (Hochschulstudium)]

Please think about your household income from the past year from all sources and combining all members of your household. Into which size category did your annual household income fall after taxes and other deductions (net)?
[10,000€ or less; 10,001–20,000€; 20,001–30,000€; 30,001–40,000€; 40,001–50,000€; 50,001–60,000€; 60,001–100,000€; More than 100,000€; Prefer not to answer]

How satisfied are you currently, generally speaking, with your life?
Please respond on a scale from 0 to 10, where 0 means completely not satisfied, and 10 means completely satisfied. [Likert scale]

What is your mood at the moment?
Please respond on a scale from 0 to 10, where 0 means very bad and 10 means very good. [Likert scale]

The next question is about the following issue: in surveys like this one, there are sometimes participants who do not read the questions carefully and simply click through quickly. This affects the quality of scientific studies. To show that you are still reading the questions carefully, please answer "Orange" to the next question about your favorite color, regardless of what your actual favorite color is.
Given the above text, what is your favorite color? [Text field]

In the following, you will make a series of decisions. At the end of the study, one decisions out of all decision you will make is randomly selected. Each decision has an equal chance of being selected. The outcome of the selected decision will then be implemented after the study exactly as described.

Since one of your decisions will actually be carried out as described, you should carefully consider each decision.

[Incentivized Risk Decision]

In the following, you will make some decisions. At the end of the study, one of the decisions you make will be selected. Each decision has an equal chance of being selected. The consequence of the selected decision will then be implemented exactly as described in the decision after the study.

Since one of your decisions will actually be implemented, you should carefully consider each decision.

On the following screens, you have the opportunity to earn additional money as a bonus payment after completing this study. In total, you will make 11 decisions. In each decision, you will choose between two options. Option A remains the same in all decisions. Option B varies between decisions. You will make one decision per screen.

Option A: If you choose Option A, you have a 50% chance of receiving a bonus payment of 100 Lifepoints and a 50% chance of receiving no payment (0 Lifepoints).

Option B: If you choose Option B, you will receive a guaranteed bonus payment (100% probability). The amount varies between decisions, ranging from 0 Lifepoints in the first decision to 100 Lifepoints in the last.

To summarize: in each decision, you can choose between a lottery and a guaranteed payment.

The Consequence of Your Decision
As previously mentioned, at the end of the study, one of your choices will be randomly selected and implemented. This means: If you chose the lottery (Option A) in the selected decision, you have a 50% chance of receiving either 100 Lifepoints or 0 Lifepoints. ; If you chose the guaranteed payout (Option B) in the selected decision, you receive the corresponding amount with certainty.

[11 decisions between Option A and B with the amount of Option B going from 0 (Decision 1) to 100 Lifepoints (Decision 11) in increments of 10. Each decision is on a seperate page.]

[Incentivized Donation Decision]

The next page is about an additional amount of 1€. You can divide this amount between yourself and a charitable organization.

The charitable organization is the **Förderkreis für krebskranke Kinder und Jugendliche e.V.** [Support Group for Children and Adolescents with Cancer]. This association provides comprehensive support to young people with cancer and their families as they navigate the challenges of the illness.

You can decide how much of the 1€ you want to keep for yourself and how much you wish to donate to the Förderkreis.

The amount you choose to donate will be transferred to the Förderkreis after the completion of today's study. The amount you decide to keep for yourself will be paid to you as a bonus in the form of Lifepoints, provided this decision is selected for payout.

[Donation decision: allocating 1€ between themselves and charity]

[Altruistic behavior]

Have you engaged in any of the following activities in the past month?

Volunteered for a charitable organization? [No; Yes]

Helped a stranger or someone you didn't know who needed assistance? [No; Yes]

Gave something as a gift to another person? [No; Yes]

Donated money to a charitable organization? [No; Yes]

[Risk behavior]

Are you currently actively practicing a sport (i.e., at least once a month)? [No; Yes]

Have you or others in your household owned stocks (funds), bonds, or warrants in the past year? [No; Yes]

Do you currently smoke, whether cigarettes, pipes, or cigars? [No; Yes]

Are you currently self-employed, or have you been in the past year? [No; Yes]

[General risk question]

How do you see yourself: are you generally a person who is willing to take risks or do you try to avoid taking risks?

Please answer on a scale from 0 to 10, where 0 means not risk-taking at all and 10 means very risk-taking. You can grade your assessment with values in between. [Likert scale]

[Domain-specific risk assessments]

One may behave differently in various areas. How would you assess your risk-taking in the following areas?
Please answer each question on a scale from 0 to 10, where 0 means not risk-taking at all and 10 means very risk-taking. You can grade your assessment with values in between.

How is your risk-taking...
...in your professional career? [Likert scale]
...concerning your health? [Likert scale]
...in leisure and sports? [Likert scale]
...when driving? [Likert scale]
...with investments? [Likert scale]

[General altruism question]

How much would you be willing to give to a good cause without expecting anything in return?
Please answer on a scale from 0 to 10, where 0 means not willing at all, and 10 means very willing. You can grade your assessment with values in between. [Likert scale]

[Domain-specific altruism assessments]

One may behave differently in various areas. How would you assess your willingness to get involved for a good cause in the following areas? Please answer each question on a scale from 0 to 10, where 0 means not willing at all and 10 means very willing. You can grade your assessment with values in between.

How is your willingness...
...to help animals in need? [Likert scale]
...to help people who suffer from illness? [Likert scale]
...to help people in other countries who are suffering from hunger? [Likert scale]
...to support measures for the preservation of nature? [Likert scale]
...to enable disadvantaged people to participate in society? [Likert scale]

[Big Five questionnaire]

Our everyday actions are influenced by the core beliefs we hold.

Below, we describe different attributes that one can have. You will likely agree with some attributes and disagree with others. For some, you might feel undecided. Please answer each question on a scale from 0 to 10, where 0 means completely disagree, and 10 means completely agree. You can grade your assessment with values in between.

I am someone who...
...is sometimes a bit rough towards others. [Likert scale]
...can forgive. [Likert scale]
...is considerate and friendly towards others. [Likert scale]
...works thoroughly. [Likert scale]
...is rather lazy. [Likert scale]
...completes tasks effectively and efficiently. [Likert scale]
...is communicative, talkative. [Likert scale]
...is reserved. [Likert scale]
...can open up, is sociable. [Likert scale]
...often worries. [Likert scale]
...gets nervous easily. [Likert scale]
...is relaxed, can handle stress well. [Likert scale]
...is original, brings in new ideas. [Likert scale]
...values artistic experiences. [Likert scale]
...has a vivid imagination and ideas. [Likert scale]

[Effort question]

What is your opinion on daylight saving time?
Do you think it's a good idea to move the clocks forward by one hour in spring and back in autumn? Please write an answer that contains at least 25 words. [Text field]

[General risk question repeated version]

Which description fits you better: Do you tend to shy away from risks, or are you generally a risk-taker?
Please answer using the following scale, where 0 means **not risk-taking at all** and 10 means **very risk-taking**. You can grade your assessment with values in between. [Likert scale]

[Domain-specific risk assessments repeated version]

One may behave differently in various areas. How would you say, how risk-taking are you in the following areas?
Please answer using the following scale, where 0 means not risk-taking at all and 10 means very

risk-taking. You can grade your assessment with values in between.

How is it...
...outside of your working hours and in physical activities? [Likert scale]
...in your professional career? [Likert scale]
...in relation to your physical condition? [Likert scale]
...with finances? [Likert scale]
...in traffic? [Likert scale]

[General altruism question repeated version]

How willing are you to donate to a charitable cause without any personal benefit?
Please answer using the following scale, where 0 means no willingness at all and 10 means very high willingness. [Likert scale]

[Domain-specific altruism assessments repeated version]

One may behave differently in various areas. How would you assess your willingness to get involved for a good cause in the following areas?
Please answer on a scale from 0 to 10, where 0 means no willingness at all and 10 means very high willingness. You can grade your assessment with values in between.

How is your willingness...
...to show engagement for environmental protection? [Likert scale]
...to help integrate people in difficult life situations into society? [Likert scale]
...to provide assistance to people in other countries who are suffering from food scarcity? [Likert scale]
...to help people who have a disease?

[Big Five questionnaire repeated version]

Our everyday actions are influenced by the core beliefs we hold. Below, we describe different attributes that one can have. You will likely agree with some attributes and disagree with others. For some, you might feel undecided.

Please answer each question on a scale from 0 to 10, where 0 means Completely disagree and 10 means Completely agree. You can grade your assessment with values in between.

I am a person, who...
...is partly a bit rough. [Likert scale]

...forgives others. [Likert scale]
...is generally friendly and considerate to others. [Likert scale]
...places value on thoroughness in work. [Likert scale]
...has a tendency towards laziness. [Likert scale]
...approaches tasks efficiently and effectively. [Likert scale]
...enjoys communicating and talking with others. [Likert scale]
...tends to hold back. [Likert scale]
...enjoys sociability and opens up. [Likert scale]
...is often worried. [Likert scale]
...tends to be nervous. [Likert scale]
...remains relaxed even under stress. [Likert scale]
...stands out with new ideas and originality. [Likert scale]
...is interested in art. [Likert scale]
...is imaginative and has creativity. [Likert scale]

[Questionnaires]

It is very important for our study that we only include responses from participants who have dedicated their full attention to this study.
Otherwise, years of hard work (from the researchers and the time of other participants) could be wasted. You will definitely receive your payout for this study, but please let us know how much effort you have put into it.
"I have put ________ effort into this study."
["almost none"; "very little"; "a little"; "quite a lot"; "a lot"]

Additionally, there are often several distractions during studies (other people, phone, music, etc.). Please indicate how much attention you dedicated to this study. Once again, you will definitely receive your payment. We appreciate your honesty!

"I dedicated ___________ of my attention to this study."
["almost none"; "very little of my attention";"some of my attention"; "most of my attention"; "all of my attention"]

How reliable are your responses in this study?
[Likert scale]

How carefully did you answer the questions in this study?
[Likert scale]

The next question concerns the following issue: In surveys like this one, there are sometimes participants who do not read the questions carefully and just click through quickly. This affects the quality of scientific studies. To demonstrate that you are reading the questions carefully, please select both the answer option "Frequently" and the answer option "Rarely" for the next question.

With this in mind: How often do you participate in surveys?
[Very frequently; Frequently; Occasionally; Rarely; Very rarely]

L.3.2 Wave 2

Wave 2 repeats:

- General risk question
- Domain-specific risk assessments
- General altruism question
- Domain-specific altruism assessments
- Big five questionnaire
- General risk question repeated version
- Domain-specific risk assessments repeated version
- General altruism question repeated version
- Domain-specific altruism assessments repeated version
- Big five questionnaire repeated version

[Anchoring questions risk]

Please read through the following three scenarios and answer the associated questions.

Scenario 1: Maria is known for considering all possible risks before taking action. It is extremely important to her to avoid unforeseen negative events, even if it means missing out on potentially positive ones. For this reason, she very rarely ventures into unknown territory if it involves uncertainties.
Based on this information, how would you rate Maria's risk-taking on a scale from 0 to 10, where 0 means not at all risk-taking and 10 means very risk-taking?
[Likert scale]

Scenario 2: Laura is someone who occasionally seeks new experiences and challenges. As long as it doesn't get out of hand, she is willing to try unfamiliar experiences, provided the potential risks are manageable. In doing so, she sometimes makes decisions that others might perceive as bold.
Based on this information, how would you rate Laura's risk tolerance on a scale from 0 to 10,

where 0 means not risk-tolerant at all and 10 means very risk-tolerant?
[Likert scale]

Scenario 3: Sophie is someone who thrives on excitement and the allure of the unknown. She is driven by the thrill of uncertainty and often seizes opportunities that involve significant risks. She is characterized by a strong willingness to push boundaries, even if it entails potential setbacks.
Based on this information, how would you rate Sophie's risk tolerance on a scale from 0 to 10, where 0 means not risk-tolerant at all and 10 means very risk-tolerant?
[Likert scale]

[Anchoring questions altruism]

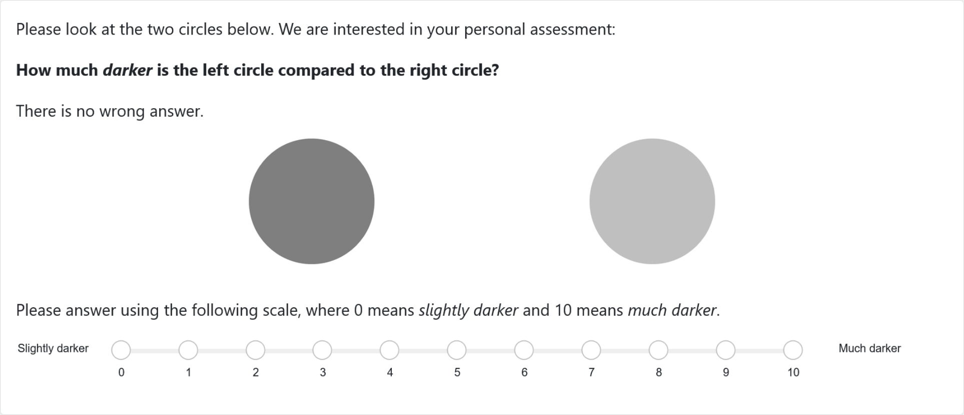
Please read through the following three scenarios and answer the associated questions.

Tim: Based on the description, Tim demonstrates a very low willingness to donate to charitable causes, as his decisions are driven by self-interest and he rarely donates unless he perceives a clear personal benefit.
On a scale from 0 to 10, where 0 means not willing to donate at all and 10 means very willing to donate, Tim's willingness to donate would likely be rated as 1. This reflects his reluctance while acknowledging a minimal possibility of donation under specific circumstances.
[Likert scale]

Markus: Markus shows a moderate willingness to donate to charitable causes, as he evaluates the potential societal impact and benefit before deciding to contribute. His willingness is situational and influenced by the perceived effectiveness of the donation.
On a scale from 0 to 10, where 0 means not willing to donate at all and 10 means very willing to donate, Markus's willingness to donate would likely be rated as 5. This reflects a balanced approach: he is open to donating under the right conditions but not highly committed to it.
[Likert scale]

Alex: Alex demonstrates a high level of altruism and a strong commitment to helping others through charitable donations, without expecting personal benefits in return. His motivation stems from the desire to contribute positively to society and prioritize the well-being of others.
On a scale from 0 to 10, where 0 means not willing to donate at all and 10 means very willing to donate, Alex's willingness to donate would likely be rated as 10. This reflects his generous and selfless nature in supporting good causes.
[Likert scale]

[Subjective scale use question 1]



Please look at the two circles below. We are interested in your personal assessment:

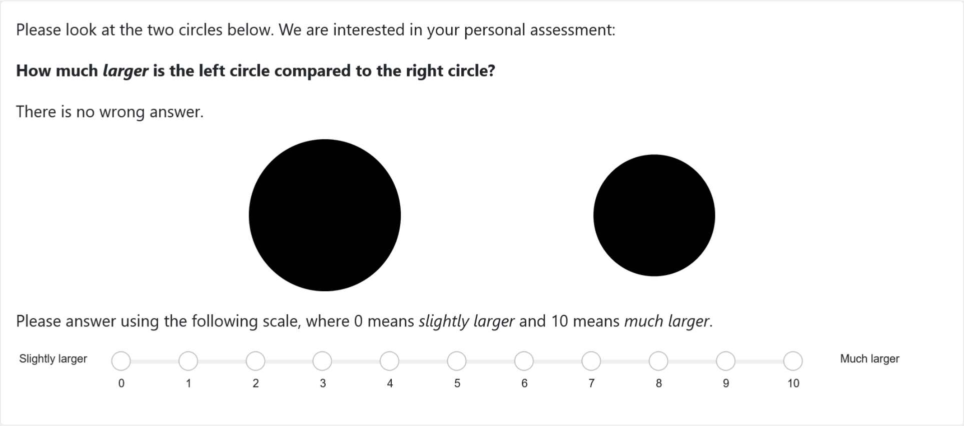
How much *darker* is the left circle compared to the right circle?

There is no wrong answer.

Please answer using the following scale, where 0 means *slightly darker* and 10 means *much darker*.

Slightly darker	○	○	○	○	○	○	○	○	○	○	○	Much darker
	0	1	2	3	4	5	6	7	8	9	10	

[Subjective scale use question 2]



Please look at the two circles below. We are interested in your personal assessment:

How much *larger* is the left circle compared to the right circle?

There is no wrong answer.

Please answer using the following scale, where 0 means *slightly larger* and 10 means *much larger*.

Slightly larger	○	○	○	○	○	○	○	○	○	○	○	Much larger
	0	1	2	3	4	5	6	7	8	9	10	