The Reliability of Questionnaires in Laboratory Experiments: What Can We Do?[§]

Forthcoming in: Journal of Economic Psychology, doi: 10.1016/j.joep.2019.102197

Irenaeus Wolff

Thurgau Institute of Economics (TWI) / University of Konstanz, Hauptstrasse 90, 8280 Kreuzlingen, Switzerland. wolff@twi-kreuzlingen.ch

1 Introduction

Questionnaires and surveys are a central research tool in many areas of economics and psychology, and they are also becoming more and more important in experimental economics. One example is the research area trying to connect personality traits or cognitive abilities with economic preferences and game-theoretic reasoning.[1] Other studies merely want to control for the effects of personality traits on choices when examining economic preferences. In either case, it is crucial that participants fill in the questionnaires faithfully. If they do not, the questionnaires do not measure what they are supposed to measure. But then, the finding that economic preferences are often unrelated to personality traits (*e.g.*, Becker et al., 2012) would be meaningless. And similarly, elicited personality traits would be meaningless as control variables. Hence, it is crucial to know how to obtain data that is as reliable as possible

[§]I am deeply indebted to Lisa Bruttel for her collaboration at an earlier stage of this project, as well as to Kate Bendrick for suggesting the specific variant of the "inBetween" payment procedure I use. Further, I would like to thank the lively research group at the Thurgau Institute of Economics (TWI) and the members of the Potsdam Center for Quantitative Research (PCQR) for helpful comments all along the way, and Marie Claire Villeval, Dirk Sliwka, Roberto Weber, as well as the participants of the 2014 GfeW meeting for fruitful discussions. Last but not least, I am indebted to Carlos Alós-Ferrer and two anonymous reviewers for their valuable comments. Konstantin Eitel provided valuable research support. His Master's thesis was part of this project. The questionnaire data is available under http://dx.doi.org/10.23663/x2620.

[1]See, *e.g.*, the substantive list of references Natalia V. Czap gathered already in 2011 (http://www-personal.umd.umich.edu/~nczap/Ref_PersonSc_Exp.htm, last accessed on 17th May, 2019). Recent examples are Kocher et al. (2019) or Proto, Rustichini, and Sofianos (2019).

Abstract:

Questionnaires eliciting personality traits and other characteristics of a person are important tools for many experimental economists. While a lot is known about how to run experiments and about how to construct and run field surveys, much less is known about how to administer such surveys in a post-experimental context. A short survey among experimental economists documents substantial heterogeneity in the procedures used, and in expectations about the effects of procedural details. I run an experiment on five aspects that are specific to the laboratory context. I find that (i) paying participants as soon as they finish the questionnaire yields a lower answer quality than waiting for all or an intermediate procedure; (ii) having participants enter their names for receipt preparation does not reduce answer quality (and does not increase the social desirability of answers); (iii) a higher overall payment increases answer quality, while (iv) framing the fixed part of participants' payment as being "for completing the questionnaire" as well as (v) progress feedback do not affect answer quality.

Keywords: Experimental economics, methods, survey, payment procedures.
JEL: C83, C91

when conducting post-experimental questionnaires. This is what this paper is about.

A lot is known about how to construct surveys (*e.g.*, Singer and Ye, 2013). We also know how to measure unreliable answers (*e.g.*, Meade and Craig, 2012). What we know little about is how to administer the questionnaires in the laboratory, and which laboratory-specific factors play a role in determining answer reliability. This paper focuses on the following five aspects: (i) the role of the payment order (in particular, whether and how it is linked to the questionnaire-completion order); (ii) the effect of asking for participants' names to prepare the receipts; (iii) the framing of the fixed element in participants' payments as compensation for showing up or for completing the questionnaire; (iv) the overall payoff participants receive; and (v) the existence of progress feedback.

Looking at these aspects is relevant. First, as I have pointed out, reliability of questionnaire data is crucial for a substantial number of studies also in experimental economics. And second, there is substantial heterogeneity with respect to how the aspects are handled in practice. So, while we may be well-aware of how to design and run incentivized experiments, there seems to be less of a consensus about how to administer post-experimental

questionnaires.

To substantiate these claims, I ran a short, anonymous survey on how experimentalists administer their post-experimental questionnaires, and on whether they would expect any of the above aspects to affect questionnaire-data quality.[2] Out of 94 respondents, 30% run post-experimental questionnaires eliciting personality traits or similar variables that go beyond the standard socio-demographic controls on a regular basis, and 56% do so occasionally. So, a vast majority of respondents are using the type of post-experimental questionnaires for which procedural details are potentially relevant. Furthermore, 59% at least "tend to agree" that such post-experimental questionnaires are becoming more and more important in the field of experimental economics, while only 11% "disagree" or "strongly disagree".[3]

Let us now turn to whether there is non-negligible heterogeneity in the administration of such questionnaires. To start with aspect (i), 24% of regular users and 32% of occasional users regularly start paying their participants as soon as they have completed the questionnaire (for 17% of regular and 13% of occasional users, this is the only payment procedure they use). In contrast, 50% of occasional users and 72% of regular users always wait until the last participant has finished before starting the cash-out. When compared to an intermediate procedure (waiting for 80% of the participants to finish, then start paying), waiting for the last participants is expected to lead to a higher data quality by 68% of the regular users but only by 33% of the occasional users. In turn, comparing the intermediate procedure to paying as soon as possible, around 60% would expect a better data quality under the intermediate procedure. Nonetheless, applying transitivity, 45% of occasional users and 33% of regular users do not expect a higher data quality when waiting for all compared to paying as soon as possible.[4]

Now take the question of whether to ask participants for their names to prepare the receipts (aspect ii). The most-widely used experimental software, z-Tree (Fischbacher, 2007), has its own pre-fabricated address form tailored to elicit participants' names. Amongst z-Tree-using respondents, 18% use this feature, whereas only 8% of those relying on other software ask for participants' names. Both among regular and occasional users, just over

[2]For recruitment, I used the ESA-discuss mailing list; 94 experimentalists from 25 countries completed the survey, among them 77 researchers from Western Europe and the US (judging by their IP addresses).

[3]One participant who "tended to disagree" did so on the grounds that post-experimental questionnaires eliciting personality traits and the like are very important already (which (s)he indicated in a final open-comments field).

[4]It was made clear that the question referred to the payment procedures "as lab standards, so that participants can foresee the procedure".

60% do not expect that asking participants for their names will have any noticeable effect on answer consistency, while the remaining 38% are split roughly evenly between those expecting an increase and those expecting a decrease in answer consistency.

Turning to aspect (iii), 63% of regular and 73% of occasional users usually frame the fixed part in participants' compensation as being "for showing up to the study" or "for showing up in time" rather than as being "for completing the questionnaire". This corresponds well to the belief of 72% of regular and 75% of occasional users that this framing will not have an impact on data quality.

Finally, progress feedback is a ubiquitous feature of the surveys we face in our everyday lives—and one that is absent in many post-experimental questionnaires. In fact, 59% of regular users and 69% of occasional users never use any kind of progress feedback. This corresponds with respondents' expectations: Only 36% of regular and 35% of occasional users believe that displaying "page x/N" on every screen would increase data quality. Similarly, only 14%/16% believe that it helps to display "page x/N" after participants have completed two thirds of the questionnaire.

This scepticism is in line with results from the literature: From their meta-analysis of experiments on progress indicators for web surveys, Villar, Callegaro, and Yang (2013) conclude that progress indicators do help to reduce the drop-out rate provided that they indicate fast progress at the beginning (and slow progress at the end). However, "linear" progress indicators do not have an effect and slow-then-fast indicators are harmful. Liu and Wronski (2017) examine a huge number of real-world surveys and conclude that even "linear" progress bars are harmful. Note, however, that this literature is focused on survey-completion rates which are not an issue in the laboratory setting I focus on. In contrast, I look at the reliability of answers depending on the progress indicator.

To examine the five aspects, I measure answer quality following a similar approach as Meade and Craig (2012), combining a number of standard proxies for dishonest or careless answering from the literature into a single answer-quality index. The proxies check for the internal consistency of answer pairs (partially reverse-coded), count how frequently a participant picks an alternative that is rarely chosen by the average participant, or count the longest string of, for example, all-left or all-right item answers. The proxies are well-correlated, and they are correlated in the expected way with further measures of diligence, such as a patience and motivation measures. By simulating a random-error benchmark it is possible then to identify those participants whose answers should be considered invalid and therefore should be excluded from further analysis.

I find that waiting for all and then paying by cubicle numbers yields a higher answer quality than paying participants as soon as they finish or a middle course between the two, which is a specific variant of waiting for many before starting to pay. In this paper, the middle course means waiting until two thirds of the participants have finished, paying this group in reverse completion order, and paying the rest in completion order afterwards. Reversing the completion order for the fastest participants is meant to take away any incentives for answering very fast. However, waiting for all also is associated with substantial time costs (in the case of this study, 15 minutes compared to paying by completion order as soon as possible). In this perspective, the third payment procedure potentially is an attractive compromise in the speed-accuracy trade-off.

Having participants enter their name into the computer for preparing the receipts does not reduce answer quality, nor does it lead to more socially desirable answers.[5] A higher payment in the experiment generally increases answer quality, but paying some amount explicitly as a reward for filling out the questionnaire has no effect. There is no clear effect of the progress report, irrespective of whether the progress indicator is displayed right from the beginning or starting only after participants have completed two thirds of the questionnaire.[6]

2 The questionnaire

In this section, I present four standard scales for measuring careless answers that enter the combined answer-quality index. I then explain how I constructed the questionnaire in order to be able to measure answer quality in the sample. Using this questionnaire, I conduct different treatments to examine the aspects of questionnaire(-administration) design pointed out in the introduction.

2.1 The unreliability index

Following the example of Meade and Craig (2012), I consider four measures for careless or erroneous answers in the questionnaire: Self-reported unreliability (*e.g.*, Meade and Craig, 2012), the VRIN inconsistency index (including

[5]The missing effect on the measured social-desirability bias may be due to the (true) assertion that names were used exclusively for the preparation of receipts, in conjunction with the strong reputation of the laboratory in terms of non-deception.

[6]Note, however, that displaying the progress report after two thirds is associated with substantially higher levels of motivation as judged by the length of answers to open-ended questions towards the end of the questionnaire.

a number of reverse-coded pairs; Pinsoneault, 1998), the rarity index according to the O'Dell (1971) principle of rare answers (which includes also some 'bogus' items with a clear correct answer, Beach, 1989), and a straightlining index (Zhang and Conrad, 2013). Self-reported unreliability directly asks participants to point out unreliable answers, while the other three indices try to detect patterns in the answers which are likely to arise if a participant answers carelessly. For all indices of careless answers, a value of zero means full reliability while a high value means maximal carelessness.

The **self-reports index** is constructed from answers to the item "You can rely on my answers—Yes/In between/No." which I asked on 9 out of 14 screens as the last question.[7] The index is constructed as a binary variable. A value of 0 indicates that the participant chose "in between" at most once, otherwise self-stating full reliability. This criterion classifies 9% of the participants as unreliable. Accordingly, 91% have an index value of 0.[8]

The VRIN **inconsistency index** counts inconsistent answer combinations to pairs of questions. Essentially, these pairs ask the same question twice using different wording (some of them reverse-coded). If the two answers of a participant are not consistent with each other, the index rises by one point.[9] In total, the questionnaire has 10 pairs of questions taken from the original MMPI-2 and 15 additional own questions.[10] I designed the questionnaire such that each two companion questions are sufficiently far apart, and only in exceptional cases on the same screen. The index is built by counting the number of inconsistencies.

The basic idea of the **rarity index** is as follows: Subjects who choose rare answers more often are more likely to have answered randomly than others. O'Dell (1971) defines a rare answer as an answer that is selected by less than 10% of the total population. For the rarity index, I use 17 questions from the 16PF questionnaire, 13 of which were used already by O'Dell (1971),[11]

[7] I did not include the question on screens where I felt it would not make much sense, such as the introductory screen of the questionnaire or a screen essentially asking participants whether they would lie to the experimenter for their own benefit.

[8] The results are virtually identical if I code those with a single "in between" answer as having a self-reports index of 0.5.

[9] The original VRIN index (Pinsoneault, 1998) includes only rare answers which violate a 10%-rarity criterion. I chose to ignore this additional criterion for consistency with other studies such as, for example, Walczyk et al. (2009).

[10] I dropped some of the question from the original VRIN because they might raise suspicion amongst the participants, *e.g.*, "I suffer from stomach trouble several times a week," or "Somebody means me ill."

[11] In the questionnaire, I included all 31 items used by O'Dell, but only 13 of them met the 10%-criterion. A possible reason for this discrepancy might be that I had to use a different version of the 16PF questionnaire (from 1967), because the 1961-version used by

3 of the additional questions for the inconsistency index, and a hypothetical question about lying, reflecting the die-rolling task from Fischbacher and Foellmi-Heusi (2013; asked on a separate form). The rarity index is then given by the count of a participant's rare answers.

The **straightlining-index** (see, e.g., Zhang and Conrad, 2013) detects a specific type of visual pattern in the answers. For example, a participant trying to finish the questionnaire as effortlessly as possible may click the left-most answer option over a whole screen. In the questionnaire, there are four screens on which the answer items are sorted from left to right. For each of these screens, I count the longest sequence of subsequent answers which have the same position on the screen. The index is then calculated as the average of the longest strings on the four different screens.

To improve the individual indices' power in identifying careless or erroneous answering, I integrate the inconsistency index, the rarity index, and the straightlining-index into a single variable. I use two measures of carelessness: A continuous variable indicating how unreliable the answers of each participant are (*e.g.*, relative to those of other participants), and a binary variable indicating whether we can rely on a participant's answers or not.

For constructing the **continuous unreliability index**, I have to determine a weighting procedure, according to which the different measures enter into the index. As I have no prior that one index should have more weight than another one, I use an unweighted average over the normalized index values. I normalize the different indices by dividing the value of the index by its maximum value as obtained in the sample of participants. Thus, the continuous unreliability index is given by the average of the three normalized index values.[12]

For the **binary unreliability index**, I simulate a distribution of index values assuming that participants make random errors.[13] Next, I compute the continuous-index distribution for the simulated agents. I identify all those as "definitely careless" who have a value of the index that is larger than 95%

O'Dell was not available to me. Hence, the overlap may have been only partial. In addition to the above 31 questions, I included another 11 items from the 16PF questionnaire, mostly to use them for an extended inconsistency index. Four of these items yielded a "rare-answer distribution."

[12] I followed the suggestion of an anonymous referee not to include the self-reports index in the continuous index.

[13] For the simulation, I assume that an agent responds to a given question by an answer that is randomly drawn from that same question's distribution of answers by the whole population. For questions that enter the inconsistency index, I sample from the distribution on the second question conditional on the first question. To obtain reliable results, I simulate 100'000 agents.

of the values exhibited by the simulated random-error agents.[14]

As a robustness check for the findings I include further measures and sets of questions into the questionnaire. I track participants' completion time for each screen of the questionnaire (see, *e.g.*, Walczyk et al., 2009, for the use of response times to identify liars).[15] I measure participants' motivation by the average length of the free-text answers to four open questions towards the end of the questionnaire (measured by the number of characters including spaces). And I include the patience scale taken from Dudley (2003) used in Bruttel and Fischbacher (2013).

The order of the different parts of the questionnaire is as follows. First, participants have to enter their name (only in one treatment variation), then there is an introductory screen explaining the questionnaire and the importance of answering carefully, followed by 11 norm-conformity questions that serve to measure social-desirability bias. Then, the main questionnaire follows which I use to measure the different indices. This main part is followed by the patience questionnaire, four open-answer questions, and a short socio-economic questionnaire.

2.2 Treatments

The unreliability index can serve two purposes. Contrasting the carelessness-index values of given sets of answers against the expected distribution under random-errors, we can identify "definitely careless" responding *after* conducting a questionnaire study. Second, by comparing different treatments we can find out how to design and administer the questionnaire best in order to maximize answer quality *before* conducting a study. Table 1 summarizes the treatments I ran for this second (and main) purpose of this study.

The study contains data from two Experiments meant to address the five aspects pointed out in the introduction: (i) the role of the payment order (in particular, whether and how it is linked to the questionnaire-completion order); (ii) the effect of asking for participants' names to prepare the receipts; (iii) the framing of the fixed element in participants' payments as a compensation for showing up or for completing the questionnaire; (iv) the overall

[14] The 95% criterion is arbitrary, but definitely conservative. In my data set, the criterion singles out 4.6% of the participants. Depending on the context of the experiment, it may be appropriate to exclude more than those 4.6% of observations. My later analysis will provide some indications. For comparison, Meade and Craig (2012) identify 10-12% as careless in their Internet survey of psychology students, in addition to 12% participants not completing the online survey.

[15] For 24 participants I do not have information about completion times due to technical problems when conducting the experiment.

Treatment	Text version	Enter Name	Show Number	N. Obs.
byCubicle	fair	no	yes	96
byFinish	quick	no	yes	123
inBetween	both	no	yes	186
byCubicle.II (control)	fair	no	yes	74
enterName	fair	yes	yes	72
fixedPayCalledShowUpFee	fair[a]	no	yes	94
noNumber	fair	no	no	83
numberAfterTwoThirds	fair	no	after form 8	77
noJustification	'none'	no	yes	79

[a]not mentioning the "payment for questionnaire."

Table 1: Overview of the treatments; the three upper treatments constitute Experiment I, the remaining six constitute Experiment II.

payoff participants receive; and (v) the existence of progress feedback.

Experiment I comprised the first three rows of Table 1 and focused on aspect (i): varying the order in which participants received their payment. Experiment II then focused on aspects (ii)-(v), also including an important control treatment for Experiment I.[16]

In the byCubicle treatment of Experiment I, all participants had to wait until everybody had completed the questionnaire before receiving their payment in the order of their cubicle numbers. I randomly started the payment procedure either with the lowest or the highest cubicle number. This procedure is the common baseline procedure for both Experiment I and Experiment II (byCubicle and byCubicle.II differ only in the experiment preceding the questionnaire, see Section 4).

In the byFinish treatment, participants were called to the exit for payment as soon as they had completed the questionnaire (first-come–first-served). While byFinish potentially sets the unintended incentive to answer as fast as possible, byCubicle avoids setting this incentive at the cost of a longer total duration of the session.[17] Treatment inBetween tries to balance these two opposing goals using the following procedure: Payout starts when two thirds of the participants have completed the questionnaire. The last finisher out of this first group gets his or her payment first, then the or-

[16]The questionnaire was virtually identical between Experiment I and Experiment II, with one minor difference: in Experiment II, I left out the fifth screen of the original questionnaire that had been included for an unrelated study. On the screen, participants read: "please choose one of the following four boxes and click on it: A, B, A, A."

[17]I will use the recorded completion times to provide an estimate of the tradeoff's dimension.

der of payment is the reversed completion order within this group. After the first group of participants have received their payment, the remaining third of participants is called to the exit for payment in the order of completion.

In order to explain the INBETWEEN procedure to the participants, I introduced this treatment with a short justification: "In order to avoid both, unnecessary waiting time and time pressure in answering the questionnaire, [...]." For a clean treatment comparison, I also included justifications in BYFINISH ("to avoid unnecessary waiting time") and BYCUBICLE ("for fairness reasons"). In principle, these justifications add an additional potential confound: The justification for BYCUBICLE suggests that the experimenter cares about fairness. This might be reciprocated by "fairer", or in this case, more reliable answers by Levine-(1998)-type altruists who care more about fairness-minded others than about egoists. To control for the potential confound, I included treatment NOJUSTIFICATION in Experiment II (final row in Table 1). In this control treatment, I left out the justification for the payment procedure altogether. If alluding to fairness leads to more reliable answers, NOJUSTIFICATION should produce less reliable answers than the replication of BYCUBICLE (*i.e.*, than BYCUBICLE.II).

The main purpose of Experiment II, however, was to examine aspects (ii)-(v) of the study. In Experiment II, I ran all treatments simultaneously within the same sessions (including BYCUBICLE.II and NOJUSTIFICATION), allocating participants to treatments randomly.

In ENTERNAME, I asked participants to type in their name for the experimenter to prepare their receipts (aspect ii). This happened on the first page of the questionnaire, which was followed by the general instruction page and the social-desirability scale.

I also tested whether framing the usual show-up fee as a payment for questionnaire completion evokes (additional) reciprocity from the participants (aspect iii). For this purpose, I used such a framing in all the treatments but FIXEDPAYCALLEDSHOWUPFEE. In FIXEDPAYCALLEDSHOWUPFEE, I framed the fixed payment part as a show-up fee, mentioning it only at the beginning of the session but not again before the questionnaire. Aspect (iv)—the effect of experimental earnings on answer reliability—will be analysed directly, so that I do not need an additional treatment without a show-up/questionnaire-completion fee.

Finally, I included two treatments to examine aspect (v), the effect of progress feedback. In NONUMBER, participants did not see which questionnaire form out of how many questionnaire forms they were filling in at any point of the questionnaire. In NUMBERAFTERTWOTHIRDS, participants saw a progress report (form X out of 14) on the top of the screen after they had completed nine of the forms. Using NUMBERAFTERTWOTHIRDS, I set out

to test whether seeing the end come closer would have a motivating effect on the participants.

3 Hypotheses

Before I can use either of the indices in treatment comparisons, I have to make sure they measure what they are intended to measure. For this purpose, I will first relate each of the partial indices to the self-stated unreliability measure. The hypothesis is that some of those who answer carelessly will also admit doing so, even if only in a downplayed way.

Hypothesis 1. The unreliability index, the inconsistency index, the rarity index, and the string-index correlate positively with each other.

Once I have established that the indices measure what they are supposed to measure, I can use them to examine whether the treatment conditions affect the reliability of participants' answers. The first hypothesis comes directly from the different incentives between byFinish and byCubicle (assuming time is a good for the participants).

Hypothesis 2. Participants paid in questionnaire-completion order (byFinish) have lower answer quality than those in byCubicle.

Conjecture: There is no difference between inBetween and byCubicle.

In enterName, participants have to type in their name "for the experimenter to prepare the receipts" (which I did, and which is the usual reason for letting participants type in their names). After doing so, participants saw the introductory page of the questionnaire, followed by the social-desirability questions. I expect a negative effect of enterName on answer quality: Participants may be more reluctant to give honest answers when they fear that the experimenters may be able to connect these answers to their personal data.

Hypothesis 3. Participants entering their name before filling out the questionnaire have lower answer quality.

Paying participants generously may improve answer quality if they have reciprocal preferences. This effect may unfold in two dimensions. First, having earned more money in the experiment may make participants spend more effort. Second, reminding them that part of their payment is explicitly intended to serve as reward for filling out the questionnaire may trigger reciprocal behavior.

Hypothesis 4. (i) Participants who get a higher payment have higher answer quality. (ii) Participants in FIXEDPAYCALLEDSHOWUPFEE have lower answer quality.

Regarding the progress report on screen, there may be two counteracting effects (*cf.*, *e.g.*, Villar, Callegaro, and Yang, 2013, for a discussion of the "first-impression" and the "surfacing" hypotheses). On the one hand, participants seeing that there will be 13 screens in total might be discouraged. Hence, not providing a number might motivate participants at the start. On the other hand, the longer into the questionnaire, the more discouraging it might be not to know the end—and the more encouraging it might be to see how much has been accomplished already. The treatment NUMBERAFTERTWOTHIRDS tries to strike a balance between the two, not discouraging participants early on, and encouraging them towards the end.

Hypothesis 5. Participants who get progress feedback after completing two thirds of all pages have a higher answer quality compared to participants who get progress feedback from the start or no progress feedback at all.

4 Procedures

A total of 884 students from various disciplines took part in the study. I appended the questionnaire to different preceding experiments, holding the preceding experiment constant for Experiment I where I could not conduct the different treatments within the same sessions. In the experiment preceding Experiment I, one of three participants could steal 5 points from another. With some probability, stealing would be revealed, in which case 10 points from the stealing player's account would be transferred to one of the other two players. The experiment was conducted as a one-shot game. For Experiment II, the preceding experiments centered on tasks in which participants saw four boxes with non-neutral labels (such as A-B-A-A or Ace-2-3-Joker). They had to perform these tasks repeatedly (on at least 15 different frames). Examples for these tasks include a discoordination game or a lottery task. After the tasks, participants played a standard trust game in full strategy method. Both preceding experiments took about 40 minutes on average. Experiment I took place from January to May 2012, Experiment II between November 2015 and November 2016.[18]

[18]Note that the questionnaire required the preceding experiments to be short, I did not want the same participants to fill it in twice, and I needed a large number of participants. These restrictions made the collection of the data somewhat harder than for the usual experiment.

The average time for filling out the questionnaire was 17 minutes, the maximum time was 42 minutes. Subjects received 5 Euros for filling out the questionnaire. At the end of each session, participants were called to the exit individually. They received their cash payment privately to maintain anonymity.

5 Results

5.1 Constructing the Unreliability Indices

The self-reports index is 0 for 91% of the participants and 1 for the remaining 9%. Figure 1 illustrates the distributions of the other indices. As you can can see from Figure 1, all three remaining indices show sufficient variance for a meaningful analysis.

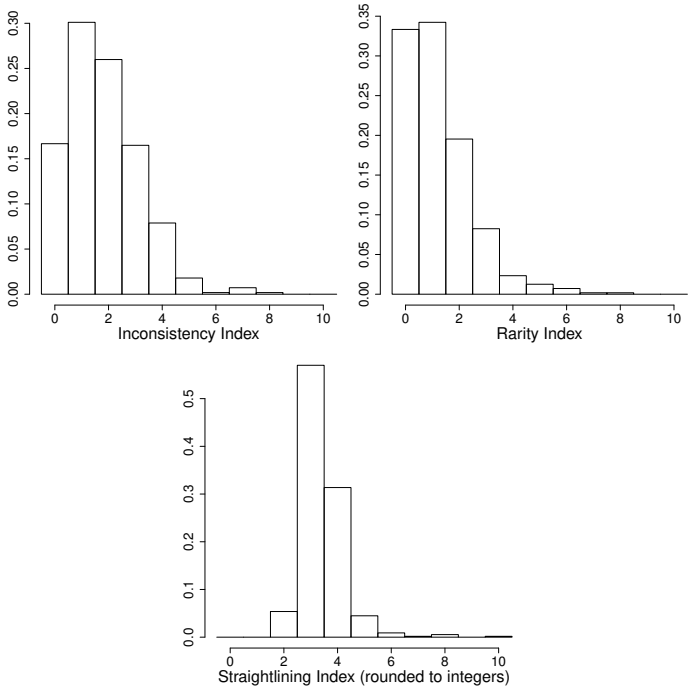

Figure 1: Distributions of index values.

Result 1. The self-reports index, the inconsistency index, the rarity index, and the string-index correlate positively with each other. In turn, the continuous unreliability index combining the inconsistency, rarity, and straightlining-indices correlates negatively with a patience index and with a

measure of motivation. Furthermore, the continuous unreliability index is negatively correlated with questionnaire-completion time (only) among the fastest quartile of the population, and negatively correlated with belief-action consistency in a preceding experiment.

Table 2 shows the correlations of the different indices for unreliable answers. All indices are significantly positively correlated with each other. The continuous unreliability index is correlated more strongly with the self-reports index ($\rho = 0.36, p < 0.001$) than any of its sub-indices.

	Self-reports	Inconsistency	Rarity	Straightlining
Self-reports	1.00			
Inconsistency	0.22***	1.00		
Rarity	0.25***	0.24***	1.00	
Straightlining	0.32***	0.18***	0.08*	1.00

Note: *** denotes significance at the 1‰ level, ** at the 1% level, * at the 5% level.

Table 2: Correlations between individuals' values on the different scales.

Patience as measured in the patience questionnaire correlates negatively with the continuous unreliability index ($\rho = -0.12, p < 0.001$). The same is true for participants who seem to be more motivated, judging by how much they write in the open-answer questions ($\rho = -0.13, p < 0.001$). The correlation between the time needed to fill in the questionnaire and unreliability for the fastest quartile of the population is substantial and negative ($\rho = -0.40, p < 0.001$), while the correlation is small (and positive) for the rest ($\rho = 0.08, p = 0.039$).

Finally, participants' carelessness-index value is related to their behaviour in the preceeding experiment.[19] The idea is that if participants answer carelessly in the questionnaire, they may have done so already in the preceding experiment. A reasonable measure for carelessness in the experiment is the degree of consistency in participants' behaviour. A type of behavioural consistency that has been discussed prominently in the literature is belief-action consistency. Fortunately, for 67 of the participants in Experiment II, I elicited both actions and beliefs.[20] If I relate the participants' carelessness-index value to their average belief-action-consistency rate over the 24 rounds they played, I find a clear and substantial negative correlation

[19] I am grateful to Marie Claire Villeval for inspiring this analysis.

[20] Of course, this was done in an incentive-compatible way: participants knew they would not be paid for their action and their belief in the same decision situation, and beliefs were incentivized by a binarised scoring rule.

($\rho = -0.297, p = 0.015$). In other words, some of the participants seem to pay only insufficient attention to the experimental tasks in general.

Summing up, all indicators are there that the continuous index provides a useful measure for participants' degree of careless answering, in line with prior uses of analogous indices in the literature.

Figure 2 illustrates the distribution of the final (continuous) unreliability index, plotted against the density function of the 100'000 simulated random-error agents in red. The dotted line indicates the 95% quantile of the random-error agents. To obtain a binary classification into reliable and unreliable answers, every participant with an unreliability-index value above this 95% quantile is classified as "definitely careless." Given that the distribution of the random agents is—and should be—shifted towards higher unreliability-index values, the 95% criterion is a rather conservative measure.

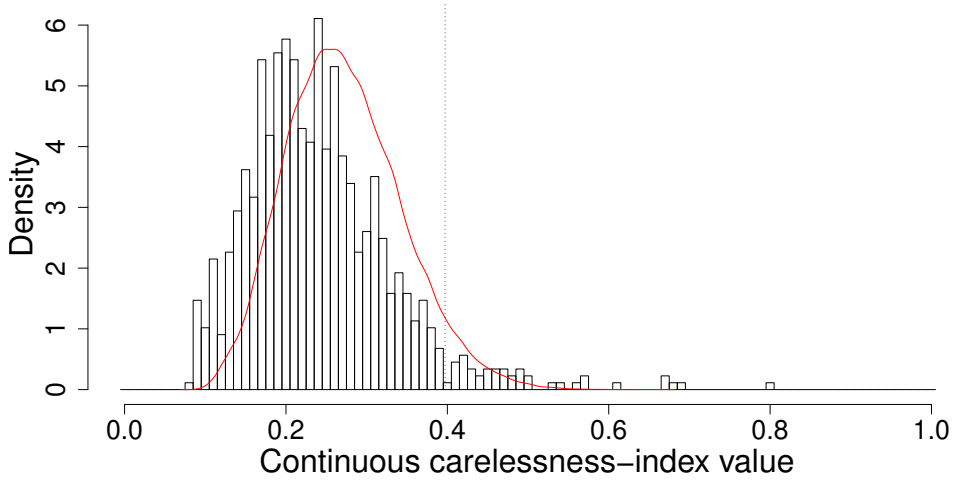

Figure 2: Frequency distribution of the continuous unreliability index, together with the density function of the 100'000 simulated random-error agents (red solid line) and its 95% quantile (dotted line).

5.2 Treatment comparisons

In this section, I focus on the question of which payment order researchers should choose when administering post-experimental questionnaires. Again, I present both continuous and binary index values for all the treatments. Table 3 summarizes their average values across treatments.

Table 4 reports the results of regression analyses testing for differences in answer quality across treatments. In the left-most column of Table 4, I regress the continuous unreliability index on the treatments, using standard

Treatment	Continuous unreliability index	"Definitely careless" (in %)	Unreliable based on on self-statements (in %)
byCubicle	0.22	3.1	5.2
byFinish	0.25	6.5	14.6
inBetween	0.23	3.8	9.7
byCubicle.II (control)	0.26	5.4	5.8
enterName	0.23	1.4	5.6
fixedPayCalledShowUpFee	0.25	6.4	3.2
noNumber	0.25	4.8	10.8
numberAfterTwoThirds	0.25	6.5	2.6
noJustification	0.23	3.8	2.5

Table 3: Average values of the continuous unreliability index, and percentages of participants characterised as "definitely careless" based on the binary unreliability index (middle) and self-stated unreliability (right), by treatment.

ordinary-least-squares regressions. The base category is byCubicle. I control for patience (as measured on the patience scale), motivation (as measured by the amount written in the free-form questions), and total completion time, as I expect them to be related to answer quality. I use the motivation and time measurements relative to the respective treatment average to account for the fact that the treatment conditions will also affect these two measures and that I am interested in the total treatment effects. In addition, I control for the total earnings from the experiment, participants' value on the norm-conformism scale, their gender, and whether they study economics. In the second column of Table 4, I regress the binary unreliability index on the same variables, using a probit regression and reporting the average marginal effects. Columns 3-6 then report ordinary-least-squares regressions of the individual sub-indices on the same regressors.

Table 3 suggests that byFinish produces the most unreliable answer sets, supporting Hypothesis 2. In Table 4, we see that byFinish indeed produces the lowest answer quality and more participants self-report giving unreliable answers. The effect on the continuous index (left-most column) is non-negligible (1/3 of a standard deviation) and seems to be driven mostly by people giving inconsistent answers. At the same time, the coefficient for inBetween is not significant. An F-test for equality of the coefficients for inBetween and byFinish is rejected ($p = 0.049$), and if I change the base category to inBetween, the coefficient for byFinish remains significant. Hence, I can conclude that the payment order byFinish induces more careless answering than either byCubicle or inBetween. If there is any difference between byCubicle and inBetween, it is too subtle to manifest

	Continuous unreliability index (OLS; std. dev.: 0.091)	Binary unreliability index (probit, av. marg. effects; std. dev.: 0.210)	Rarity (OLS; std. dev.: 1.261)	Inconsistency (OLS; std. dev.: 1.404)	Strings (OLS; std. dev.: 0.767)	Self-repor unreliability std. dev.: 1
(Intercept)	0.228 (0.011)***		1.207 (0.156)***	1.415 (0.173)***	3.650 (0.093)***	0.091 (0.03
byFinish	0.032 (0.012)**	0.034 (0.038)	0.298 (0.171)	0.601 (0.189)**	−0.066 (0.102)	0.098 (0.03
inBetween	0.012 (0.011)	0.001 (0.028)	0.132 (0.158)	0.317 (0.175)	−0.163 (0.094)	0.050 (0.03
EnterName	−0.027 (0.016)	−0.049 (0.008)***	−0.269 (0.229)	−0.382 (0.253)	−0.040 (0.137)	−0.015 (0.04
Total Earnings (in Euros)	−0.001 (0.001)*	−0.001 (0.001)	−0.016 (0.008)*	−0.014 (0.009)	−0.004 (0.005)	−0.003 (0.00
fixedPayCalledShowUpFee	−0.014 (0.015)	0.005 (0.035)	−0.040 (0.211)	−0.243 (0.234)	−0.103 (0.126)	−0.027 (0.04
noNumber	−0.003 (0.015)	−0.010 (0.031)	0.014 (0.215)	−0.058 (0.238)	−0.036 (0.128)	0.056 (0.04
numberAfterTwoThirds	−0.009 (0.016)	0.025 (0.044)	−0.201 (0.221)	0.041 (0.244)	−0.058 (0.132)	−0.039 (0.04
NoJustification	−0.013 (0.016)	−0.007 (0.032)	−0.255 (0.222)	−0.043 (0.246)	−0.037 (0.132)	−0.036 (0.04
Experiment II	0.042 (0.015)**	0.018 (0.037)	0.532 (0.211)*	0.677 (0.234)**	−0.162 (0.126)	0.036 (0.04
Total time†	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	−0.000 (0.000)**	−0.000 (0.00
Motivation†	−0.000 (0.000)***	−0.001 (0.000)**	−0.003 (0.001)*	−0.004 (0.001)**	−0.002 (0.001)*	−0.001 (0.00
Norm Conformism	−0.000 (0.001)	0.000 (0.002)	0.000 (0.013)	−0.000 (0.014)	−0.014 (0.008)	0.003 (0.00
Patience	−0.002 (0.001)**	−0.001 (0.001)	−0.011 (0.008)	−0.022 (0.009)*	−0.014 (0.005)**	−0.003 (0.00
Female	0.006 (0.007)	0.006 (0.016)	−0.043 (0.095)	0.116 (0.105)	0.112 (0.057)*	−0.050 (0.02
Economist	0.002 (0.007)	0.017 (0.017)	−0.103 (0.101)	0.192 (0.112)	−0.027 (0.060)	0.015 (0.02
R^2	0.053		0.028	0.047	0.065	0.049
Adj. R^2	0.035		0.010	0.029	0.047	0.031
Num. obs.	806	806	806	806	806	806
RMSE	0.090		1.254	1.390	0.750	0.266
Log Likelihood		-138.280				
Deviance		276.561				
AIC		308.561				
BIC		383.634				

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$; †relative to the treatment average, to control for treatment effects on the measures.

Table 4: Regressing the unreliability indices on treatment conditions and further controls (note: all significant coefficients remain significant if we take out all individual controls). I include the analogous regression of self-reported unreliability for completeness.

itself in the data clearly. Note also that the effect is not driven by the fact that I alluded to fairness in justifying the byCubicle treatment to participants. Looking at the coefficients for the noJustification treatment, we see that they are all negative. Not mentioning "fairness" thus seems to reduce answer unreliability, if at all.

Result 2. Participants paid in questionnaire-completion order (byFinish) have lower answer quality than those in byCubicle and inBetween.

Up to this point, it seems that I unambiguously should recommend researchers to use the byCubicle procedure when administering post-experimental questionnaires. However, this payment order also causes substantial time costs via two channels. First, the payment procedure delays the time when payments start. While payment in byFinish is normally completed straight after the last participant finished the questionnaire, payment in byCubicle only starts at that point. This already causes a time difference of about 10 minutes, during which participants sit around waiting. Second, the procedure in byFinish provides incentives to fill in the questionnaire faster than the procedure in byCubicle. Comparing completion times by treatment, I find

that the last participant in a byCubicle-session on average needs about 4.5 minutes longer than the last participant in a byFinish-session (1491 seconds compared to 1227 seconds).[21]

All in all, the improvement in answer quality in byCubicle comes at the cost of about 15 additional minutes duration. If this is considered crucial, inBetween may be a viable compromise between reliable answers and completion speed. It shortens the payment procedure by more than five minutes compared to byCubicle and causes only insignificantly more careless answers.

Some experimenters ask participants to enter their name in the beginning of the questionnaire in order to print automated receipts. According to the regression analyses in Table 4, this practice is not harmful for answer quality. If at all, enterName leads to less participants being classified as 'definitely careless' (effect size: 0.23 of a std. dev.).[22] In contrast to what we might have expected, this effect does not seem to be driven exclusively by a reduced level of inconsistency, as the effect on rare answers seems to be of roughly comparable size (21 *vs.* 27% of a std. dev.).

Result 3. Participants entering their name before filling out the questionnaire do not have lower answer quality.

The higher the total earnings of a participant are, the better is answer quality according to the data (paying 60 rather than 5 Euros is associated with an estimated decrease in the continuous unreliability index of around 80% of a std. dev.). However, framing the showup fee as a payment for answering the questionnaire does not have an impact on the unreliability index. Thus, it seems that it is not a reciprocity motive that helps improving answer quality, but rather satisfaction with the experiment in general and their payment in particular that supports participants' motivation to fill out the questionnaire carefully.[23]

Result 4. (i) Participants who get a higher payment have higher answer

[21] Note that this comparison rests on a simulation of 1'000 hypothetic sessions that controls for varying session sizes. The estimate provided is for a session size of 24 participants.

[22] Interestingly, enterName does not even have a significant effect on the social desirability of answers to the norm-conformism scale ($p = 0.666$ for the coefficient in a regression of norm conformism on treatments and total earnings). This would suggest that the laboratory's strict no-deception policy pays off in that participants trust the announcement that I will not store their names together with their questionnaire responses.

[23] This corresponds to the finding that answer quality was lower in Experiment II. In this study, the preceding task was rather repetitive, so their overall satisfaction was presumably lower, which in turn may have reduced their willingness to invest effort into the questionnaire.

quality. (ii) Framing the showup fee as a reward for filling out the questionnaire does not improve answer quality further.

The progress report has no clear effect on answer quality, neither if it is shown throughout the questionnaire nor if it is provided only in the last part of the questionnaire. This is despite the fact that participants in NUMBERAFTERTWOTHIRDS show clearly the highest values on the motivation scale (in a regression of motivation on treatments, it is the only treatment that has a significantly positive coefficient at the 5-percent level, with an average of 17 characters above the BYCUBICLE level).

Result 5. Participants who get progress feedback do not have a higher answer quality than those who do not.

There are two more factors that influence the answer quality in the data within each treatment: Participants' value on the patience scale, and their motivation. Both of them point into a plausible direction: being more patient or motivated leads to more careful answering (going from the least patient/motivated to the most patient/motivated participant decreases the continuous unreliability index by 73%/120% of a std. dev.).

6 Discussion and conclusions

This paper is about the reliability of answers to post-experimental questionnaires in laboratory experiments. In the typical experiment, we may want to identify unreliable answers *ex post*, while methodological studies like mine test procedures to prevent them *ex ante*. I design a number of treatments to test for their effects on answer reliability, assessing the reliability of answers by following the approach of papers like Meade and Craig (2012). In doing so, I focus on five aspects that are rather specific to economic experiments, inspired by the heterogeneity in procedures used by different researchers and laboratories. Running a small questionnaire among experimental economists with 94 respondents, I provide a rough idea of the extent of this heterogeneity.

First, among those researchers who regularly or occasionally use post-experimental questionnaires to elicit more than the standard socio-economic demographics, roughly one in seven always pays participants as soon as they complete the questionnaire, and more than a quarter use this procedure regularly. This study shows that paying as soon as possible clearly leads to less reliable answers than waiting for all or using an intermediate procedure. Surprisingly, around 40% of the above researchers would not have expected any difference in data quality.

Second, one in seven researchers asks participants to type in their names in order to prepare the receipts for payout. Judging by the survey responses, the general view in the profession seems to be that doing so will not affect data quality. While my results are not conclusive on this question, they do suggest that asking participants may actually improve data quality, as judged by the lower number of participants classified as 'definitely careless'. Contrary to what we might have expected, I do not find any effect on a social-desirability scale, and the effect seems to be driven by more than just an increase in answer consistency.

Third, paying participants more increases data quality. However, it does not seem to matter what participants are being paid for (as would be the case if reciprocity followed mental-accounting principles; aspect four). This is well in line with the expectations of 75% of the surveyed researchers that framing the fixed part of participants' payments as being "for completing the quesitonnaire" rather than "for showing up" will not affect data quality. Having said that, roughly a third of the respondents sometimes or always do frame at least a part of the fixed part of participants' payment as being "for completing the questionnaire".

Finally, roughly a third of the questionnaire users expect that the type of progress report used in this study will improve data quality. In fact, providing progress feedback after two thirds of the questionnaire forms does increase measured motivation at the end of the questionnaire. However, I find no evidence for a positive effect on overall data-quality for either continuous or delayed progress feedback. This corresponds well to findings from the survey literature, where linear (*i.e.*, truthful) progress feedback does not affect survey-completion rates (*e.g.*, Villar, Callegaro, and Yang, 2013).

The unexpected effect of the preceding experiment suggests that participants' experience from that experiment influences answer quality: Quality may deteriorate when the experiment is too repetitive. At the same time, the measured answer quality also conveys new insights into the behaviour in the preceding experiment. The literature has long documented notable inconsistencies in participant behaviour (*e.g.*, Tversky, 1969; Nyarko and Schotter, 2002). Among the participants for whom I have a measure of behavioural consistency in the experiment, unreliability in the questionnaire and consistency in the preceding experiment are strongly negatively correlated. This suggests that some participants pay insufficient attention to the experimental tasks in general. I therefore contribute to explaining the puzzle of why participants so often act inconsistently in economic experiments: Some of them simply seem to care too little, or are unable to focus their attention for long enough.

A final word on the size of the reported effects seems in place. While

most of the effects are not huge, most of them are non-neglibigle.[24] We also have to keep in mind that *precise* measures of unreliability do not exist, so that I had to rely on noisy proxies. In that perspective, being able to detect treatment differences is an achievement in itself. The documented effects might well under-estimate the true differences in how much care participants take when completing post-experimental questionnaires.

Technical acknowledgements

All experiments were computerized using z-Tree (Fischbacher, 2007), participants were recruited using ORSEE (Greiner, 2015) with Mozilla Firefox. The statistical analyses were done using R (R Development Core Team 2001, 2012; Ihaka 1998) in combination with RKWard (Rödiger et al., 2012) and Stata 13. Some of this was done on a computer running on KDE-based (KDE eV, 2012) Kubuntu, which required the use of wine for the programming of the experiments. The article was written using Kile and TeXnicCenter.

References

Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology: Interdisciplinary and Applied*, 123, 101–103.

Becker, A., T. Deckers, T. Dohmen, A. Falk, and F. Kosse (2012). The Relationship Between Economic Preferences and Psychological Personality Measures. *Annual Review of Economics* 4, 453–478.

Bruttel, L., and U. Fischbacher (2013). Taking the initiative. What characterizes leaders? *European Economic Review* 64, 147–168.

Dudley, K.C. (2003). Empirical Development of a Scale of Patience, Dissertation, West Virginia University.

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics* 10(2), 171–178.

Fischbacher, U., and F. Foellmi-Heusi (2013). Lies in disguise. An experimental study on cheating. *Journal of the European Economic Association* 11(3), 525–547.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association* 1(1), 114–125.

[24] Only the treatment-controlled motivation measure moves the expected data-quality index by more than a standard deviation when going from the least to the most motivated individual.

Kocher, M.G., D. Schindler, S.T. Trautmann, and Y. Xu (2019). Risk, Time Pressure, and Selection Effects. *Experimental Economics* 22(1), 216–246.

Levine, D.K. (1998). Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics* 1, 593–622.

Liu, M., and L. Wronski (2017). Examining Completion Rates in Web Surveys via Over 25,000 Real-World Surveys. *Social Science Computer Review* 36(1), 116–124.

Meade, A.W., and S.B. Craig (2012). Identifying Careless Responses in Survey Data. *Psychological Methods* 17(3), 437–455.

Nyarko, Y., and A. Schotter (2002). An Experimental Study of Belief Learning Using Real Beliefs. *Econometrica* 70, 971–1005.

O'Dell, J. (1971). Method for detecting random answers on personality questionnaires. *Journal of Applied Psychology* 55(4), 380–383.

Pinsoneault, T.B. (1998). A variable response inconsistency scale and a true response inconsistency scale for the jesness inventory. *Psychological Assessment* 10(1), 21–32.

Proto, E., A. Rustichini, and A. Sofianos (2019). Intelligence, Personality, and Gains from Cooperation in Repeated Interactions. *Journal of Political Economy* 127(3), 1351–1390.

Singer, E., and C. Ye (2013). The use and effects of incentives in surveys. *The Annals of the American Academy of Political and Social Science* 645(1), 112–141.

Tversky, A. (1969). Intransitivity of Preferences. *Psychological Review* 76(1), 31–48.

Villar, A., M. Callegaro, and Y. Yang (2013). Where Am I? A Meta-Analysis of Experiments on the Effects of Progress Indicators for Web Surveys. *Social Science Computer Review* 31(6), 744–762.

Walczyk, J.J., K.T. Mahoney, D. Doverspike, and D.A. Griffith-Ross (2009). Cognitive lie detection: response time and consistency of answers as cues to deception. *Journal of Business and Psychology* 24, 33–49.

Zhang, C., and F.G. Conrad (2013). Speeding in web surveys: the tendency to answer very fast and its association with straightlining. *Survey Research Methods* 8(2), 127–135.