This is the accepted version of a manuscript published by Oxford University Press. | Pölzler, Thomas (forthcoming): Insufficient Effort Responding in Experimental Philosophy. In: Lombrozo, Tania; Knobe, Joshua; Nichols, Shaun (eds.): **Oxford Studies in Experimental Philosophy, Volume 4**. Oxford: Oxford University Press.

Insufficient Effort Responding in Experimental Philosophy

THOMAS PÖLZLER
Department of Philosophy, University of Graz

ABSTRACT: Providing valid responses to a self-report survey requires cognitive effort. Subjects engaging in insufficient effort responding (IER) are unwilling to take this effort. Compared to psychologists, experimental philosophers so far seem to have paid less attention to IER. This paper is an attempt to begin to alleviate this shortcoming. First, I explain IER's nature, prevalence and negative effects in self-report surveys in general. Second, I argue that IER might also affect experimental philosophy studies. Third, I develop recommendations as to how experimental philosophers should (and should not) try to prevent IER. Fourth, I develop recommendations as to how experimental philosophers should (and should not) try to detect IER. Fifth, I sketch how experimental philosophers ought to proceed once a subject has been identified as an insufficient effort responder. And finally, I report the results of an online survey that addresses experimental philosophers' current knowledge, consideration and assessment of IER.

KEY WORDS: experimental philosophy; insufficient effort responding; methodology; attention checks; MTurk

In the last two decades analytic philosophy has seen the rise of a novel interdisciplinary approach going by the name of "experimental philosophy" (see Knobe and Nichols 2017). Experimental philosophers conduct scientific studies that are designed to test the empirical premises of arguments for philosophical conclusions.[1] These studies haven taken different forms (see Rose and Danks 2013). In what follows I will focus on what has made up the bulk of the extant experimental philosophy research, namely studies that target ordinary people's

[1] These conclusions can be either substantive (as in experimental philosophy's "positive program") or methodological (as in its "negative program," which purports to undermine traditional philosophical methods such as the method of cases) (see Kauppinen 2007). In addition, many experimental philosophers have also used scientific data to illuminate human cognition and to reveal patterns of intuitions that require explanation (see Knobe 2016; Knobe and Nichols 2007) — a project that might be claimed to go beyond testing the empirical premises of arguments for philosophical conclusions.

intuitions about philosophical concepts (such as intentional action, free will, and knowledge). These studies share the following features:

- **Participants:** ordinary people (in the sense of philosophical laypersons)
- **Hypotheses:** hypotheses about the content, causes and underlying psychological mechanisms of these people's intuitions about philosophical concepts
- **General Method:** quantitative methods of the social sciences, in particular psychology
- **Specific Method:** self-report surveys: subjects are asked to respond to hypothetical scenarios by reporting their intuitions about them

One thing that puzzled me when I first contributed to an experimental philosophy study in the above sense was how little cognitive effort many subjects seem to have taken in responding to it. For example, some subjects completed the study in less than half the time that would have been necessary to even only read all materials; and many failed very simple attention and comprehension checks.

In the psychology literature the phenomenon that I encountered has recently mostly been referred to as "insufficient effort responding" (IER) (Huang et al. 2012).[2] More and more psychologists have come to regard this phenomenon as a threat to the validity and reliability of self-report surveys (e.g., Curran 2016; Huang et al. 2012, 2015; Meade and Craig 2012). Yet, in experimental philosophy IER so far seems to have received less attention. This paper is an attempt to begin to alleviate this shortcoming. My aim is to synthesize the current state of the art regarding IER; to infer preliminary and rough guidelines for experimental philosophers (including references for further investigation); and, most of all, to raise awareness for the issue.

Here is how I will proceed. First, I will explain IER's nature, prevalence and negative effects in self-report surveys in general. Second, I will argue that IER might also affect experimental philosophy studies in particular. Third, I will develop recommendations as to how experimental philosophers should (and should not) try to prevent IER. Fourth, I will develop recommendations as to how experimental philosophers should (and should not) try to detect

[2] Other common labels include "careless responding," "satisficing," "random responding," "content-independent responding," and "content-nonresponsitivity." Note, however, that most of these labels denote somewhat narrower or otherwise slightly different phenomena (see Huang et al. 2012, as well as Sec. 1).

IER. Fifth, I will sketch how experimental philosophers ought to proceed once a subject has been identified as an insufficient effort responder. And finally, I will report the results of an online survey that addresses experimental philosophers' current knowledge, consideration and assessment of IER.

1 Insufficient Effort Responding in Self-Report Surveys

Providing valid responses to a self-report survey requires cognitive effort. Ideally, subjects would go through the following five steps (see Krosnick 1991; Tourangeau et al. 2010): (1) they read the survey's instructions, questions, answer options or items; (2) they attempt to comprehend them, i.e., to grasp their intended meaning; (3) they compute and/or search their memory for relevant information; (4) they form judgements in response to the questions; and (5) they respond in ways that reflect these judgements. Unfortunately, subjects are not always willing to conform to this model. Sometimes they respond without having completed some or even any of its steps, even though they would have been able to do so.[3] This is when IER occurs.

The label "insufficient effort responding" was introduced by Huang et al. (2012). In line with the understanding assumed here, they define the phenomenon exclusively by reference to its underlying cause, namely a lack of motivation to take sufficient cognitive effort (independently of subjects' response patterns, intentions, etc.):

> To provide a comprehensive depiction of the phenomenon of interest, we propose the label of insufficient effort responding (IER), defined as a response set in which the respondent answers a survey measure with low or little motivation to comply with the survey instructions, correctly interpret item content, and provide accurate responses. (Huang et al. 2012: 100)

In order for IER to be a potential problem for self-report surveys it must be both prevalent in these surveys and have negative effects. Recent psychological research suggests that both of these conditions might be met.

[3] To reemphasize, as it is understood here, IER only pertains to cases in which subjects were not *willing* to complete all steps of the survey response model, and not to cases in which they were not *able* to do so. Poor quality responses that result from factors such as low cognitive abilities or insufficient grasp of the survey's language hence do not qualify as IER.

1.1 Prevalence

So far only few studies have attempted to determine the prevalence of IER in self-report surveys. Given that some samples exhibit more IER than others, and that there are no established IER indicators and indicator thresholds (Hauser and Schwartz 2016; see also Sec. 4), it might not come as a surprise that these studies' estimates vary considerably (see Table 1).

Studies	Minimum	Average	Maximum
Kurtz and Parish 2001		10.6%	
Johnson 2005		3.5%	
Ehlers et al. 2009		5%	
Curran et al. 2010	5%		50%
Meade and Craig 2012	10%		12%
Maniaci and Rogge 2014	3%		9%
Klein et al. 2014[4]	7%		47%
Hauser and Schwarz 2016[5]	5%		61%

Table 1: Some estimates of the proportion of IER subjects in self-report surveys

Some earlier studies (Johnson 2005; Ehlers et al. 2009; Curran et al. 2010) used indicators that can only detect one form of IER, and only if it occurs over the course of a number of items. They thus likely underrepresented the proportion of IER subjects (for this argument see Meade and Craig 2012). Some later studies (Klein et al. 2014; Hauser and Schwarz 2015) implausibly considered failure in even only one attention check as sufficient for IER. I propose that the most well-grounded estimates therefore lie somewhere in between these extremes. It seems that on average around 10% of subjects in self-report surveys do not show sufficient effort over considerable parts of these surveys (Meade and Craig 2012).

[4] Klein et al.'s data was analyzed by Hauser and Schwarz (2016).

[5] Hauser and Schwartz's minimum and maximum values pertain to different samples, namely MTurkers and students, respectively. In other experiments they also found that MTurkers fared better in terms of IER (4% vs. 74% on a novel attention check, and 74.5% vs. 87.8% on a structurally unusual attention check). For IER-related differences between crowdsourcing and student samples see also Sec. 3.1.

1.2 Effects

Researchers widely agree that when IER occurs at the rate reported above it can have serious psychometric effects (Huang et al. 2012, 2015; Curran et al. 2015; Leiner 2016; Liu et al. 2013).

For a long time it has been assumed that IER always or almost always results in random responding, i.e., in patterns which are not sensitive to the survey's content or form in any way (e.g., Beach 1989; Charter 1994; Curran et al. 2010; Johnson 2005). To the extent to which it takes this form IER can attenuate correlations, and hence increase the likelihood of Type II errors; i.e., effects that would have been significant if subjects had taken sufficient effort may be found to be not significant (e.g., Clark et al. 2003; McGrath et al. 2010). Moreover, random responding can also lower reliability estimates and cause errors in scale development and factor analysis (e.g., Johnson 2005; Meade and Craig 2012; Woods 2006).

To provide a concrete example, in an article entitled "Random responding from participants is a threat to the validity of social science research results" Osborn and Blanchard (2011) report a study on the effectiveness of two educational interventions. In the pre- and post-intervention parts of this study 40.0% and 29.5% of students were found to engage in random responding, respectively. With these students included in their analysis, Osborn and Blanchard did not find any significant difference in the extent to which the two educational interventions influenced students' test scores. But once the insufficient effort responders were removed, the effect reached the level of statistical significance ($p < 0.0001$), and its size increased by no less than 42%. That is, only addressing IER allowed Osborn and Blanchard to detect the substantial benefits of one method of instruction over the other.

In addition, it is important to stress that IER may not always result in random response patterns. Subjects who fail to complete all steps of the response model introduced above may sometimes rather choose the first answer option that seems reasonable to them (without considering the others); answer in the same way several times in a row; form answer patterns like straight lines, diagonal lines or zigzags; pick the "don't know" option even though they have an intuition about the matter at hand; and so on. IER of this kind cannot only attenuate, but also inflate correlations. In other words, it can lead researchers to reject the null hypothesis even though it is true (Type I error) (Huang et al. 2013, 2015; Woods 2006).

2 Insufficient Effort Responding in Experimental Philosophy

Our above considerations suggest that under certain circumstances IER can be a problem for self-report surveys. But is it also a potential problem for experimental philosophy studies in particular?

So far the prevalence and effects of IER have mainly been investigated with regard to studies in personality psychology.[6] It might be argued that these results do not generalize to experimental philosophy, either because experimental philosophy studies are relevantly different from personality psychology studies or for statistical reasons. In what follows I will argue that both of these arguments are weak. Then I will present some positive anecdotal evidence for the hypothesis that IER is also a potential problem for experimental philosophy studies.

2.1 The Relevant Differences Objection

Previous research has suggested that IER increases with a survey's length (e.g., Meade and Craig 2012; see also Sec. 3.3) and complicatedness (Gage et al. 1957). According to the most plausible version of the first objection, experimental philosophy studies tend to involve less IER than personality psychology studies because they are shorter and simpler.

Proponents of this objection are right that the average experimental philosophy study is short; shorter than the average personality psychology study.[7] But this only means that, all other things being equal, these studies are *somewhat* less prone to IER. It does not by itself imply that serious psychometric effects cannot occur (see Sec. 2.3). After all, in addition to survey length, IER is also sensitive to various other factors, such as the formulation of survey materials, personality traits, complicatedness, etc. (e.g., Bowling et al. 2016; Gage et al. 1957). Moreover, not *all* experimental philosophy studies are short. To provide an example from my own research, the study that I mentioned in the introduction to this paper took subjects on average around 50 minutes to complete (Pölzler and Wright 2020b; see also, e.g., Wagner et al. under review; Wright et al. 2013, 2014; Wright 2018).

[6] For example, many early studies focused on the Minnesota Multiphasic Personality Inventory, and in recent years many researchers have used items from the Revised NEO Personality Inventory.

[7] Think again of classic experimental philosophy studies like Knobe 2003 on intentional action, in which subjects are only presented one single vignette, and only asked two questions about this vignette. In contrast, the Revised NEO Personality Inventory (Costa and McCrae 1992) requires that subjects rate 60 distinct items.

The above objection's second sub-claim — that experimental philosophy studies are also generally simpler than personality psychology studies — strikes me as wrong in the first place. IER has often been tested based on the Revised NEO Personality Inventory (Costa and McCrae 1992). In such studies subjects are asked to indicate their level of agreement with items such as "I am a productive person who always gets the job done," or "I am sometimes completely absorbed in music I am listening to" (Costa and McCrae 1992). Many materials in experimental philosophy studies are equally or more difficult to process than this personality psychology research. Think, for example, of two of the most famous studies, in which subjects had to decide whether Gettier cases (cases in which a person has a justified true belief but their justification is false) qualify as knowledge (Weinberg et al. 2001), and whether an action's good/bad side-effects qualify as intentional (Knobe 2003).

2.2 The Statistical Objection

A second potential reason for assuming a low prevalence of IER in experimental philosophy (compared to personality psychology) is statistical. Psychologists recently had to confront the worrisome fact that only 36.1% of a selection of 100 studies could be replicated, with personality psychology ranging below average (Open Science Collaboration 2015). In contrast, the XPhi-replicability project's recent attempt to replicate 40 experimental philosophy studies resulted in a replication rate of about 70% (Cova et al. 2018). Doesn't this suggest that experimental philosophy is less affected by IER than personality psychology?

I take it that such an inference would be hasty. A first thing to note is that IER is most often problematic in that it attenuates effects. It prevents them from being detected as statistically significant (Sec 1.2). This means that those studies that are most often and most strongly affected by IER are studies with null effects, i.e., studies that fail to yield statistically significant effects. But both the Open Science Collaboration and the XPhi-replicability project did not test many studies of this kind. For example, only three of the tested experimental philosophy studies showed null effects (Cova et al. 2018). We therefore still lack sufficient data about the replicability of those studies that were most likely and most strongly to be affected by IER, and therefore cannot draw reliable conclusions about differing IER proportions from those replication attempts that have so far been made.[8]

[8] To restate and further clarify my above argument, I acknowledge that IER may have contributed to some replication failures (in psychology, experimental philosophy and elsewhere) in minor ways. However, since these

Moreover, none of the most common explanations for experimental philosophy's high replication rate cites a comparably lower level of IER either. Consider the XPhi-replicability project's proposed explanations (Cova et al. 2018). According to this project, experimental philosophy studies were more likely to replicate because they (1) showed larger effect sizes than psychology studies; (2) are typically survey-based and less costly (which means that they often involve larger samples, have more subjects in each condition, and can be more easily re-run and replicated); (3) mostly study how the content of certain stimuli (hypothetical scenarios) affects subjects' behavior (as opposed to contextual or demographic variables); and (4) are run by researchers who are more prone to thinking about methodological issues.

2.3 Anecdotal Evidence

Recent studies suggest that an IER subjects proportion of only 5% may suffice to attenuate or inflate correlations (Huang et al. 2012, 2015). So even if IER's prevalence in experimental philosophy were lower than in personality psychology (as claimed by the relevant difference and the statistical objections discussed above) this would still not rule out negative psychometric effects. However, to further substantiate that experimental philosophy actually likely involves a comparably high IER rate let me also present some positive anecdotal evidence.

In the introduction I have already touched upon IER results from my own experimental philosophy research. Here I would also like to mention a recent study that seems particularly unlikely to provoke IER, because it was very short (average completion time: 5.3 minutes) and simple (subjects were asked to select whether they agree/disagree/neither agree nor disagree with moral sentences or whether they take these sentences to be true/false/neither true nor false) (Pölzler and Wright 2020a).[9] In this study 17.92% failed an attention check that required them to pick the farthest left of five horizontally lined fields; 8.65% failed an attention check that required them to pick any of five horizontally lined fields that is not farthest to the left; 11.54% failed an attention check that required them to memorize an important part of

failures where mostly found in non-null effect studies, and since IER mainly affects null effect studies, a number of other factors plausibly contributed in much stronger ways. Only once null effects get published more frequently, and are attempted to be replicated more frequently, IER can become a more powerful explanation of replication failures, and replication rate differences across disciplines may be partly traced back to differing rates of IER.

[9] As the study's results were not helpful in testing the targeted hypothesis, the study is only described in a footnote of this paper (Section „Moral Truth").

a sentence that they had been asked to rate on the previous page; and 48.22% finished at a pace that probably did not allow them to read through all materials carefully.

Following the 2018 Buffalo Experimental Philosophy Conference, some participants were willing to share IER-related data from some of their most recent experimental philosophy research with me as well. In studying stakes effects on knowledge Francis et al. (forthcoming) found that around 10% of their subjects failed at least one of two comprehension checks[10], and that around 4% of their responses probably originated from virtual private server farms or bots[11]. Park et al. (forthcoming) found that 13.41% of their lay subjects answered an open-ended question in a way that suggested a lack of competence. And in Roberts et al. (2018) 18.04% of subjects failed at least one of two simple comprehension checks about the content of scenarios.

In sum, the above considerations support that just as with personality psychology, IER is a potential problem for experimental philosophy studies as well. Researchers in this area should thus address IER.

3 Preventing Insufficient Effort Responding (in Experimental Philosophy)

One way of addressing IER is to detect and thereafter deal with it, for example by excluding subjects from analysis. This strategy has inevitable methodological downsides that will be discussed below (Sec. 5). It is thus advisable to try to prevent as much IER as possible before it even occurs (Ward and Pond 2015; Ward and Meade 2018).

In what follows I will introduce the most common measures that psychologists have so far taken or suggested to prevent IER: (1) choosing particular kinds of samples, (2) sufficiently compensating subjects, (3) shortening surveys, (4) simplifying surveys, (5) having proctors that oversee the completion of a survey, (6) including CAPTCHAS into online surveys, and (7) formulating instructions in certain kinds of ways. I will assess these measures' effective-

[10] Experiment 1: 13.39%, Experiment 2: 6.84%, Experiment 3: 10%, Experiment 4: 7.96%.
[11] Experiment 1: 6.67%, Experiment 2: 2.50%, Experiment 3: 0.08%, Experiment 4: 5.83%.

ness and provide some thoughts as to whether and how they may be useful in experimental philosophy studies in particular.

3.1 Sample

In recent years both psychologists and experimental philosophers have increasingly drawn on crowdsourcing samples, such as from Amazon Mechanical Turk. Some researchers have argued that these samples are more prone to IER than traditional student samples — say, because subjects from crowdsourcing samples often participate for exclusively or mainly financial reasons (Goodman et al. 2013). The available evidence does not support this hypothesis. Almost all relevant studies suggest that crowdsourcing subjects are not more, and probably even less, likely to engage in IER than students (see Hauser and Schwarz 2016; Klein et al. 2014; Paolacci et al. 2010).[12,13] Thus, IER-related considerations do not tell against experimental philosophers using crowdsourcing samples. If anything they support this practice.[14]

3.2 Compensation

It is natural to assume that if a subject feels that he or she receives too little money, research credits, etc. for participating in a study then this subject is more likely to show insufficient effort. A second way in which one might try to prevent IER is thus by increasing subjects' compensation. In self-report surveys this strategy again does not seem to improve data quality to any noteworthy extent (e.g., Burhmester et al. 2011; Horne et al. 2013; Meade and Warde 2018).[15] However, there is some evidence that the effectiveness of increased compensation is a function of subjects' economic or cultural situation. In particular, MTurkers from India showed considerably less IER upon better payment (Litman et al. 2015). This fact should be kept in mind by experimental philosophers doing cross-cultural research (which has recently become more common; see, e.g., Machery et al. 2017; Rose et al. forthcoming). Apart from

[12] The only study that suggests that students perform better than crowdsourcing participants (Goodman et al. 2013) is methodologically flawed (as its MTurk sample included many non-native English speakers, see Hauser and Schwarz 2016).

[13] Within Amazon Mechanical Turk workers are assigned rates that reflect the proportion of their HITS (human intelligence tasks) that were approved. There is evidence that requiring an approval rate of at least 95% decreases IER (Peer et al. 2014).

[14] To reemphasize, this section only addresses the extent to which the particular phenomenon of IER is a problem for crowdsourcing samples. It does not purport to account for other potential advantages or disadvantages of these samples, such as ethical concerns regarding compensation or representativeness of the general population.

[15] In contrast, there is some evidence that increasing compensation reduces IER in studies whose tasks have objectively correct answers (Aker et al. 2012).

that, the above findings suggest that increasing payment is not an effective strategy for preventing IER.[16]

3.3 Proctors

Today most experimental philosophy studies are conducted online. Some researchers worry that in this unsupervised setting subjects are more likely to engage in IER than in the presence of a proctor (say, because they do not feel sufficiently accountable or more easily give in to distractions) (Johnson 2005; Meade and Craig 2012). This worry appears to be ungrounded too. Several studies suggest that the physical presence of a proctor at best only minimally decreases IER, if at all (see Curran 2016; Francavilla et al. forthcoming; Klein 2014). The same holds true for "virtual proctors" as well: stylized human beings that are shown in the left margin of the computer screen throughout online surveys (Francavilla et al. forthcoming; Ward and Pond 2015).[17]

3.4 Survey Length

As mentioned in Sec. 2.1, IER has been found to be to some extent determined by the length of surveys. In personality psychology studies (which tend to be long) it occurs more often in the middle or towards the end. For instance, Meade and Craig (2012) report that while at the beginning of their study less than 5% failed attention checks, towards the end this proportion increased to 25%. Responses also became continuously more alike over time, in the sense that subjects were more likely to choose the same answer option several times in a row (see also, e.g., Baer et al. 1997; Berry et al. 1991).

The above findings suggest that one way of decreasing IER in experimental philosophy studies is to shorten them; for example, by using between-subject designs or limiting the number of tasks or items. However, this strategy must of course be balanced against countervailing considerations. Many studies can only be valid or reach a sufficient degree of statistical power if they involve a within-subjects design or a high number of tasks and items (such as with the study that I mentioned in the introduction). Moreover, the available evidence only

[16] Needless to say, there are strong *moral* reasons to compensate subjects fairly, e.g., to pay them at least their country's minimum wage.

[17] On an IER-unrelated note, experimental philosophers might also want to decide against having proctors because they might influence subjects' responses, especially if a survey is about personal or moral issues.

suggests that shortening long studies decreases IER; it has not yet addressed studies that are short (or rather short) from the beginning.

3.5 Complicatedness

Another feature that may increase self-report surveys' propensity to IER (also mentioned in Sec. 2.1) is their being complicated, i.e., difficult to understand (Gage et al. 1957). The empirical evidence for this effect is scarce. From a theoretical perspective, however, it almost seems like a truism. The more complicated a task or item the more cognitive effort it takes to read it, comprehend it, relate it to relevant memories, form a judgement about it, and respond to it in a way that reflects this judgement. All other things being equal, subjects are therefore less likely to complete all of these steps with regard to complicated than with regard to simple tasks and items.

Some experimental philosophy studies are quite complicated, such as the studies about Gettier cases and intentional actions mentioned above.[18] In these cases simplifying tasks and items may help to decrease IER. For example, one might substitute less by more familiar words (as indicated by word familiarity databases) and complex for simpler words (with the help of synonyms dictionaries), shorten sentences, structure questions in the most comprehensible way, introduce additional paragraph breaks, and highlight important terms. Analogously to the shortening recommendation, however, experimental philosophers of course should not sacrifice reliability or validity in employing this strategy (sometimes there just is no simple/r way of testing what one means to test); and simplifying materials that are already quite simple may not have any effect at all.

3.6 CAPTCHAs

The least effortful way of completing an online survey is by running a dedicated bot, i.e., a program that answers questions automatically.[19] In recent months more and more researchers

[18] Parts of my own research on folk moral objectivity are even more complicated. For example, in one task subjects were asked to choose among descriptions of moral realism and variants of anti-realism, such as "When a person says that something is morally right or wrong, good or bad, etc. she intends to state a fact. Such facts exist – and they are independent from what anybody thinks about them. For example, an action that is morally wrong is wrong no matter what anyone thinks. So it would still be wrong even if you yourself, or the majority of the members of your culture, thought that it is not morally wrong" (Pölzler and Wright 2020b).

[19] On the definition given in Sec. 1, running a bot qualifies as an instance of IER because those who engage in this behavior are unmotivated to complete any of the survey response model's steps. But even if one rejects this

have reported results and paradata which suggest the increased presence of such programs on MTurk (Stokel-Walker 2018). A simple and effective way of preventing some of this worst form of IER is by including a CAPTCHA ("Completely Automated Public Turing test to tell Computers and Humans Apart") at the beginning of one's survey — for example, to have subjects report letters that can be seen in a blurred image or to have them identify the location of certain objects in a picture. On many survey platforms prebuilt CAPTCHAs can be implemented within seconds. This measure of preventing IER is thus highly advisable for any experimental philosophy study (and indeed any survey using online crowdsourcing samples).

3.7 Instructions

A final way in which one might try to prevent IER is by including certain information in one's survey's instructions. Studies suggest that neither introducing the researchers and their expertise nor informing subjects that they will receive feedback about the quality and usage of their responses has any effect (Ward and Meade 2018).[20] Two other kinds of instruction-based measures, in contrast, are to be recommended.[21] First, experimental philosophers might appeal to subjects' conscience. Ward and Meade (2018) found less IER when they briefly informed subjects about the amount of work that goes into designing and conducting a study, had them type their initials next to six requirements (such as about the study's length), and finally asked them to type their name next to the statement "I promise to carefully read each item and to provide an honest response" (Ward and Meade 2018: 14-15).[22] Second, it may also prove effective to increase subjects' perception of meaningfulness; for example, by explaining the study's purpose and significance (even if only in vague and general terms), and thanking subjects for their participation (see Chandler and Kapelner 2013).[23]

classification, the issue is definitely related to IER, and is so important that it certainly warrants being mentioned.

[20] To be precise, these measures did not have any effect on objective IER measures. They did influence subjects' IER self-reports.

[21] There is another instruction-based measure that proved effective in deterring IER. Subjects had to provide ten reasons for sufficient effort responding and declare when did not respond with sufficient effort to a survey in the last year they. This measure seem impracticably long and effortful to me.

[22] Having subjects provide their names undermines the (perceived) anonymity of their responses. Hence, to me it seems preferable to have them type the above statement into a box, without revealing their name.

[23] Note that Chandler and Kapelner found an effect of increasing perceptions of meaningfulness in studies that have objectively right answers. It is not yet clear whether this effect generalizes to self-report surveys.

4 Detecting Insufficient Effort Responding (in Experimental Philosophy)

Even with the best prevention measures in place some IER is bound to happen. Experimental philosophers should thus also use strategies to *identify* such responses (so that they can deal with them). Seven indicators have so far been particularly common in self-report surveys: (1) attention checks, (2) response time measures, (3) comprehension checks, (4) self-reports on survey effort, (5) response pattern analyses, (6) response consistency analyses, (7) the identification of atypical responses, and (8) open-ended questions. In what follows I will introduce these indicators, assess their effectiveness, and provide some thoughts as to whether and how they may be useful in experimental philosophy studies in particular. Before that, however, three general remarks are in order.

First, each of the indicators that will be discussed below can only detect specific forms of IER. Experimental philosophers are thus well-advised to use more than one of them (Berinsky et al. 2013; Curran 2016; Meade and Craig 2012; Thomas and Clifford 2017). Second, different indicators suit different kinds of studies (depending on their length, usage or number of items, etc.) (DeSimone et al. 2014). And third, each of the indicators requires stipulating a threshold that distinguishes sufficient from insufficient effort. At the moment there are no established values for these thresholds. As research about IER indicators' effectiveness and potential unwanted effects is still at its beginning, and as one should also combine several of these indicators, I will here assume a conservative approach. That is, I will recommend low individual thresholds that flag only the worst IER subjects, as measured by each individual indicator (see Curran 2016; Leiner 2016). Note, however, that in the end the question of where to set these thresholds also depends on assumptions about the kind of intuitions that are relevant for philosophy. For example, if one believes that only intuitions that have resulted from thorough reflection can be philosophically relevant (e.g., Kauppinen 2007), one should set stricter IER limits than if one targets more immediate intuitions.

4.1 Attention Checks

The most well-known and common method for detecting IER are attention checks. These checks have typically taken four distinct forms: (1) instructional manipulation checks, (2) instructed-response items, (3) bogus/infrequency items, and (4) stand-alone checks.

Instructional Manipulation Checks

Instructional Manipulation Checks (Oppenheimer et al. 2009) measure the extent to which subjects attend to a task's instructions. A statement embedded in these instructions asks subjects to proceed in an unusual way. For example, it requires clicking on the page title to be forwarded to the next page or writing a predefined text in a box entitled "other":

> […] in order to demonstrate that you have read the instructions, please ignore the […] items below, as well as the continue button. Instead, simply click on the title at the top of this screen […] to proceed to the next screen. (Oppenheimer et al. 2009: 868)
>
> If you are reading these instructions please write "I read the instructions" in the "other" box. (Pennycook et al. 2014: 6)

Instructed-Response Items

Instructed-Response Items (e.g., Gummer et al. forthcoming; Kam and Chan 2018) measure subjects' attention to a task's items. At first sight they look like regular items. However, subjects are asked to provide specific predefined responses to them. Here are again two examples:

> Please choose *strongly disagree* for this item (Kam and Chan 2018: 83)
>
> click strongly agree (Gummer et al. forthcoming: 2)

Bogus/Infrequency Items

Another common way of testing subjects' attention to item content is to include items that most people would regard as absurd or highly probable/improbable (e.g., Beach 1989; Huang et al. 2014). Here are some examples of such "bogus" or "infrequency" items (with the first not to be recommended for reasons specified below):

> I work twenty-eight hours in a typical work day. (Huang et al. 2014)
>
> I was born on February 30th. (Beach 1989)
>
> I have never used a computer. (Breitsohl and Steidelmüller 2018: 291)

Stand-Alone Checks

In some experimental philosophy studies instructions are too short for instructional manipulation checks to work. The studies also may not ask subjects to rate various items in close succession. In these cases stand-alone attention checks (that make up a separate survey page or

task) may be more suitable. For example, as reported above, subjects could be presented five fields and asked:

> Please pick the response farthest to the left. (Pölzler and Wright 2020b)

General Discussion

In formulating any of the above attention checks several recommendations should be kept in mind. First, subjects may over time become familiar with standardized versions of these checks. Their effectiveness in detecting IER can therefore be increased by introducing novel content or structures (Thomas and Clifford 2017). Second, attention checks' required responses should vary within studies (e.g., not always "strongly disagree", or the farthest left response), so that they are more likely to detect IER subjects who continuously provide the same response (Desimone et al. 2015; Huang et al. 2014; Meade and Craig 2012). Third, there must not be plausible alternative readings of the attention checks that justify differing responses (such as, e.g., with the purported bogus sentence "I work twenty-eight hours in a typical work day" which might be read as expressing, in a hyperbolic way, that one is hard-working, see Breitsohl and Steidelmüller 2018). And fourth, all checks should be formulated in ways that minimize unwanted effects, such as emotions, amusement or answers based on social desirability (for this problem see again the above bogus sentence about a twenty-eight hours work day).

How many attention checks may a subject fail before he or she is classified as having engaged in IER? Nobody can pay full attention all of the time. Flagging subjects for failing just one check therefore seems too rigid (at least if the checks aren't extremely simple) (Berinsky et al. 2013; Thomas and Clifford 2017; Paolacci et al. 2010). Curran (2016) suggests inferring IER from failure in more than 50% of a study's attention checks. This, on the other hand, seems too liberal (at least if the checks aren't extremely difficult). A subject who fails as many checks as that likely does not show sufficient effort. In my view, for most kinds of attention checks, appropriate thresholds will lie somewhere in between these values; such as, for example, at a failure rate of 25%.[24]

[24] As indicated above, in setting their threshold researchers should also account for their attention checks' relative difficulty. For example, Instructional Manipulation Checks are generally more difficult to pass than other kinds of attention checks (as they are harder to get right by accident).

Attention checks have been found to be effective in detecting IER (e.g., Breitsohl and Steidelmüller 2018; Gummer et al. 2018; Huang et al. 2012; Leiner 2016). If properly done such checks also do not affect subsequent response behavior (e.g., Breitsohl and Steidelmüller 2018; Huang et al. 2015; Oppenheimer 2009).[25] Negative effects are especially unlikely if the survey instructions explicitly mention that there will be an assessment of subjects' carefulness (Breitsohl and Steidelmüller 2018). I therefore recommend that experimental philosophers make use of properly designed attention checks in their studies.

4.2 Response Time

A second common way of detecting IER is based on subjects' completion times (of the survey as a whole, individual pages or individual items).[26] If a subject answers very fast then he or she probably did not go through all stages of the idealized response process, as introduced in Sec. 1. Specifying what is *too* fast is challenging. Matjasic et al. (2018) helpfully distinguish two approaches: the statistical and the cognitive. Here I will discuss these approaches by the example of survey — as opposed to page or item — completion times (as survey completion times have been most commonly appealed to). (Note, however, that subjects sometimes enter a survey before they actually complete it (to enable participation) or take breaks during completion (to surf the internet, eat, etc.). These outliers at the level of individual pages must be substituted by alternative values, e.g., the page's median completion time (Leiner 2016).)The statistical approach determines completion time thresholds by considering the statistical properties of a survey's response times. For example, researchers have assumed that IER occurs when a subject's completion time is two standard deviations below the survey's mean completion time (e.g., Heerwegh 2003), is lower than the first quartile minus 1.5 interquartile ranges (e.g., Funke et al. 2011), is lower than the first percentile (e.g., Gummer and Roßmann 2015), or is lower than the fifth percentile (e.g., Harms et al. 2017) (for an overview see Matjasic et al. 2018).

The cognitive approach, in contrast, appeals to psychological properties. Huang et al. (2012) suggest that subjects cannot sufficiently process individual items in less than 2 seconds. This threshold has been taken up by several other researchers. However, it seems

[25] Neither do they make subjects feel tricked or insulted, nor do they reduce their reliance on conversational norms or instill a more deliberative mind-set.

[26] Most survey tools (such as Qualtrics) automatically measure survey completion times. To also measure page or item completion times, respective measures need to be manually implemented.

somewhat arbitrary; the average length of items varies considerably across studies; and many studies (especially in experimental philosophy) are not item-based at all. More appropriate and helpful specifications therefore rather appeal to general reading speed. For example, one may draw on Taylor's (1965) classic study of US students, according to which the average reading time per minute is 300 words, and set an IER threshold at half (150 words/minute) or two thirds (200 words/minute) of this value.

So far the statistical approach to setting completion time thresholds has been more common than the cognitive one. Under some circumstances (such as when subjects receive differing numbers of tasks) it is indeed more feasible or appropriate. In the majority of cases, however, the cognitive approach seems preferable. Response time indicators are motivated by the idea that completing a survey faster than at a certain speed just cannot be done without engaging in IER. This is the idea of an absolute threshold. While this threshold may be difficult to determine, the statistical approach involves the danger of having a slow sample, and thus flagging subjects who actually took enough time; as well as of having a fast sample, and thus failing to flag subjects who went too fast.

In any case, response time indicators have been found to be highly effective in detecting IER (Leiner 2016; Zhang and Conrad 2014). They are also invisible to subjects, and hence cannot possibly affect their responses in any way. I therefore strongly advise experimental philosophers to make use of these indicators.

4.3 Comprehension Checks

A third possible way of testing for IER are comprehension checks. Comprehension checks measure whether subjects have fully grasped the meaning of certain study materials or relevant concepts or claims. Failure to do so can have two distinct sources: (1) the subjects did not take sufficient effort, or (2) even though they took sufficient effort, they were not able to grasp the relevant meanings.

The above distinction suggests that comprehension checks are not pure IER indicators, in the sense of measuring *only* IER. This need not be regarded as a problem. First, the simpler comprehension checks are the more likely they capture IER rather than cognitive inability. One can thus render them almost pure. And second, for some experimental philosophy studies measuring cognitive ability is independently worthwhile. These studies' manipulations only work if subjects have a very good grasp of the meaning of certain materials. Moreover, it has been argued that for experimental research to potentially have implications for philosophy,

subjects need to be competent with regard to the concepts at issue anyhow (see Kauppinen 2007; Pölzler 2018)[27].

On the other hand, there are also experimental philosophy studies whose ambition is more empirical, e.g., to determine why people have certain kinds of philosophical intuitions. In these cases the results of comprehension checks may have to be utilized and interpreted in somewhat different ways. After all, it could be that the reason why people have certain intuitions is precisely that they lack in comprehension, either in the sense of being unwilling or unable to grasp certain philosophical propositions. By discarding comprehension check failures as IER straight away researchers could thus rob themselves of the possibility of gaining important insights into folk philosophical cognition. For an example of a comprehension check, consider again the folk metaethics study mentioned in the introduction (Pölzler and Wright 2020b). After having explained the difference between normative and metaethical sentences, we asked subjects a theoretical question that tested their understanding of this distinction, and had them classify a number of sentences as either normative ethical or metaethical. Subjects who failed in any of these checks were shown the instructions again, and had to complete the task for a second time. High failure rates at this second attempt were then treated as *pro tanto* evidence for IER.

So far comprehension checks have not received any attention in the general IER literature. There is thus no data about their effectiveness, compared to other indicators. Yet, from a theoretical perspective such checks again seem to be recommendable to experimental philosophers, adjusted to the aims of their respective studies; especially on the basis of conservative thresholds (e.g., IER equals failure in more than 25% of these checks).

4.4 Self-Reports

The most straightforward way of testing for IER is to simply ask subjects whether they engaged in it. For example, Meade and Craig (2012) ended their survey with a self-report scale involving items such as "I put forth _____ effort towards this study," "I gave this study ____ attention" (five response options), as well as with the question "In your honest opinion, should we use your data in our analyses in this study?" This measure did not turn out to be helpful. Answers to Meade and Craig's questions correlated only to a negligibly low degree

[27] Kauppinen of course argues that quantitative research cannot ensure that subjects have sufficient conceptual competence.

with the results of more objective IER indicators, such as attention checks or response time measures.[28]

4.5 Response Patterns

Sometimes IER manifests in patterned responses. The most common case (more common, e.g., than diagonal lines or zigzags) is when subjects give the same answer several times in a row. To detect such "straightliners" researchers have so far mainly used two measures: (1) LongString Analysis and (2) Scale Straightlining Analysis.

LongString Analysis

LongString Analysis (e.g., Johnson 2005) determines the longest sequence of identical responses by a subject. If this value exceeds a certain threshold the subject is classified as an IER responder. The threshold's most appropriate location depends both on the survey's length and content. Following Curran (2016), for example, one may flag subjects who give consecutive identical responses over half or more than half of a scale (or survey).

Scale Straightlining Analysis

Scale Straightlining Analysis (see Leiner 2016) counts the number of completely straightlined scales (or tables). Again, if this number is above a certain threshold the subject is classified as an IER responder. And again, the appropriate value of this threshold depends on the survey's length (in particular, the number of scales) and content.

Straightlining is particularly likely to occur with matrix tables (which allow subjects to rate multiple items in response to one question). In the absence of countervailing reasons such tables are thus to be avoided (Vannette 2016). But even in non-matrix settings LongString Analysis and Scale Straightlining Analysis may be effective ways of detecting IER (Leiner 2016; Meade and Craig 2012). While they may not be applicable to typical short experimental philosophy studies (as these studies only involve few questions, tables, scales or items), longer studies may be assessed in terms of these indicators. A helpful Excel macro for calculating LongString can be found at Landers 2016.

[28] One might argue that given special features of typical experimental philosophy studies (such as high conceptual complexity) self-reports function as a reliable IER indicator for at least those studies. This is an interesting empirical hypothesis that could be taken up by future IER research.

4.6 Response Consistency

During a survey subjects' intuitions normally do not considerably change. If they provide different responses to questions or items with the same or similar content this may therefore be taken as another indicator for IER. Response consistency has so far mainly been measured in four ways: (1) identical responses, (2) odd-even consistency, (3) psychometric consistency, and (4) semantic consistency.

Identical Responses

The simplest way of testing for consistency is to provide subjects with the same question or item several times (Buechley and Ball 1952). If their responses differ (considerably) this may be due to IER.

Odd-Even Consistency

In the odd-even consistency test (e.g., Huang et al. 2012, 2014; Johnson 2005) items on unidimensional scales are divided into odd (item 1, item 3, etc.) and even (item 2, item 4, etc.). For each subject the researchers then calculate the extent to which their odd and even responses correlate. If the value is low this indicates low individual reliability, and hence potential IER.

Psychometric Consistency

Psychometric consistency tests (Johnson 2005) identify those item or response pairs that correlated most strongly across the whole sample. Then they investigate how these items are correlated within subjects. If the within-subject correlation strays too far from the sample correlation this is again taken to indicate IER.

Semantic Consistency

To run semantic consistency tests (e.g., Goldberg and Kilkowski 1985) some of a study's questions, answers or items must be formulated in terms of semantic synonyms or antonyms. For example, after asking whether a person in a hypothetical scenario is "happy", one may present subjects with the same scenario at a later point, asking them whether the person is "joyful". If subjects' responses fail to be sufficiently correlated they are flagged for IER.

Evidence on the above strategies for detecting IER is scarce. Odd-even consistency (see Leiner 2016), semantic consistency (see Kurtz and Parrish 2001), and psychometric consistency (see Meade and Craig 2012) may not be particularly effective in detecting IER. Moreover, many experimental philosophy studies are not structured in a way that allows for odd-even tests; and semantic consistency tests seem ill-suited for these studies in that researchers typically target intuitions about specific concepts (e.g., happiness as opposed to joyfulness), and wording matters much. These tests are hence not be recommended. However, experimental philosophers may occasionally provide fully identical content, as in this case differing responses are a more reliable IER indicator.

4.7 Atypical Responses

IER is also a potential cause of subjects providing extreme (statistically unlikely) responses. Various statistics allow measuring how far a subject's responses stray from the sample mean across all of a survey's items (e.g., Mahalanobis 1936). Values that exceed a certain threshold (e.g., p=.05, p=.01, or p=.001, Meade and Craig 2012) have sometimes been taken to indicate IER. However, the effectiveness of this indicator is again unclear (see Meade and Craig 2012 versus Leiner 2016). Some statistics also require much power and presuppose a data distribution approximating normality (Meade and Craig 2012). Finally, especially in experimental philosophy studies some subjects simply may hold extreme views. In light of this lack of evidence and worries I do not recommend classifying atypical responses as IER either (see Curran 2016 for a similar conclusion).[29]

4.8 Open-Ended Questions

As explained in Sec. 3.6, there is reason to believe that more and more surveys have recently been completed by bots. MTurkers have also been found to use virtual private servers that enable them to participate even if they are not qualified to do so (for example, because of their geographical location or language) or to provide duplicate responses (Dennis et al. 2018). There are various suggestions as to how to detect this most serious threat to the integrity of survey data. One of the most accurate and simple indicators appear to be open-ended questions. If responses are provided by a bot or an unqualified person the answers to these questions are often either irrelevant, non-sensical or pasted from online sources such as Wikipedia

[29] That said, outliers may of course be removed for non-IER reasons.

(see Dennis et al. 2018; Francis et al. 2018). Experimental philosophy studies might hence involve at least one open-ended question for the purposes of IER detection.[30] This question need not be substantive. It can also be part of the instructions — as with the suggestion to have subjects type "I promise to carefully read each item and to provide an honest response" in a box (Sec. 3.7) — or part of gathering demographic data.

5 Dealing with Insufficient Effort Responding (in Experimental Philosophy)

Suppose a subject has been identified as an insufficient effort responder. What next? Unfortunately, there are no established protocols for this case. In what follows I will therefore explain how experimental philosophers might deal with IER, given current preliminary recommendations in psychology and some reflections of my own. In particular, I will investigate whether IER subjects should be excluded from analysis, how such exclusions should be reported, and whether insufficient effort responders should be compensated for participating in surveys.

5.1 Removing Data

Should data from subjects who were identified as insufficient effort responders be excluded from analysis? When experimental philosophers run a survey, they attempt to get at subject's intuitions about philosophical concepts. The responses of IER subjects mostly do not reflect these intuitions. In Sec. 1.2 I explained how this can have negative psychometric effects: attenuating correlations, and in some cases also inflating them. Hence, there are strong reasons for excluding IER subjects from analysis.

Excluding IER subjects from analysis has three downsides that need to be addressed. First, it increases researchers' degrees of freedom. By varying the consideration, weight or thresholds of IER indicators they could engage in p-hacking or other questionable practices that allow them to achieve statistically significant effects. In light of this possibility I strongly encourage experimental philosophers to define objective and transparent exclusion rules prior to running their studies (see Curran 2016; Thomas and Clifford 2017). Compliance with this

[30] Going beyond IER detection, the inclusion of qualitative methods in experimental philosophy has recently been convincingly advocated by Andow (2016).

recommendation can and should be demonstrated by preregistering studies — including these rules — at dedicated websites such as aspredicted.org or osf.io.

A second downside of excluding IER subjects is that it reduces sample size. This, in turn, reduces statistical power, i.e., the probability of rejecting a false null hypothesis (Ward and Meade 2018; Ward and Pond 2015). In planning their studies experimental philosophers need to take this effect into consideration. In particular, they will often have to start out with larger samples than they used to: at least 10%, or better even 20% larger than with studies that did not exclude IER subjects.

Finally, excluding subjects on the basis of IER indicators may introduce sampling bias, and thus limit generalizability (e.g., Berinsky et al. 2013; Meade and Craig 2012; Ward and Pond 2015).[31] The evidence concerning this effect has been somewhat contradictory. It is possible that IER is correlated with traits such as having a low level of education (e.g., Zhang and Conrad), being male (e.g., Berinsky et al. 2013), being less intelligent (e.g., Holbrook et al. 2007), and being young (e.g., Berinksy et al. 2013). To be able to detect such potential correlations experimental philosophers might gather a number of demographic features of their samples.

5.2 Reporting Results

Even if one excludes only one single subject based on IER indicators, this must be clearly reported. Often such exclusions will not have any effect on one's results. In this case a brief statement such as "The exclusion of insufficient effort responders that were identified based on [list indicators] did not have any significant effect on these results" suffices. If IER and non-IER subjects provided different responses, however, these differences must be explained in detail. Moreover, correlations between the IER subjects and demographic characteristics might be reported as well (Berinsky et al. 2013; Curran 2016; Desimone et al. 2015; Thomas and Clifford 2017). This could, for example, take the following form:

> The exclusion of insufficient effort responders that were identified based on [list indicators] did have a significant effect on these results. In particular, insufficient effort responders were more/less likely to [explain effect]. We found a significant correlation between insuffi-

[31] To a lesser extent, this worry also applies to some of the IER prevention strategies discussed in Sec. 3. For example, choosing a less IER-prone sample or formulating instructions in certain ways might also lead to sampling bias.

cient effort responders and [list demographic characteristics]. This means [explain potential implications for generalizability].

The requirement to report potential effects of excluding IER subjects has an important consequence for the timing of this exclusion. Sometimes subjects are kicked out of a survey immediately after failing on one or several indicators (such as attention or comprehension checks). In this case differences between IER and non-IER subjects in subsequent parts of the study cannot be investigated and reported. Given the above considerations, experimental philosophers should rather have every subject complete their study, irrespectively of their performance on IER indicators (Thomas and Clifford 2017).

5.3 Compensating Subjects

Anecdotal evidence as well as my below survey (see APPENDIX) suggest that most experimental philosophers fully compensate subjects even if they were found to have engaged in IER. From each individual researcher's perspective this practice may be rational. It spares them gaining special permissions for non-compensation from their universities' Institutional Review Boards (where such permissions are required), as well as dealing with the potential angry backlash of subjects whom they refused compensation (and judging from what I experienced and heard, such a backlash is to be expected).

However, if even the worst IER subjects are compensated this clearly threatens the integrity of subject pools such as MTurk. Why should one take sufficient effort if one will receive compensation anyhow: rushing through the survey, always picking the same response, etc.? I hence recommend that, given conservative indicator thresholds, experimental philosophers do not — or at least not fully — compensate IER subjects.[32] For ethical, legal and pragmatic reasons this should be clearly stated at the beginning of the study. For example, after having disclosed the usage of attention and comprehension checks, one's instructions may include a sentence like: "Poor performance on these attention and comprehension checks will result in you being disqualified from payment."

[32] Another possible solution is to compensate all subjects at a low level, and provide additional compensation to sufficient effort responders (but not to insufficient effort responders).

Conclusion

Insufficient effort responding has been considered a serious threat to the integrity of psychological self-report surveys. In this paper I first explained this phenomenon and argued that it is a potential problem for experimental philosophy as well. Drawing on current research in psychology, I then developed guidelines for how (and how not) to address IER in experimental philosophy studies. These guidelines are rough and preliminary. At least, however, they may provide a helpful basis for further discussion. Here is a brief summary of the central recommendations that emerged:

Preventing IER

- increase compensation in cross-cultural research (poor countries)
- keep surveys short (but do not additionally shorten already short surveys, and do not sacrifice reliability/validity for shortness)
- keep surveys simple (but do not additionally simplify already simple surveys, and do not sacrifice reliability/validity for simplicity)
- use CAPTCHAs
- in survey instructions appeal to subjects' conscience (researchers' work amount, participation requirements, careful responding declaration) and promote meaningfulness (significance and gratitude)

Detecting IER

- use attention checks, and disclose this in survey instructions (instructional manipulation checks, instructed-response items, bogus/infrequency items or stand-alone checks; keeping in mind the recommendations in Sec. 4.2)
- use response time measures with absolute thresholds (or, under some circumstances, with statistical thresholds)
- use comprehension checks, adjusted to the aims of studies
- in longer studies use response pattern analyses (especially in the case of matrix tables; which should, however, if possible be avoided)
- maybe use response consistency analysis; in particular, the analysis of identical content
- include open-ended questions

- use several different (versions of) the above indicators, and those that suit the survey in question
- for each indicator set conservative thresholds

Dealing with IER

- remove IER subjects from (at least part of) analyses
- report exclusions, as well as results with IER subjects included (even if there is no difference)
- preregister exclusion rules
- start with somewhat larger samples
- potentially report IER correlations with demographic characteristics
- do not kick IER subjects out during the study
- do not (fully) compensate IER subjects, and state this in the survey instructions

Following recommendations such as these may be another important step in improving the methodology of experimental philosophy studies. In any case, I hope that this paper has raised awareness for the problem of insufficient effort responding and initiates further discussion.

APPENDIX: A Survey of Experimental Philosophers

In the main part of this paper I explained how experimental philosophers *should* proceed with regard to IER. This appendix addresses the question of how they have *actually* proceeded. What do they know about IER? How big of a problem do they think it is? To what extent and in what ways do they attempt to prevent, detect, and deal with it? To shed light on these questions I developed and conducted an online survey.

Participants

Invitations to participate in the survey were sent out to all persons whose studies were tested by the X-Phi Replicability project (Cova et al. 2018), all persons who gave presentations at the Buffalo Experimental Philosophy Conferences (2012 to 2018), all persons who gave presentations at the German Experimental Philosophy Group Workshops (2015 and 2017), all persons who gave presentations at the Meetings of the Spanish Experimental Philosophy As-

sociation (2017 and 2018), and all persons who gave presentations at the last two conferences of the Experimental Philosophy Group UK (2017 and 2018). In total, 260 persons received invitations.[33]

The survey was completed by 53 persons. This low response rate (20,39%) is mainly explained by the fact that I asked invitees to only take the survey if they have ever actively contributed to an experimental philosophy study in the narrow sense assumed in this paper, i.e., a survey on lay persons' views about a philosophical question. Many invitees did not meet this condition. They had only contributed to experimental philosophy studies in a broad sense or to theoretical work in experimental philosophy (as evinced both by their CVs and occasional response e-mails).

Participants' mean age was 40 years. 7 were full professors, 10 were associate professors, 14 were assistant professors, 7 were postdocs, 3 were lecturers, 7 were PhD students, and 4 identified as "other". The majority's primary affiliation was only with a philosophy department (25 participants), only with a psychology department (10 participants) or with multiple departments (8 participants). 28 participants held a PhD in philosophy, 11 in psychology, and 8 in other fields of the humanities or sciences. 14 participants listed MA or BA degrees in both philosophy and a science. Finally, many participants selected multiple experimental philosophy subfields as their primary area of research, with the experimental philosophy of ethics (29 participants) being selected most often, followed by the experimental philosophy of action (11 participants) and the experimental philosophy of language and mind (10 participants each).

The study was preregistered at aspredicted.org. To detect IER I measured response times, included several attention checks, and performed a scale straightlining analysis. Subjects were excluded if they completed the survey in less than 04:00 minutes, failed two or more attention checks or straightlined one or more survey pages. On the basis of these criteria one subject was excluded. This left a total of 52 participants for analysis.

Methods

At the beginning of the survey participants were informed about its purpose and received an explanation of IER (which was similar to the explanation provided in the first paragraph of

[33] This number excludes 40 persons whose e-mail addresses were not available online and 19 undeliverable e-mails.

Sec. 1). Then they were asked questions pertaining to four categories: (1) preventing IER, (2) detecting IER, (3) dealing with IER, and (4) general questions.[34]

Preventing Insufficient Effort Responding

The part on preventing IER asked subjects how often they had used certain strategies for the specific purpose of preventing IER in studies that belong to experimental philosophy (in the sense of surveys on lay persons' philosophical views). They were presented the following items (including brief explanations) which they could rate on a scale from "never" to "always":

(P1) using online crowdfunding samples
(P2) using student samples
(P3) sufficiently compensating subjects
(P4) shortening surveys
(P5) simplifying surveys
(P6) having a physical proctor oversee the completion of a survey
(P7) having a virtual proctor oversee the completion of a survey
(P8) including CAPTCHAS
(P9) explaining the survey's purpose
(P10) warning against negative consequences of IER
(P11) having subjects sign or type a carefulness or honesty declaration
(P12) introducing yourself and/or your expertise
(P13) informing subjects that they will receive feedback about their responses
(P14) thanking subjects
(P15) explaining the survey's significance

Detecting Insufficient Effort Responding

The part on detecting IER asked how often subjects had used strategies for the specific purpose of detecting IER in studies that belong to experimental philosophy (in the sense of surveys on lay persons' philosophical views). Subjects were presented the following items,

[34] Some of these questions were taken or inspired by a survey that Liu et al. (2013) conducted among the members of the Society for Industrial and Organizational Psychology.

which again included brief explanations, and again were supposed to be rated on a scale from "never" to "always":

(D1) instructional manipulation checks
(D2) instructed-response
(D3) bogus/infrequency items
(D4) stand-alone checks
(D5) absolute response time threshold
(D6) statistical response time threshold
(D7) comprehension checks
(D8) self-reports on survey effort
(D9) Scale Straightlining Analysis
(D10) LongString Analysis
(D11) response consistency analyses
(D12) odd-even consistency
(D13) psychometric consistency
(D14) semantic consistency
(D15) identical responses
(D16) identification of atypical responses
(D17) open-ended questions

Dealing with Insufficient Effort Responding

The part on dealing with IER asked subjects how they usually proceed if they detect IER. The items looked as follows, again including brief explanations, and again to be rated on a scale from "never" to "always":

(W1) I exclude IER subjects from (at least parts of) my analyses
(W2) I define exclusion rules prior to running the survey
(W3) I preregister my exclusion rules at dedicated websites such as aspredicted.org or osf.io (at least since very recently)
(W4) In anticipation of IER I start with a larger sample size
(W5) I report that I excluded IER subjects in my paper
(W6) I report to what extent IER is correlated with demographic characteristics

(W7) I report what difference excluding IER subjects makes to the results

(W8) I immediately kick subjects out of the survey when they fail on one or several IER indicators

(W9) I fully compensate IER subjects

General

Finally, these were the survey's general questions (*verbatim*):

(G1) How many empirical studies in any area of research (not only experimental philosophy) do you contribute to per year?

(G2) How many experimental philosophy studies do you contribute to per year?

(G3) For how many years have you contributed to conducting empirical studies in any area of research (not only experimental philosophy)?

(G4) For how many years have you contributed to conducting experimental philosophy studies?

(G5) To what extent do you think IER impacts the findings of your own experimental philosophy studies?

(G6) To what extent do you think IER impacts the findings of experimental philosophy studies in general (not only your own studies)?

(G7) In your estimate, what proportion of subjects in your experimental philosophy studies engage in IER throughout at least a *quarter* of the tasks that they complete?

(G8) In your estimate, what proportion of subjects in your experimental philosophy studies engage in IER throughout at least *half* of the tasks that they complete?

(G9) Compared to survey research in other disciplines, do you think that IER is a bigger or smaller problem in experimental philosophy?

(G10) Which of the following practices have you been familiar with before you were invited to participate in this survey (whether or not you have ever used them)? [list of IER prevention strategies]

(G11) Which of the following practices have you been familiar with before you were invited to participate in this survey (whether or not you have ever used them)? [list of IER detection strategies]

(G12) Before you were invited to participate in this study, could you have explained the difference between an instructional manipulation check, and instructed-response item and a bogus/infrequency item (three kinds of attention checks).

(G13) How many scholarly articles that primarily address IER (e.g., how to detect or respond to it) have you read so far?

(G14) How often have you explicitly discussed IER (e.g., how to detect or respond to it) with your colleagues so far?

(G15) How much has your knowledge and consideration of IER increased over your career as an empirical researcher?

(G16) How much do you think experimental philosophy would benefit from researchers becoming more knowledgable about IER?

(G17) Do you think you have sufficiently addressed IER in your experimental philosophy research (in terms of preventing, detecting and dealing with it)?

Both the order of the first three categories (preventing/detecting/dealing with) and the order of all individual items were randomized. At the end of each survey page subjects could provide comments on the questions and the answers that they had given to them. To test a number of hypotheses about what predicts IER attitudes (see next section) I created five scales. These scales reflect subjects' experience in conducting (experimental philosophy) studies (G1-G4), their knowledge of IER (G10-G14), their estimates of IER's effect on experimental philosophy studies (G5-G9, G16, G17), their usage of IER prevention strategies (P1-P15), their usage of IER detection strategies (D1-D17), and their usage of the most appropriate IER prevention and detection strategies (P4, P5, P8, P9, P11, P14, P15, D1-D7, D9, D10, D15, D17).

Results

The following three charts present how the survey's participants have so far attempted to prevent, detect, and deal with IER in their experimental philosophy studies.

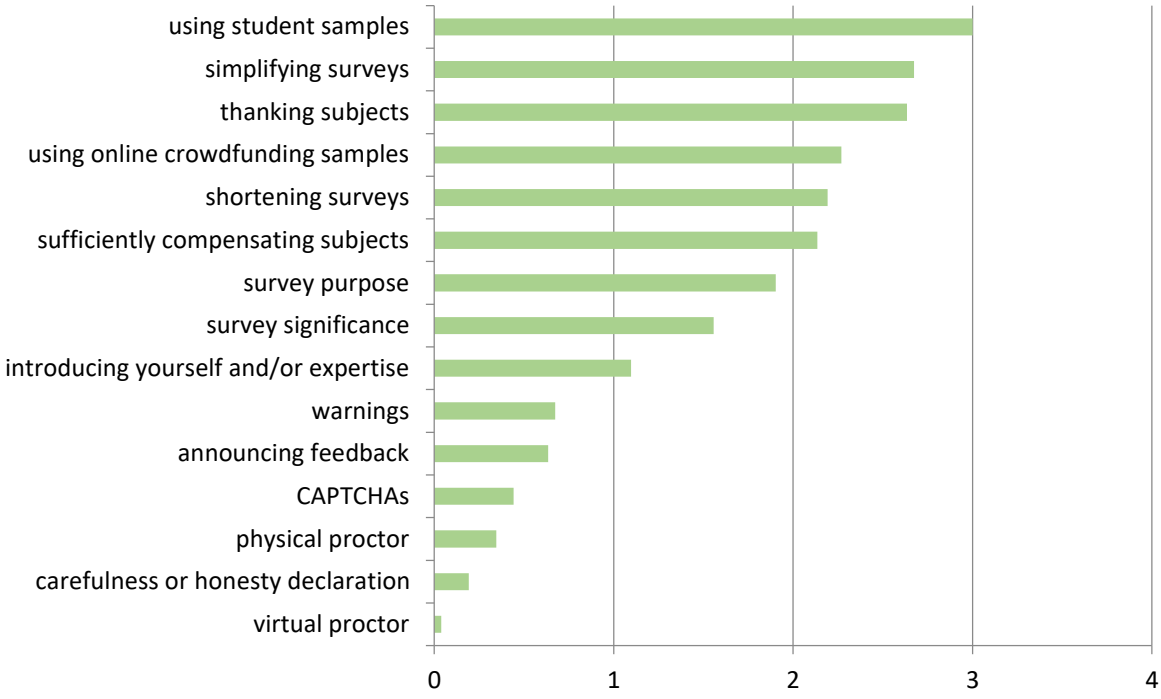

Figure 1: Usage of IER prevention strategies in experimental philosophy studies: 0 = Never, 1 = Sometimes, 2 = About half the time, 3 = Most of the time, 4 = Always

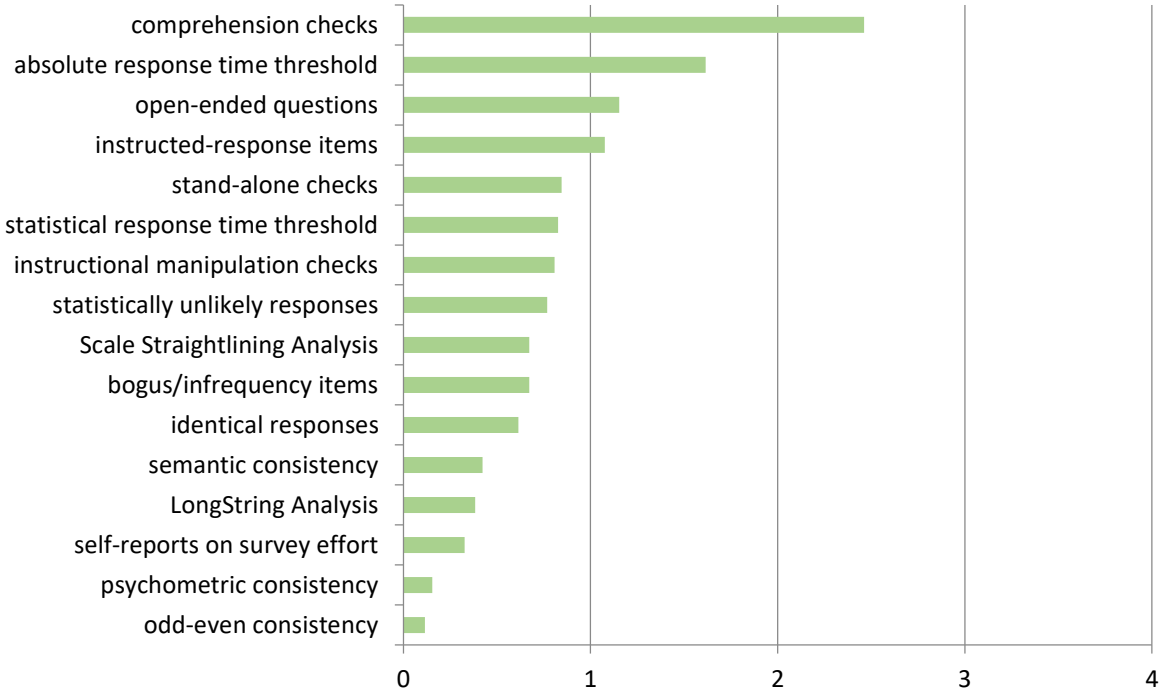

Figure 2: Usage of IER detection strategies in experimental philosophy studies: 0 = Never, 1 = Sometimes, 2 = About half the time, 3 = Most of the time, 4 = Always

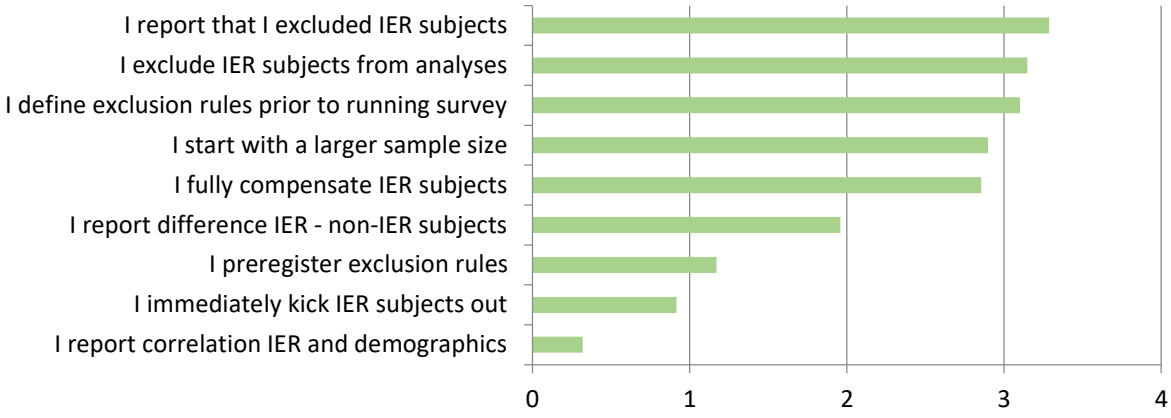

Figure 3: Dealing with IER in experimental philosophy studies: 0 = Never, 1 = Sometimes, 2 = About half the time, 3 = Most of the time, 4 = Always[35]

Experience in conducting (experimental philosophy) studies

Participants stated that they contribute to 2.67 empirical studies a year, most of which (2.13) are experimental philosophy studies. They have been engaged in empirical research for 2.85 years, and in experimental philosophy research for 2,54 years. In their perception, their knowledge and consideration of IER has "moderately" increased over their career as a researcher.

Knowledge of IER

All participants indicated that they have been familiar with at least some IER prevention and some IER detection strategies prior to completing the survey; most notably, with thanking subjects and explaining surveys' significance and purpose (regarding prevention), and with comprehension checks, attention checks and response time measures (regarding detection). They also stated that they have explicitly discussed IER with their colleagues several times (M = 3.44), and have read some scholarly articles that address it (M = 2.02).

Effects of IER

In the participants' view, IER only "somewhat" or "lowly" impacts the findings of their own and others' experimental philosophy studies — to the same extent to which it impacts self-report surveys in other areas of research. They believe that they "might" or "probably" have

[35] At first sight it may seem paradoxical that participants declared to more often *report* the exclusion of IER subjects than to *actually* exclude such subjects. A plausible explanation of this result is that the reporting item was read in a conditional sense: *If* I exclude IER *then* I report this exclusion.

sufficiently addressed IER in their own studies. At the same time, however, participants also estimated that 21.71% of subjects in their experimental philosophy studies engage in IER throughout at least a quarter of the tasks that they complete, and 15.53% throughout at least half of these tasks; and they stated that experimental philosophy would benefit "moderately" to "a lot" from researchers becoming more knowledgeable about IER.

Prior to conducting the survey I formulated several hypotheses about what predicts experimental philosophers' knowledge, usage, proper usage, and effect estimates regarding IER. Most importantly, I was interested in the impact of affiliation and academic training.

It turned out that experimental philosophers with a primary or secondary affiliation in psychology, cognitive science or some other science had more knowledge about IER ($t(50) = -2.807$, $p = 0.007$), used IER prevention and detection strategies more often ($t(50) = -1.710$, $p = 0.094$), and rated IER's impact on experimental philosophy studies to be lower than experimental philosophers who are only affiliated with philosophy or other humanities departments ($t(50) = 1.916$, $p = 0.061$). Experimental philosophers also tended to have more knowledge about IER if they held a BA, MA or PhD in psychology, cognitive science or some other science ($t(50) = -2.657$, $p = 0.011$).

	Knowledge M (SD)	Usage M (SD)	Proper Usage M (SD)	Est. Effect M (SD)
Psychology, Cognitive Science or other Science Affiliation/s	4,77** (1,30)	2,21* (0,38)	2,67 (0,51)	2,51* (0,34)
Only Philosophy or other Humanities Affiliation/s	3,72 (1,41)	2,03 (0,39)	2,43 (0,57)	2,67 (0,25)
Psychology, Cognitive Science or other Science Degree/s	4,69** (1,45)	2,13 (0,44)	2,58 (0,57)	2,55 (0,34)
Only Philosophy or other Humanities Degree/s	3,68 (1,25)	1,11 (0,32)	2,52 (0,54)	2,65 (0,25)

Table 2: Effect of being affiliated with a psychology, cognitive science or other science department (primarily or secondarily) and of having a degree in any of these fields (BA, MA or PhD) on experimental philosophers' knowledge about IER, their usage of IER prevention and detection strategies, their proper usage of these strategies, and their estimates of IER's effect on experimental philosophy studies. *p < 0.05 **p < 0.01

Interestingly, and to be further discussed below, by themselves neither affiliation nor academic training predicted whether an experimental philosopher *properly* addresses IER (in the way suggested in this paper's main part). The variable that correlated most strongly with proper usage was knowledge about IER as such, as measured in terms of the prevention and detection strategies that an experimental philosopher is familiar with, the number of articles that he or she has read about it, and the number of discussions that he or she has had about it (r = 0.643, p = 0.000). Proper usage also correlated with research experience, albeit more weakly (r = 0.337, p = 0.015).

Finally, the exclusion of the survey's insufficient effort responder (as detected by response time measures, attention checks, and scale straightlining analysis) did not have any notable effect on the results of the above analyses, except slightly decreasing power.[36]

Discussion

As explained in the paper's main part, research suggests that IER can be a serious threat to self-report surveys' validity and reliability (e.g., Curran 2016; Huang et al. 2012, 2015; Meade and Craig 2012). Experimental philosophers appear to be somewhat aware of this threat. They have some knowledge about IER, they believe that it has at least some impact on their studies' results, and they sometimes use some strategies for preventing and detecting it. That said, the survey also suggests that IER's effects might still be somewhat underestimated by or unclear to experimental philosophers (especially given their divergent and sometimes intrapersonally inconsistent explicit estimates); and just like researchers in other areas, they have not always addressed it in the most appropriate ways.

As to the prevention of IER, experimental philosophers often use some of the strategies that were recommended in Sec. 3. Two such strategies, however, have so far widely been neglected. Experimental philosophers have only rarely included carefulness/honesty declarations and CAPTCHAs into their surveys. This is particularly noteworthy in the case of CAPTCHAs, as they are a very simple and effective tool to prevent some of the worst IER, namely responses from bots. Finally, the results of the survey's prevention part also reflect a common misunderstanding about the relative effort exerted by different kinds of subjects. Experimental philosophers have most often attempted to prevent IER by using student samples — even

[36] All significant effects also turned out to be significant with the IER subject included, with very similar levels of significance, effect sizes and correlation coefficients. No additional effects crossed the threshold of significance.

though, as argued in Sec. 3.1, students are no less and perhaps even more prone to such behavior than subjects from crowdsourcing platforms.

The list of experimental philosophers' most-used IER detection practices is topped by comprehension checks, response time measures and attention checks. If implemented properly these strategies are indeed reliable and effective. Differing from my recommendations, experimental philosophers have so far hardly drawn on presenting subjects identical questions or items; and they have rarely used (and mostly are not even familiar with) post-hoc analyses such as LongString Analysis and Scale Straightlining. The latter is in stark contrast to how psychologists attempt to detect IER. For example, in a similar survey the members of the *Society for Industrial and Organizational Psychology* have stated that they use post-hoc analyses more regularly than any other IER detection strategy (Bowling et al. 2013).[37]

If a subject turns out to have engaged in IER experimental philosophers typically exclude this subject from (some) analyses and report such exclusions in their paper — but not always. This result is puzzling. Why would a researcher take the effort of detecting IER subjects but then include them in all analyses? The only plausible answer that I can think of is that such decisions are made when the prevalence of IER is found to be very low. In any case, it would be good for experimental philosophers to *always* exclude IER subjects and to *always* report these exclusions, as well as to preregister their exclusion rules. To preserve the integrity of subject pools like MTurk the practice of fully compensating IER subjects (employed "most of the time") should be rethought too.

The (perhaps obvious) main result of my statistical analysis is that the more experimental philosophers know about IER the more likely they are to properly address it in their studies. This knowledge sometimes arises from a degree in psychology, cognitive science or some other science. Thus, if controlled for knowledge, such degrees turn out to predict that an experimental philosopher uses the most recommendable IER prevention and detection strategies, and deals with IER subjects in the most appropriate ways ($B = 0.271$, $p = 0.09$; when knowledge is entered as a covariate in an ANOVA degree predicts proper usage too). But knowledge about IER can of course be acquired in different ways too (such as by reading or having discussions about it). In line with this article's motivation, it is therefore important to make experimental philosophers more aware of and encourage them to learn about IER.

[37] Very likely, this difference is at least partly explained by the fact that industrial and organizational psychologists use long lists of items (which are prone to patterned responses), while experimental philosophers typically present their subjects with a small number of more detailed cases.

The results of this survey must be taken with a grain of salt. Its sample size is small. Moreover, it presupposes that participants correctly recalled, synthesized and reported information about a number of studies and other events that may date back years. Hopefully, however, the survey nevertheless provided a first rough glimpse into experimental philosophers' knowledge, consideration and assessment of IER.

References

Andow, James (2016): Qualitative tools and experimental philosophy. *Philosophical Psychology*, 29 (8), 1128-1141.

Aker, A.; El-Haj, M.; Albakour, M.-D.; & Kruschwitz, U. (2012): Assessing crowdsourcing quality through objective tasks. Paper presented at the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey.

Baer, R. A.; Ballenger, J.; Berry, D. T. R.; Wetter, M. W. (1997): Detection of random responding on the MMPI-A. *Journal of Personality Assessment* 68 (1), 139-151.

Beach, D. A. (1989): Identifying the random responder. *The Journal of Psychology* 123 (1), 101-103.

Berinsky, A. J.; Margolis, M. F.; Sances, M. W. (2013): Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* 58 (3), 1-15.

Berry, D. T. R.; Wetter, M. W.; Baer, R. A.; Widiger, T. A.; Sumpter, J. C.; Reynolds, S. K.; Hallam, R. A. (1991): Detection of random responding on the MMPI-2: Utility of F, back F, and VRIN scales. *Psychological Assessment: A Journal of Consulting and Clinical Psychology* 3 (3), 418-423.

Breitsohl, Heiko; Steidelmüller, Corinna (2018): The impact of insufficient effort responding detection methods on substantive responses: Results from an experiment testing parameter invariance. *Applied Psychology* 67 (2), 284-308.

Buechley, R.; Ball, H. (1952): A new test of "validity" for the group MMPI. *Journal of Consulting Psychology* 16 (4), 299-301.

Buhrmester, M.; Kwang, T.; Gosling, S. D. (2011): Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6 (1), 3-5.

Chandler, D.; Kapelner, A. (2013): Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90, 123-133.

Charter, R. A. (1994): Determining random responding for the Category, Speech-Sounds Perception, and Seashore Rhythm tests. *Journal of Clinical and Experimental Neuropsychology* 16 (5), 744-748.

Clark, M. E.; Gironda, R. J.; Young, R. W. (2003): Detection of back random responding: Effectiveness of MMPI-2 and personality assessment inventory validity indices. *Psychological Assessment* 15 (2), 223-234.

Costa, P. T.; McCrae, R. (1992): *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Model (NEO-FFI) Professional manual*. Odesa: Psychological Assesment Center.

Cova, Florian; Strickland, Brent; et al. (2018): Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*.

Curran, P. G. (2016): Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* 66, 4-19.

Curran, P.; Kotrba, L.; Denison, D. (2010): Careless responding in surveys: Applying traditional techniques to organizational settings. Poster presented at the 25th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Dennis, S. A.; Goodson, Brian M.; Pearson, C. (2018): MTurk workers' use of low-cost "virtual private servers" to circumvent screening methods: A research note. Working paper.

Desimone, J. A.; Harms, P. D.; Desimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior* 36 (2), 171-181.

Ehlers, C.; Greene-Shortridge, T. M.; Weekley, J. A.; Zajack, M. D. (2009): The exploration of statistical methods in detecting random responding. Poster presented at the annual meeting for the Society for Industrial/Organizational Psychology, Atlanta, GA.

Francavilla, N. M.; Meade, A. W.; Young A. L. (forthcoming): Social interaction and internet-based surveys: Examining the effects of virtual and in-person proctors on careless response.

Francis, Kathryn B.; Beaman, Philip C.; Hansen, Ned (forthcoming): Stakes, Scales, and Skepticism. *Manuscript under review.*

Funke, F.; Reips, U. D.; Randall K. T. (2011): Sliders for the smart: Type of rating scale on the Web interacts with educational level. *Social Science Computer Review* 29 (2), 221-231.

Gage, N. L.; Leavitt, G. S.; Stone, G. C. (1957): The psychological meaning of acquiescence set for authoritarianism. *The Journal of Abnormal and Social Psychology* 55 (1), 98-103.

Goldberg, L. R.; Kilkowski, J. M. (1985): The prediction of semantic consistency in self-descriptions: characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology* 48 (1), 82-98.

Goodman, J. K.; Cryder, C. E.; Cheema, A. (2013): Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making* 26 (3), 213-224.

Gummer, T.; Roßmann, J. (2015): Explaining interview duration in web surveys: A multilevel approach. *Social Science Computer Review* 33 (2), 217-234.

Gummer, T.; Roßmann, J.; Silber, H. (forthcoming): Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*.

Harms, C.; Jackel, L.; Montag, C. (2017): Reliability and completion speed in online questionnaires under consideration of personality. *Personality and Individual Differences* 111, 281-290.

Hauser, D. J.; Schwarz, N. (2016): Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods* 48 (1), 400-407.

Heerwegh, D.; Loosveldt, G. (2006): An experimental study on the effects of personalization, survey length statements, progress indicators, and survey sponsor logos in web surveys. *Journal of Official Statistics* 22 (2), 191-210.

Holbrook, A. L.; Krosnick, J. A.; Moore, D.; Tourangeau, R. (2007): Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. *Public Opinion Quarterly* 70 (4), 325-348.

Horn, R. G.; Karim, M. N.; Behrend, T. S.; Sharek, D. J.; Wiebe, E. N. (2013): Mechanical Turk: Compensation rate and data quality. Poster presented at the 25th annual meeting of the Association of Psychological Science, Washington, DC.

Huang, J. L. ; Curran, P. G. ; Keeney, J.; Poposki, E. M.; DeShon, R. P. (2012): Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology* 27 (1), 99-114.

Huang, J. L.; Bowling, N. A.; Liu, M.; Li, Y. (2015): Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology* 30 (2), 299-311.

Johnson, J. A. (2005): Ascertaining the validity of individual protocols from webbased personality inventories. *Journal of Research in Personality* 39 (1), 103-129.

Kam, C. S. C.; Chan, G. H. (2018): Examination of the validity of instructed response items in identifying careless respondents. *Personality and Individual Differences* 129, 83-87.

Kauppinen, A. (2007): The rise and fall of experimental philosophy. *Philosophical Explorations* 10 (2), 95-118.

Klein, R. A.; Ratliff, K. A.; Vianello, M.; Adams Jr, R. B.; et al. (2014) : Investigating variation in replicability. *Social Psychology* 45, 142-152.

Knobe, J. (2003): Intentional action and side effects in ordinary language. *Analysis* 63 (3), 190-193.

Knobe, J. (2016): Experimental philosophy is cognitive science. In: Sytsma, J.; Buckwalter, W. (eds.): *A Companion to Experimental Philosophy*. Oxford: Blackwell, 37-52.

Knobe, J. ; Nichols, S. (2017): Experimental philosophy. In: Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/experimental-philosophy/.

Knobe, J.; Nichols, S. (2007): An experimental philosophy manifesto. In: Knobe, J.; Nichols, S. (eds.): *Experimental Philosophy*. Oxford: Oxford University Press, 3-14.

Krosnick, J. A. (1991): Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5 (3), 213-236.

Kurtz, J. E.; Parrish, C. L. (2001): Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment* 76 (2), 315-332.

Leiner, D. J. (2016): Too fast, too straight, too weird: Post-hoc identification of meaningless data in internet surveys. Working Paper.

Litman, L.; Robinson, J.; Rosenzweig, C. (2015): The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods* 47 (2), 519-528.

Liu, M.; Bowling, N. A.; Huang, J. L.; Kent, T. A. (2013): Insufficient effort responding to surveys as a threat to validity: The perceptions and practices of SIOP members. *TIP: The Industrial-Organizational Psychologist* 51 (1), 32-38.

Landers, Richard N. (2016): *Calculating LongString in Excel to Detect Careless Responders. NeoAcademic*. http://neoacademic.com/2016/12/21/calculating-longstring-excel-detect-careless-responders/.

Machery, E.; Stich, S. P.; Rose, D.; Chatterjee, A.; Karasawa, K.; Struchiner, N.; Sirker, S.; Usui, N.; Hashimoto, T. (2017): Gettier across cultures. *Noûs* 51 (3), 645-664.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India* 12, 49-55.

Maniaci, M. R.; Rogge, R. D. (2014): Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality* 48 (1), 61-83.

Matjasic, M.; Vehovar, V.; Lozar, K. M. (2018): Web survey paradata on response time outliers: A systematic literature review. *Metodološki zvezki* 15 (1), 23-41.

McGrath, R. E.; Mitchell, M.; Kim, B. H.; Hough, L. (2010): Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin* 136 (3), 450-470.

Meade, A. W.; Craig, S. B. (2012): Identifying careless responses in survey data. *Psychological Methods* 17 (3), 437-455.

Open Science Collaboration (2015): Estimating the reproducibility of psychological science. *Science* 349 (6251), 1-8.

Oppenheimer, D. M.; Meyvis, T.; Davidenko, N. (2009): Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45, 867-872.

Osborn, J. W.; Blanchard, M. R. (2011): Random responding from participants is a threat to the validity of social science research results. Frontiers in Psychology 1, 1-7.

Paolacci, G.; Chandler, J.; Ipeirotis, P. G. (2010): Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5 (5), 411-419.

Peer, E.; Vosgerau, J.; & Acquisti, A. (2014): Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* 46 (4), 1023-1031.

Pennycook, G.; Trippas, D.; Handley, S. J.; Thompson, V.A. (2014): Base rates: Both intuitive and neglected. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 40, 544-554.

Pölzler, Thomas (2018): *Moral Reality and the Empirical Sciences*. New York: Routledge.

Pölzler, Thomas; Wright, Jennifer C. (2020a): An Empirical Argument against Moral Non-Cognitivism. *Inquiry*.

Pölzler, Thomas; Wright, Jennifer C. (2020b): Anti-Realist Pluralism: A New Approach to Folk Metaethics. *Review of Philosophy and Psychology* 11 (1), 53-82.

Roberts, Pendaran; Andow, James; Schmidtke, Kelly A. (2018): Lay intuitions about epistemic normativity. *Synthese* 195 (7), 3267-3287.

Rose, D.; Danks, D. (2013): In defense of a broad conception of experimental philosophy. *Metaphilosophy* 44 (4): 512-532.

Rose, D.; Machery, E.; Stich, S. P.; et al. (forthcoming): Nothing at stake in knowledge. *Noûs*.

Stokel-Walker, C. (2018): Bots on Amazon's Mechanical Turk are ruining psychology studies. *NewScientist* August, 10.

Taylor, S. E. (1965): Eye movements in reading: Facts and fallacies. *American Educational Research Journal* 2 (4), 187-202.

Thomas, K. A.; Clifford, S. (2017): Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior* 77, 184-197.

Tourangeau, R.; Rips, L. J.; Rasinski, K. A. (2009): *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Vannette, D. (2016): Why using grid questions is probably hurting your data. https://www.qualtrics.com/blog/why-using-grid-questions-is-probably-hurting-your-data/.

Ward, M. K.; Meade A. W. (2018): Applying social psychology to prevent careless responding during online surveys. *Applied Psychology* 67 (2), 231-263.

Ward, M. K.; Pond, S. B. (2015): Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior* 48, 554-568.

Weinberg, J. M.; Nichols, S.; Stich, S. (2001): Normativity and epistemic intuitions. *Philosophical Topics* 29 (1/2), 429-460.

Woods, C. (2006): CR to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment* 28 (3), 186-191.

Zhang, C.; Conrad, F. G. (2014): Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Method* 8 (2), 127-135.

Acknowledgments

This work was funded by the Austrian Science Fund (FWF) under research grant J 4163-G24. My early thinking on insufficient effort responding was shaped by helpful discussions with Jen Wright and Josh Knobe. Jen also provided invaluable support in conducting and analyzing the survey. For helpful comments on previous versions of the manuscript or survey I would like to thank Johannes Wagner, Norbert Paulo, Lieuwe Zijlstra and two anonymous reviewers for Oxford Studies in Experimental Philosophy. Kathryn Francis, James Andow and John Park kindly provided me with evidence about insufficient effort responding in some of their recent experimental philosophy studies. Finally, thanks to all those experimental philosophers who took the time to complete my survey.