No. 110

THURGAU INSTITUTE
OF ECONOMICS
at the University of Konstanz

Lisa Bruttel
Irenaeus Wolff

Incentives and Random Answers in Post-Experimental Questionnaires

Research Paper Series
Thurgau Institute of Economics
and Department of Economics
at the University of Konstanz

Member of

thurgauwissenschaft

www.thurgau-wissenschaft.ch

Incentives and Random Answers in Post-Experimental Questionnaires[§]

Lisa Bruttel[a] and Irenaeus Wolff[b]

[a] *University of Potsdam, Department of Economics and Social Sciences, August-Bebel-Str. 89, 14482 Potsdam, Germany, lisa.bruttel@uni-potsdam.de*

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

Abstract:

Post-experimental questionnaires are becoming increasingly important in experimental research in economics. Yet, the question of how these questionnaires should be administered remains largely up to the individual researcher. In this paper, we focus on two aspects of the design and evaluation of questionnaire data. First, we present a procedure that can help researchers to identify careless answers *ex post*. Second, we provide recommendations on how a questionnaire should be set up *ex ante* in order to maximize answer quality.

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

1 Introduction

Post-experimental questionnaires are becoming increasingly important in experimental research in economics to understand the decisions of participants and explain them. Yet, the question of how we should administer these questionnaires remains largely up to the individual researcher. In this project, we design a questionnaire that allows to construct an answer-quality index combining several existing measures for detecting dishonest or careless answers. These measures, for example, check for the internal consistency of answer pairs, count how frequently a participant picks an alternative that is rarely chosen by the average participant, or count the longest string of all-left or all-right item answers.

[§]We 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, as well as Marie Claire Villeval, Dirk Sliwka, and Roberto Weber, and the participants of the 2014 GfeW meeting for fruitful discussions. Konstantin Eitel provided valuable research support. His Master's thesis was part of this project.

The data we obtain from running the questionnaire in the laboratory supports the hypothesis that our index indeed measures answer quality. On top, by simulating a random-error benchmark we identify those participants whose answers should be considered invalid and therefore should be excluded from further analysis.

In a second step, we use our answer-quality index to examine experimentally how to administer a questionnaire in economic experiments optimally. In particular, we study which is the best rule for the order in which participants are paid, considering three potential procedures: paying participants by their cubicle numbers after all participants have completed the questionnaire; paying participants as soon as possible in the order of questionnaire completion; or a middle course between the two. The middle course means waiting until two thirds of the participants have finished, paying these two thirds in reverse completion order, and paying the rest in completion order afterwards. Furthermore, we study whether telling participants to enter their name into the computer reduces answer quality, and whether monetary incentives as well as a progress report on the computer screen improve it.

We find that paying by cubicle numbers yields the highest answer quality. However, this procedure also is associated with substantial time costs (in our case, 15 minutes compared to paying by completion order as soon as possible). Judging by our data, our third payment procedure is an attractive compromise in the speed-accuracy trade-off. Entering the name into the computer does not reduce answer quality, nor does it lead to a social-desirability bias. 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. The effect of the progress report seems to be weak.

There is a vast literature on survey methodology in different social sciences (see, for example, the broad overview in Singer and Ye, 2013). However, up until recently, there were only few studies that are directly relevant for this paper, and most of them focus on one single measure of carelessness. For example, O'Dell (1971) tries to detect random answers by identifying participants who frequently give answers which are given rarely in the remaining population; Pinsoneault (1998) developed an inconsistency index he called VRIN (Variable Response Inconsistency). Walczyk et al. (2009) use response times and inconsistent answers to identify liars, and Zhang and Conrad (2013) show a correlation between speeding and straightlining. In our study, we combine all of these measures to a powerful tool for the identification and analysis of careless answers in questionnaires conducted in a laboratory setting. A similar approach was pioneered in Meade and Craig (2012), who look at the reliability of Internet surveys filled in by students of psychology under obligatory participation.

Within experimental economics, there is little past research on exactly our

topic, but there are a number of related studies. Schmelz and Ziegelmeyer (2015) and Arechar, Molleman, and Gächter (forthcoming), for example, evaluate the reliability of data obtained in online experiments by the degree to which it replicates the findings from the same studies conducted in the laboratory. Bigoni and Dragone (2012) study whether using on-screen instructions with forced inputs improve comprehension, and hence, data quality. Furthermore, there is some research on how to elicit beliefs best in order to obtain reliable answers (e.g., Trautmann and van de Kuilen, 2015).

2 The questionnaire

In this section, we present the different existing scales for measuring careless answers, which we will use to construct our combined index. We then explain how we constructed the questionnaire in order to be able to measure answer quality in our sample. Having constructed the questionnaire, we conduct a number of experimental treatments. These treatments vary the order in which participants receive their payment for participation in the experiment, whether participants have to enter their name, whether the showup fee is framed as a special payment for the questionnaire, and two variants of a progress report on the computer screen. The treatments test how the procedure affects the value of the answer-quality index and the likelihood of this index value passing a critical value.

2.1 Constructing the carelessness index

In total, we consider four measures for careless or erroneous answers in the questionnaire: a self-reported unreliability measure, the VRIN inconsistency index, the rarity index according to the O'Dell principle of rare answers, and a straightlining index.

The **unreliability index** is constructed from answers to the item "You can rely on my answers." which we asked on 9 out of 14 screens as the last question.[1] This question offered the three possible answers "yes," "in between," and "no." The unreliability index is constructed as a binary variable. A value of 0 indicates that the participant self-stated full reliability whenever asked this question. In the ensuing analysis of our data, we will allow for one "in between" answer for an index value of 0. This more conservative criterion classifies 9% of the participants as unreliable. Accordingly, 91% have an index value of 0.[2]

[1] We did not include the question on screens where we 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.

[2] The results are virtually identical if we code those with a single "in between" answer as

The unreliability index may not perfectly filter out careless answers. To improve index quality, we will use the other three measures. These other measures are more indirect, because they 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 VRIN **inconsistency index** counts inconsistent answer combinations to pairs of questions. Essentially, these pairs ask the same question two times using two different wordings. If the two answers of a participant are not consistent with each other, the index rises by one point.[3] In total, we have 10 pairs of questions taken from the original MMPI-2 and 15 additional own questions.[4] For the 15 additional questions, we combine questions, for example, for a point estimate of some number (*e.g.*, the appropriate price for a ticket to the cinema) with a question asking whether some value is lower or higher than the appropriate value ("8 Euros are definitely too much for a ticket"). All pairs of questions are distributed over the whole 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 **rarity index** builds on O'Dell's principle of rare answers. The basic idea 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, we use 17 questions from the 16PF questionnaire, 13 of which were used already by O'Dell (1971),[5] 3 of the estimation questions which were included 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). Similar to the inconsistency index, the rarity index is constructed by counting the number of rare answers given by each participant.

The **straightlining-index** (see, e.g., Zhang and Conrad, 2013) detects a spe-

having an unreliability index of 0.5.

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

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

[5]In the questionnaire, we 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 we had to use a different version of the 16PF questionnaire (from 1967), because the 1961-version used by O'Dell was not available to us. Hence, the overlap may have been only partial. In addition to the above 31 questions, we 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."

cific type of visual pattern in the answers. For example, a participant trying to finish the questionnaire as effortless as possible may click the left-most answer option over a whole screen. In our questionnaire, there are four screens in which the answer items are sorted from left to right. For each of these screens, we 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 test whether each of the indices indeed measures what we are interested in, we use their correlation as a quality check. As we will see in Section 5.1, all the indices are correlated with each other. To improve their power in identifying careless or erroneous answering, we next integrate them into a single variable. We use two measures of carelessness: a continuous variable indicating how unreliable the answers of one 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 index of answer quality, we have to determine a weighting procedure, according to which the different measures enter into the index. As we have no prior that one index should have more weight than another one, we use an unweighted average over the normalized index values. We normalize the different indices such that the maximal value obtained in our sample is 1. This is done by dividing the value of the index by its maximum value as obtained in our sample of participants. For the unreliability index, no such normalization is necessary, because by construction it only takes the values 0 and 1. The average of the four normalized index values delivers our final (continuous) index value.

Using the continuous index, we will compare our different treatments to derive recommendations about how to administer questionnaires best from an *ex-ante* perspective. Another possible usage of the index is the identification of careless answers in order to exclude them from the analysis in an experimental dataset. For this second purpose, a binary index is more suitable. To construct such a binary index, we simulate a distribution of index values assuming that participants make random errors.[6] Next, we compute the preliminary index distribution for the simulated agents. We identify all those as "definitely careless" who have a value of the continuous index that is larger than 95% of the values exhibited by the simulated random-error agents.

The 95% criterion is arbitrary, but definitely conservative. Depending on the

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

context of the experiment, it may be appropriate to exclude more than 5% observations. Our later analysis will provide some indications.[7]

As a robustness check for our findings we include further measures and sets of questions into the questionnaire. We track participants' completion time for each screen of the questionnaire.[8] We estimate 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 we included a patience scale taken from the dissertation by Dudley (2003) as 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 of the questionnaire), then there is an introductory screen explaining the questionnaire and the importance of answering carefully, followed by a self-developed norm-compliance scale and the main questionnaire which we use for the construction of the different indices. Finally, there is the patience questionnaire, four open-answer questions, and a short socio-economic questionnaire.

2.2 Treatments

Our study tries to answer two related research questions. With the carelessness index introduced above, we apply and combine different principles from psychological research to measure the reliability of given sets of answers *after* conducting a questionnaire study. Measuring the reliability of sets of answers against the expected distribution under random-errors, we show how to identify definitely careless responding. Second, by comparing different treatments to be described below, we want to find out how to design and administer the questionnaire best in order to maximize answer quality *before* conducting the study.

The study was conducted in two waves with different treatments. The questionnaire itself was virtually identical between Experiment I and Experiment II, with one minor difference: in Experiment II, we left out the fifth screen of the original questionnaire that had been included for an unrelated study (Wolff, 2016). On the screen, participants read: "please choose one of the following four boxes and click on it: A, B, A, A."

In Experiment I, we varied the order in which participants received their payment. In the byFinish treatment, participants were called to the exit for payment as soon as they completed the questionnaire (first-come–first-served). In the

[7]For comparison: in an Internet survey with Psychology student participants, Meade and Craig (2012) identify 10-12% of their participants as careless, in addition to 12% participants not completing the online survey.

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

BASELINE treatment, all participants had to wait until everybody completed the questionnaire before receiving their payment by cubicle number. We randomly started the payment procedure either with the lowest or highest cubicle number.

While BYFINISH potentially sets the unintended incentive to answer as fast as possible, BASELINE avoids setting this incentive at the cost of a longer total duration of the session.[9] Treatment CAPICO (capped inverse completion order) tries to balance these two opposing goals using the following procedure: The payment procedure 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 order of payment is the reversed completion order in this group. After these two thirds 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 CAPICO to our participants, we 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, we also included justifications in BYFINISH ("to avoid unnecessary waiting time") and BASELINE ("for fairness reasons").

In Experiment II, we ran the treatments simultaneously within the same sessions, allocating participants to treatments randomly. We fixed the procedure to BASELINE and varied other treatment variables in Experiment II.

In ENTERNAME, we asked participants to type in their name for the experimenter to prepare their receipts. This happened on the first page of the questionnaire, which was followed by the general instruction page and the social-desirability scale. In NONUMBER, participants were not shown which questionnaire form out of how many questionnaire forms they were filling in at each 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, we set out to test whether seeing the end come closer would have a motivating effect on the participants.

We also test whether framing the usual show-up fee as a payment for questionnaire completion evokes (additional) reciprocity in our participants. For this purpose, we used such a framing in all our treatments but PAYMENTCALLED-SHOWUPFEE, in which we left out the according statement on the questionnaire-announcement screen, not reminding them of the show-up fee, either.[10]

Two further treatments served as control conditions. Given that the preceding experiment was different in Experiment II, we replicated BASELINE. Furthermore, NOJUSTIFICATION was a control treatment in which we left out the justifi-

[9] We will use the recorded completion times to provide an estimate of the dimensions of the tradeoff.

[10] Note that we can analyse the effect of experimental earnings on answer reliability directly, so that we do not need an additional treatment without a show-up/questionnaire-completion fee.

cation for the payment procedure, to test whether mentioning the fairness aspect in our initial experiment triggered more fair behavior, i.e. reliable answers, in the questionnaire.

3 Hypotheses

Before we can use either of our indices in treatment comparisons, we have to make sure they measure what they are intended to measure. For this purpose, we will first relate each of our partial indices and their combination 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 we have established that our indices measure what they are supposed to measure, we can use them to examine whether our treatment conditions affect the reliability of participants' answers. The first hypothesis comes directly from the different incentives between byFinish and Baseline (assuming time is a good for the participants).

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

Conjecture: There is no difference between capICO and Baseline.

In enterName, participants have to type in their name "for the experimenter to prepare the receipts" (which we did, and which is the usual reason for letting participants type in their names). Participants in this treatment entered their name before the introductory page of the whole questionnaire, which was followed immediately by our norm-conformity questions. We expect a negative effect of enterName on answer quality. This is due to the notion that 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 PAYMENTCALLEDSHOWUPFEE have lower answer quality.

Regarding the progress report on screen, there may be two counteracting effects. 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 different treatments described above. We appended the questionnaire to different preceding experiments, holding the preceding experiment constant for Experiment I where we 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. The experiments took place in the *Lakelab*, the laboratory for experimental economics at the University of Konstanz. Experiment I took place from January to May 2012, Experiment II between November 2015 and November 2016.[11]

[11]Note that our questionnaire required the preceding experiments to be short, we did not want the same participants to fill it in twice, and we 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. Table 1 summarizes the treatments and the number of observations.

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

Table 1: Number of observations per treatment. [a]not mentioning the "payment for 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 Descriptive Statistics and Construction of the Carelessness Index

Figure 1 illustrates the distributions of the different indices in our sample. The inconsistency index has its mode at 1 inconsistent answer pair (out of 15). The rarity index shows that most participants give either no or one rare answer, 20% give two and 13% give three or more rare answers. For the straightlining index, most participants have values of 3 or 4 items in a row, and 12% have index values of 5 or higher.

Table 2 shows the intra-personal correlations of the different indices for careless answers. All indices are significantly positively correlated with each other.

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

We conclude from the analysis so far that the combined index is a good measure for careless answers. Figure 2 illustrates the distribution of the final (continuous) carelessness index, plotted against the density function of the 100'000

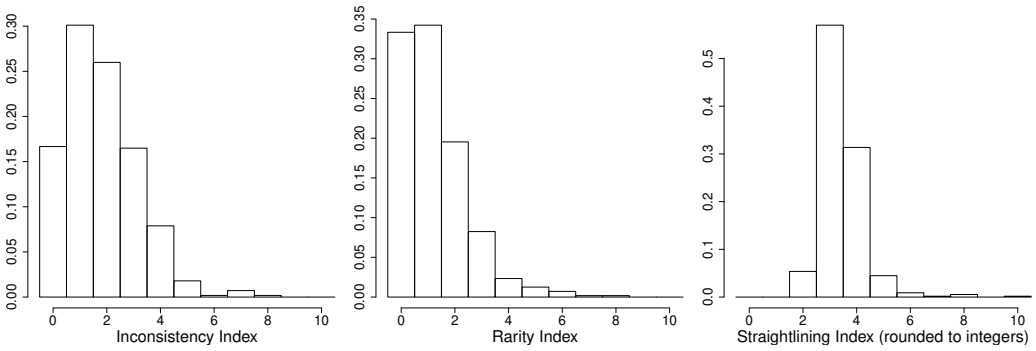

Figure 1: Distributions of index values.

	Unreliability	Inconsistency	Rarity	Straightlining
Unreliability	1.00			
Inconsistency	0.23***	1.00		
Rarity	0.27***	0.27***	1.00	
Straightlining	0.39***	0.20***	0.16***	1.00

Table 2: Correlations between the individual values of the different scales. *** denotes significance at the 1% level, ** at the 5% level and * at the 10% level.

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 a carelessness-index value above this 95% quantile is classified as "definitely careless." We will use both the continuous and the binary index in the treatment comparisons in the next section.

Before we look at the treatment comparisons, we provide some further indications that our carelessness index measures what it is supposed to measure. For this purpose, we look at the relationships between the index values and patience, motivation, and time. Patience as measured in the patience questionnaire correlates negatively with unreliable answers ($\rho = -0.08, p = 0.072$), and definitely careless agents are less patient (35.4 *vs.* 37.1 points on the patience scale, $p = 0.012$ in a t-test). The same holds true for participants who seem to be more motivated, judging by how much they write in the open-answer questions ($\rho = -0.19, p < 0.001$, or 32.2 letters *vs.* 48.2 letters, $p < 0.001$). While there seems to be a correlation between the time needed to fill in the questionnaire and carelessness overall ($\rho = -0.08, p = 0.051$), the correlation is much stronger for the fastest participants. Among those who need less than 14 minutes to complete the questionnaire, the correlation is substantial ($\rho = -0.34, p < 0.001$). Furthermore, there are clearly more definitely careless participants amongst those who

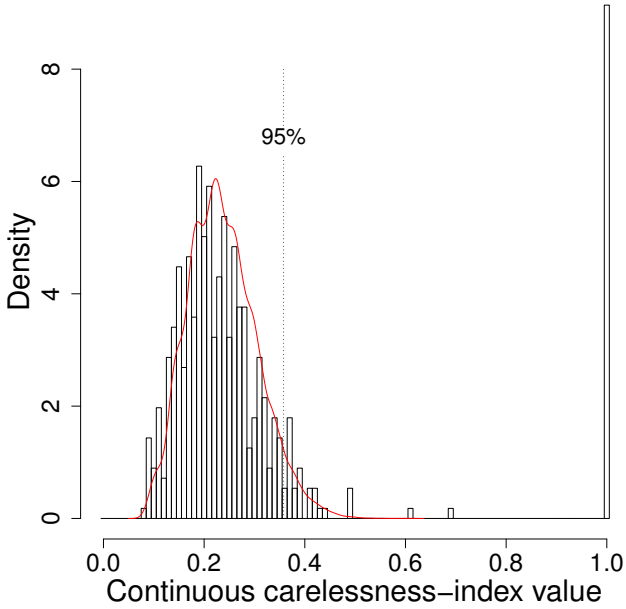

Figure 2: Frequency distribution of the continuous carelessness index, together with the density function of the 100'000 simulated random-error agents (red) and its 95% quantile (dotted line).

need less than 14 minutes than amongst those who need more (30% *vs.* 14%, $p = 0.004$ in a Boschloo-test)

We illustrate the relationship between total time used and carelessness-index value in Figure 3. The lowess smoother depicted in red clearly illustrates the negative relationship among the fastest participants. Even if we use only the preliminary index, those three participants who answer faster than a fast reader would take to merely read the questionnaire (those who take less than 8 minutes) display an average value of 0.57, compared to a value of 0.24 among everybody else.

As a final exercise, we set out to relate participants' carelessness-index value to their behaviour in the preceeding experiment.[12] 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 our participants in Experiment II, we elicited both actions and beliefs.[13] If we relate the participants' carelessness-index value

[12]We are grateful to Marie Claire Villeval for inspiring this analysis.

[13]Of course, this was done in an incentive-compatible way: participants knew they would not

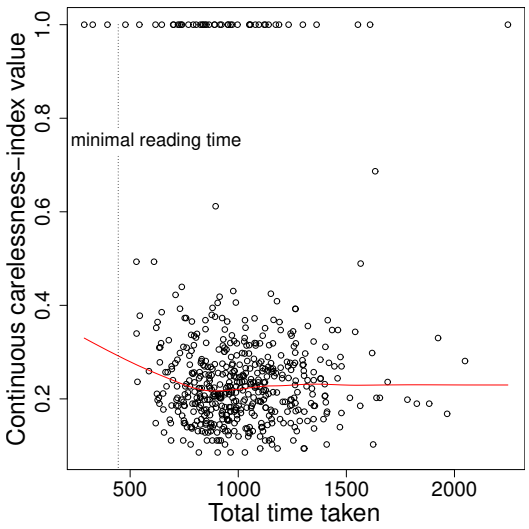

Figure 3: Continuous carelessness index by total time taken (lowess smoother in red)

to their average belief-action-consistency rate over the 24 rounds they played, we find a clear and substantial negative correlation ($\rho = -0.306, p = 0.012$). Comparing the average consistency rate of those who are definitely careless by our binary index (5 out of 67) to those who are not, the average belief-action-consistency rate is 45.8% amongst the definitely careless, compared to 70.2% amongst the rest (a Wilcoxon Mann-Whitney test yields $p = 0.036$, a $t-$test $p = 0.065$). In other words, some of the participants seem to pay only insufficient attention to the experimental tasks in general.

Summing up, the evidence presented so far strongly supports the claim that our carelessness index indeed provides a useful measure for participants' degree of careless answering. Having established the validity of our measure, we now turn to using the index for examining the question of how post-experimental questionnaires should be administered.

5.2 Treatment comparisons

In this section, we focus on the question of which payment order researchers should choose when administering post-experimental questionnaires. For ease of argumentation, we will use "answer quality" as our variable of interest in this subsection. Answer quality is simply $1-$ continuous carelessness index. Again,

be paid for their action and their belief in the same decision situation, and beliefs were incentivized by a binarised scoring rule that is proper for any degree of risk aversion.

we present both continuous and binary index values for all our treatments. Table 3 summarizes their average values across treatments.

Treatment	Answer quality	Of sufficient quality (in %)
Baseline	0.73	88.5
byFinish	0.64	77.2
capICO	0.69	84.4
enterName	0.73	94.4
paymentCalledShowUpFee	0.73	81.9
noNumber	0.68	84.3
numberAfterTwoThirds	0.74	93.5
Baseline.II	0.68	83.8
noJustification	0.75	94.9

Table 3: Average values of the continuous and binary answer-quality indices by treatment.

Table 4 reports the results of regression analyses testing for differences in answer quality across treatments. In the left half of Table 4, we regress the answer-quality index on the treatments, using standard ordinary-least-squares regressions. The base category is Baseline. The first model controls only for whether the questionnaire was appended to Experiment I or Experiment II, the second model additionally controls for other influences. In particular, we 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 we expect them to be related to answer quality. We use the motivation and time measurements relative to the respective treatment average to account for the fact that our treatment conditions will also affect these two measures and that we are interested in the total treatment effects. In addition, we control for the total earnings from the experiment, participants' value on the norm-conformism scale, their gender, and whether they study economics. In the right half of Table 4, we regress the corresponding binary index of displaying sufficient answer quality (1 - binary carelessness index) on the same variables, using a probit regression and reporting the average marginal effects.

Table 3 suggests that byFinish produces the most careless answer sets, supporting Hypothesis 2. In Table 4, we see that byFinish indeed produces the lowest answer quality and somewhat fewer classifications as "of sufficient answer quality." At the same time, the coefficient for capICO never reaches significance in any of the specifications (note, however, that $0.1 < p < 0.15$ in both OLS regressions). If we change the base category to capICO, the coefficient for byFinish remains significant in the continuous-index specifications. Hence, we can conclude that the payment order byFinish induces more careless answering

	continuous index (OLS)		binary index (probit, marg. effects)	
(Intercept)	0.822 (0.011)***	0.807 (0.013)***		
byFinish	−0.048 (0.015)***	−0.049 (0.014)***	−0.092 (0.056)	−0.097 (0.055)·
capICO	−0.020 (0.013)	−0.021 (0.013)	−0.038 (0.045)	−0.040 (0.044)
EnterName	0.032 (0.018)·	0.024 (0.019)	0.060 (0.033)·	0.038 (0.050)
Total earnings		0.002 (0.001)*		0.004 (0.002)·
paymentCalledShowUpFee	0.027 (0.017)	0.017 (0.018)	0.023 (0.040)	−0.017 (0.059)
noNumber	0.004 (0.017)	−0.012 (0.018)	−0.010 (0.048)	−0.070 (0.070)
numberAfterTwoThirds	0.027 (0.017)	0.016 (0.019)	0.052 (0.035)	−0.001 (0.058)
NoJustification	0.039 (0.017)*	0.019 (0.019)	0.065 (0.031)*	0.026 (0.051)
Experiment II	−0.041 (0.016)*	−0.040 (0.018)*	−0.039 (0.047)	−0.023 (0.055)
Total time†		0.000 (0.000)		−0.000 (0.000)
Motivation†		0.000 (0.000)***		0.002 (0.000)***
Norm Conformism		−0.000 (0.001)		0.000 (0.003)
Patience		0.002 (0.001)**		0.004 (0.002)*
Female		0.008 (0.008)		0.032 (0.023)
Economist		−0.005 (0.008)		−0.033 (0.025)
R^2	0.023	0.063		
Adj. R^2	0.014	0.045		
RMSE	0.107	0.105		
Num. obs.	884	806	884	806
Log Likelihood			-290.671	-252.614
Deviance			581.342	505.229
AIC			599.342	537.229
BIC			642.403	612.302

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$; †relative to the treatment average, to control for treatment effects on the measures; ‡cf. ftn. 7.

Table 4: Regressing answer quality on treatment conditions and further controls (probit: probability of providing sufficient quality).

than either Baseline or capICO. If there is any difference between Baseline and capICO, it is too subtle to manifest itself in our data clearly.

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

Up to this point, it seems that we unambiguously should recommend researchers to use the Baseline 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 Baseline 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 Baseline. According to the comparison of completion times in Table 5, participants in Baseline on average need about 4.5 minutes longer than those in byFinish (1491 seconds compared to 1227 seconds).

All in all, the improvement in answer quality in Baseline comes at the cost of

average	simulated average maximum	Treatment
937	1491	Baseline
834	1227	byFinish
920	1528	capICO
1025	1600	Baseline.II
1082	1632	noJustification

Table 5: Questionnaire-completion times (excluding the payment procedure) in seconds. For the simulated average maximum time within a session, we used a simulation to control for varying session sizes (the average of the 4 maxima of 6-people sessions will be lower than the maximum of one 24-people session). The average maximum values in the table are an average of 1'000 hypothetic sessions of 24 participants randomly drawn with replacement from the corresponding treatment population. Note: times in the first three rows are adjusted for the time used on the form that was missing in Experiment II.

about 15 additional minutes duration. If this is considered crucial, capICO may be a viable compromise between reliable answers and completion speed. It shortens the payment procedure by more than five minutes compared to Baseline and causes only insignificantly more careless answers.

Many experimenters ask participants to enter their name in the beginning of the questionnaire in order to print automated receipts. According to our regression analyses in Table 4, this practice is not harmful for answer quality. If at all, enterName has a positive effect on answer quality.[14]

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 our data. This is a very robust finding, which is stable in all specifications. However, framing the showup fee as a payment for answering the questionnaire does not have an impact on the carelessness index. Thus, it seems that it is not a reciprocity motive that helps improving answer quality, but rather general satisfaction with the experiment in general and their payment

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

in particular that supports participants' motivation to fill out the questionnaire carefully.[15]

Result 4. (i) Participants who get a higher payment have higher answer 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. If at all, it may be advisable to show the progress report only in the last part of the questionnaire, as indicated by the positive coefficients in Table 4 ($p-$values for the first and third model are $0.1 < p < 0.15$). In fact, participants in NUMBERAFTERTWOTHIRDS show clearly the highest values on our 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 BASELINE level).

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

Table 4 also shows that NOJUSTIFICATION is associated with the highest values on the answer-quality index. This is something we did not expect. We expected that briefly alluding to "fairness" when explaining the payment order in BASELINE would trigger more careful answering than NOJUSTIFICATION, if anything. Potentially, participants did not see the connection between paying by cubicle and fairness, instinctively disapproved the misuse of the fairness ideal, and therefore reacted in a negative way. This is pure speculation, however.

There are two more factors that influence the answer quality in our data: 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.

6 Summary and conclusions

This paper is about careless answers in post-experimental questionnaires. We want to identify careless answers *ex post*, and we want to use procedures to prevent them *ex ante*. In the paper, we took four different principles used in the social sciences to identify careless or dishonest answering and combined them into a single, powerful index. We designed a questionnaire that allows to get measures

[15] This corresponds to the finding that answer quality was lower in Experiment II. In this study, the preceding task was rather annoying to the participants, so their overall satisfaction was presumably lower and this reduced their willingness to invest effort into the questionnaire.

on the corresponding scales. All four scales that had been claimed to indicate unreliability of answers in prior research correlate well in our study. Further, our combined continuous carelessness index also correlated as expected with additional measures like questionnaire-completion times, values on a patience scale, and a measure of motivation. In light of the findings, we are confident that our carelessness index is valid in the sense that it measures what it was designed to measure.

In a next step, we simulated a random-error benchmark to identify "definitely careless" participants. We identify the "definitely careless" by singling out answer sets that yield index values so high that the likelihood of observing values that are at least as extreme under the random-error benchmark is less than five percent.

Both the continuous and the binary carelessness index provide a way of assessing the consequences of different procedures of how to administer questionnaires after economic experiments. We find that the byFinish payment procedure leads to significant increases in carelessness compared to the other two procedures, while capICO leads to only insignificantly higher carelessness compared to Baseline. At the same time, in particular Baseline is associated with substantial costs in terms of overall session time, and conceivably, also of participant annoyance. In light of these findings, capICO lends itself to being a useful compromise unless either care or speed are of utmost importance.

Asking participants to enter their name before answering the questions does not reduce answer quality. Furthermore, we find that the experience participants make in the preceding experiment influences answer quality: quality improves when the experiment was sufficiently interesting and reasonably paid. At the same time, our 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, for repeated choices between risky gambles, or Nyarko and Schotter, 2002, for belief-action consistency). In our data, answer quality is strongly negatively related to behavioural consistency in the experiment. This suggests that some of the participants pay only insufficient attention to the experimental tasks in general. We 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.

Technical acknowledgements

All experiments were computerized using z-Tree (Fischbacher, 2007), participants were recruited using ORSEE (Greiner, 2015) with Mozilla Firefox. The sta-

tistical analyzes 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

Bigoni, M., and D. Dragone (2012). Effective and efficient experimental instructions. *Economics Letters* 117(2), 460–463.

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

Crowne, D. P., and D. Marlowe (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology* 24, 349–354.

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.

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.

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.

Trautmann, S.T., and G. van de Kuilen (2015). Belief elicitation: a horse race among truth serums. *Economic Journal* 125, 2116–2135.

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

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.

Wolff, I. (2016). Elicited Salience and Salience-Based Level-k. *Economics Letters* 141, 134–137.

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.

THURGAU INSTITUTE
OF ECONOMICS
at the University of Konstanz

Hauptstrasse 90
Postfach
CH-8280 Kreuzlingen 2

T +41 (0)71 677 05 10
F +41 (0)71 677 05 11

info@twi-kreuzlingen.ch
www.twi-kreuzlingen.ch