Saving the armchair by experiment: what works in economics doesn't work in philosophy

Boudewijn de Bruin^{1,2}

Accepted: 21 September 2020 / Published online: 7 October 2020
© The Author(s) 2020

Abstract Financial incentives, learning (feedback and repetition), group consultation, and increased experimental control are among the experimental techniques economists have successfully used to deflect the *behavioral challenge* posed by research conducted by such scholars as Tversky and Kahneman. These techniques save the economic armchair to the extent that they align laypeople judgments with economic theory by increasing cognitive effort and reflection in experimental subjects. It is natural to hypothesize that a similar strategy might work to address the *experimental* or *restrictionist challenge* to armchair philosophy. To test this hypothesis, a randomized controlled experiment was carried out (for incentives and learning), as well as two lab experiments (for group consultation, and for experimental control). Three types of knowledge attribution tasks were used (Gettier cases, false belief cases, and cases in which there is knowledge on the consensus/orthodox understanding). No support for the hypothesis was found. The paper describes the close similarities between the economist's response to the behavioral challenge, and the expertise defense against the experimental challenge, and presents the experiments, results, and an array of robustness checks. The upshot is that these results make the experimental challenge all the more forceful.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11098-020-01559-z>) contains supplementary material, which is available to authorized users.

✉ Boudewijn de Bruin
b.p.de.bruin@rug.nl

¹ Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands

² Economics, University of Groningen, Nettelbosje 2, 9747 AE Groningen, The Netherlands

Keywords Experimental control · Experimental philosophy · Expertise defense · Gettier · Group consultation · Incentives · Knowledge and belief · Learning · Randomized controlled trial (RCT)

1 Introduction

Do armchair philosophers have an *experimental* argument against experimental philosophers? One could argue that we shouldn't hope for something of that sort: a wealth of experiments demonstrate that not only laypeople judgments, but also those of philosophers are prone to be influenced by irrelevant perturbing factors, and that in fact expert judgments are sometimes plainly wrong, for even professional epistemologists tend to attribute knowledge to the protagonist of fake barn type Gettier cases (Horvath and Wiegmann 2016), to give only one example. How likely then is it that new experiments would generate empirical evidence backing philosophical expertise?

The starting point of this paper is that one experimental strategy has not yet been deployed. To see what strategy that is, consider how neo-classical economists responded to the *behavioral challenge*, which was launched by Herbert Simon—who used the word “armchair” to lambaste the economic methodology of old long before it became popular in philosophy (Simon and Bartel 1986).

In a famous experiment, two other pioneers of behavioral economics, Amos Tversky and Daniel Kahneman (1983), asked subjects to consider Linda:

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Tversky and Kahneman asked the folk which proposition they believe is more probable: (i) that Linda is a bank teller, or (ii) that Linda is a bank teller who is active in the feminist movement. Now surely if you've learned something about probability, you know that the right answer is (i). But 85% of the subjects in the experiment gave (ii) as an answer.

Probability and utility theory form the unifying theoretical core of economics, and hence empirical evidence showing that people diverge from theory constitutes a clear challenge to the field, just like the experimental challenge to philosophy.

Economists responded to the challenge with arguments that are strikingly similar to the expertise and reflection defense that philosophers have marshaled against the *experimental challenge* (or *restrictionist challenge* (Alexander and Weinberg 2007)). Economists argued that the subjects in Tversky and Kahneman's experiment (and other behavioral experiments) lacked experience, were not reflecting on their answers, likely missed essential features of the vignettes they were confronted with, and so simply didn't understand the question. If, the economists maintained, experimental subjects could be made to pay attention to the relevant details of the scenarios, reflect on them, and come to understand what was being asked of them, then they would naturally answer correctly.

That is exactly what they set out to test empirically. Economists used a repertoire of experimental techniques to encourage participants to pay attention and reflect, and to make them understand what the experiments are about: they told participants that they would pay out a financial bonus if they answered the question correctly (that is, they used *financial incentives*); they familiarized participants with the type of question prior to the experiment (so they introduced an element of *learning* in the experiment); they allowed participants to consult with each other about the question before answering (*group consultation*); and they made sure that the experiment took place in controlled conditions in a lab, rather than online where all sorts of things may distract participants from the task (*experimental control*).

These techniques didn't reduce error rates to zero, but altogether the effects were quite impressive. For instance, when participants were offered a bonus of \$4 for a correct answer to the question about Linda, only 33% chose (ii).

The thought behind this paper was that what had saved the economic armchair should save the philosophical armchair just as much. That is, I expected that with standard experimental techniques from economics, laypeople judgments would converge to those of professional philosophers. To test this *convergence hypothesis*, I ran a number of knowledge attribution experiments (including Gettier cases), testing the effects of financial incentives and learning (randomized controlled trial, RCT, total sample size 1443), group consultation (sample size 661), and experimental control (RCT, sample size 191).

Frankly, I expected that this would yield a response to the experimental philosophy challenge just as forceful as the economist's response to the behavioral challenge. What I found was entirely different, though. Financial incentives, learning, group consultation, and experimental control didn't bring laypeople judgments closer to those of philosophers, and so what looked as a promising strategy to restore confidence in armchair philosophy has made the experimental challenge stronger—at least, it is stronger than the behavioral challenge to economic orthodoxy.

The remainder of the paper is structured as follows. Section 2 provides some background about the experimental challenge and ways to deflect it as well as about the behavioral challenge and the economist's response. Section 3 describes the vignettes as well as the regression model used to estimate the hypothesized effect. Sections 4–8 introduce and motivate the experimental techniques (incentivization, learning (one short and without incentives, the other longer and with incentives), group consultation, and experimental control), and provide evidence that these techniques are *prima facie* plausible tools to meet the experimental challenge. These sections also discuss the results of the experiments. Section 9 compares the results with those of others (among others, I consider an array of covariates such as age, gender, education, etc.), anticipates a number of objections, points out various limitations of the study, and makes some suggestions for future research.

Before proceeding, it is important to note the following. I have decided to present the results of all five experiments in one paper rather than in several separate papers because I think that this makes the argument stronger and more coherent. The downside is that this paper would have become very long had I reported all usual information about each experiment (materials, methods, procedures, data, models,

results, etc.). That's why I have chosen to focus on the general argument in the main paper, and defer these details to supplementary materials, which together with the data are available online.

2 Background

A wealth of survey research attests to the fact that laypeople often express different opinions about philosophical cases than professionals. Consider, for instance, the vignette used in all experiments reported here. It is due to Starmans and Friedman (2012). Participants were asked whether (i) Peter only believes that there is a watch on the table, or (ii) Peter really knows that there is a watch on the table.

Peter is in his locked apartment reading, and is about to have a shower. He puts his book down on the coffee table, and takes off his black plastic watch and leaves it on the coffee table. Then he goes into the bathroom. As Peter's shower begins, a burglar silently breaks into the apartment. The burglar takes Peter's black plastic watch, replaces it with an identical black plastic watch, and then leaves. Peter is still in the shower, and did not hear anything.

Now surely if you've learned something about epistemology, you know that the consensus answer is (i). But 72% of the subjects in Starmans and Friedman's (2012) experiment gave (ii) as an answer—a finding I replicate. The experimental challenge stares us starkly in the face.

Several strategies have been deployed to fend off the challenge. Some authors have pointed to weaknesses in the way the experiments were designed, carried out, and statistically analyzed. Woolfolk (2013), for instance, arraigns the use of self-reported data and complains about statistical errors in published experimental philosophy research. But as self-reported data are part and parcel of much of psychology and economics, and experimental and statistical sophistication among philosophers is growing, this strategy isn't too effective in the long run. That's why a more prominent line of defense that armchair philosophers put up against the experimentalists accepts the empirical results largely as given.

This line has come to be known as the *expertise defense*.¹ The thought is that the data experimental philosophers have gathered do not jeopardize the status of more traditional philosophical practice because they concern judgments of people who

¹ The expertise defense is developed by such authors as Bach (2019), Deutsch (2009), Devitt (2011), Egler and Ross (2020), Grundmann (2010), Hales (2006), Hofmann (2010), Horvath (2010), Iriekfe (2020), Kauppinen (2007), Ludwig (2007, 2010), Sosa (2007, 2009) and Williamson (2007, 2009, 2011). Also see Alexander (2016), Horvath and Wiegmann (2016), Nado (2014a, 2014b), Seyedsayamdost (2019) and Williamson (2016). The expertise defense was challenged on conceptual and empirical grounds by, among others, Buckwalter (2016), Clarke (2013), Hitchcock (2012), Machery (2011), Machery et al. (2013), Mizrahi (2015), Ryberg (2013), Schulz et al. (2011) and Weinberg et al. (2010). Of particular relevance to the present paper: Weinberg et al. (2010) called upon philosophers to develop and test empirical hypotheses about philosophers' alleged competence in thought experimentation, but most subsequent experimenting has not led to anything near a defense of the armchair position. See, e.g., Horvath and Wiegmann (2016) (but also for empirical evidence showing that philosopher intuitions are in a sense better than those of laypeople), Schulz et al. (2011), Schwitzgebel (2009), Schwitzgebel and

have no philosophical expertise. We should expect laypeople to diverge just as much from experts if we asked them questions about mathematics (or physics, or law, and so on); yet no sensible mathematician would find this worrying, or even faintly intriguing, and simply explain the divergence by the training they, and not laypeople, have had. The main idea of the expertise defense is then that philosophers, too, have a specific kind of expertise that justifies our setting aside laypeople judgments, just like the mathematicians (physicists, lawyers, etc.) ignore the folk. Unlike the folk, we have learned “how to apply general concepts to specific examples with careful attention to the relevant subtleties,” as Williamson (2007, 191) has put it, and this makes us far better at answering questions about Gettier cases and their ilk.

Advocates of the expertise defense often point up an additional difference between professional judgments and the data of the experimentalists: whereas laypeople only give their spontaneous and thoughtless opinions about the vignettes they are confronted with, professional philosophers engage in painstaking reflection and deliberation and put in great cognitive effort with close attention to details before reaching judgment. Ludwig (2007, 148) likens this type of reflection to what is needed to grasp mathematical truths: to see that one can construct a bijection between the set of numbers and the set of odd numbers, he claims, “a quick judgment is not called for, but rather a considered reflective judgment in which the basis for a correct answer is revealed.” Our experience with this type of reflection—for Sosa (2007) and Kauppinen (2007) joined with dialogue and discussion—is what grounds our authority on matters philosophical. As a result, there is no merit whatsoever in “questioning untutored subjects” (Ludwig 2007, 151), who “may have an imperfect grasp of the concept in question [and] rush in their judgements” (Kauppinen 2007, 105) with their “characteristic sloppiness” (Williamson 2007, 191).

Here is my key observation (again): the expertise defense bears more than superficial similarity to the way economists have responded to the behavioral challenge. Take Ken Binmore, a prominent economist:

But how much attention should we pay to experiments that tell us how *inexperienced* people behave when placed in situations with which they are *unfamiliar*, and in which the incentives for *thinking things through carefully* are negligible or absent altogether? (1994, 184–185, my emphasis)

Yet, while the philosophical debate has rumbled on with characteristic armchair methods, the economists were quick to finish their debate by bluntly counter-challenging behavioral economics on *empirical* grounds. Binmore asked:

Footnote 1 continued

Cushman (2012), Tobia et al. (2013) and Vaesen et al. (2013). Also see Drożdżowicz (2018), Hansson (2020) and Seyedsayamdost (2019).

Does the behavior survive when the incentives are increased? (1994, 185)

So he asked: if we pay people to pay attention, do they still deviate from economic theory?²

Binmore also asked:

Does it survive after the subjects have had a long time to familiarize themselves with all the wrinkles of the unusual situation in which the experimenter has placed them? (1994, 185)

That is: what happens if we make sure that participants understand what they are being asked to do? Binmore concluded that if both questions are answered affirmatively,

the experimenter has probably done no more than inadvertently to trigger a response in the subjects that is adapted to some real-life situation, but which bears only a superficial resemblance to the problem the subjects are really facing in the laboratory. (1994, 185)

So if the deviation from theory disappears when *financial incentives* and *learning* are introduced, then the relevance of the behavioral challenge is probably negligible.

Such reasoning is, I think, very attractive. Just as anyone paying attention to the Linda the Bank Teller task should be able to give the right answer, a person sufficiently alert and engaged should be expected to give the consensus answer to a Gettier task. Also: the average advocate of the expertise defense should hold that knowledge attribution cases are very much of a piece with Linda the Bank Teller (and other cases such as the gambler's fallacy, etc.).³ And if the average advocate of the expertise defense agrees that no economist should be worried by what laypeople think about such cases, they might conclude by analogy that no philosopher should be afraid of the folk Gettier holdouts.

These are then the ideas that motivated the experiments. I expected that when *financial incentives* and *learning* are introduced, as Binmore suggested, the deviation from theory disappears, and I also considered two other mechanisms of more recent date, not mentioned by Binmore: *group consultation* and *experimental control*.

I was very surprised to see this *convergence hypothesis* founder.⁴

² Binmore means performance-based payments, not appearance fees.

³ Williamson (2007) briefly discusses the Linda the Bank Teller experiment. Also see Jackson (2011) (not a critic of experimental philosophy *per se*). The gambler's fallacy is another behavioral bias often discussed in this context (Hales 2006; Hofmann 2010; Jackson 2011; Ludwig 2007).

⁴ For additional support for the convergence hypothesis, see Buckwalter (2016) and Turri (2013).

3 Vignettes and model

3.1 Vignettes

I carried out three experiments: (i) a large randomized controlled trial (RCT) to study the impact of financial incentives and learning (online UK sample, Sects. 4–6); (ii) an experiment in a large lecture hall to study group consultation (student sample, Sect. 7); and (iii) a lab experiment to study experimental control (student sample, Sect. 8).⁵

To compare results across the experiment as well as with the existing literature, I used the same vignettes in all experiments, drawn from Starmans and Friedman (2012). Participants were randomly assigned to one of three conditions: *Gettier Case*, *Non-truth Case* (where veridicality doesn't hold) and *Knowledge Case* (where there is knowledge on the textbook understanding).⁶

The Gettier Case and the Non-truth Case involve the vignette from above (Peter and the watch), where in the Non-truth Case “an identical black watch” is replaced by “a banknote.” In both cases, participants were confronted with the following two alternatives, and asked the question “Which of the following is true?”:

Peter really knows that there is a watch on the table.

Peter only believes that there is a watch on the table.

In the Knowledge Case, participants read the text from the Gettier Case, but their task involved the following two new alternatives:

Peter really knows that there is a book on the table.

Peter only believes that there is a book on the table.

In all three conditions the “really knows” and “only believes” options appeared in random order.

If laypeople follow epistemological consensus, we should expect that participants attribute knowledge to Peter only in the Knowledge Case. In the Non-truth Case, there is no watch on the table, so Peter has a false belief. Since knowledge entails truth, Peter should not be seen as possessing knowledge. In the Gettier Case, there is a watch on the table. However, it is not the watch that Peter left on the table. The burglar's actions have ensured that Peter's environment is Gettiered, which means that if subjects follow orthodoxy, they will deny knowledge to Peter. Only in the Knowledge Case does Peter possess knowledge, as there is in fact a book on the table, and the burglar's actions have not affected the book.⁷ Subsequently,

⁵ Details can be found in online supplementary materials.

⁶ The Non-truth and Knowledge Case figure as control cases, testing other conditions of knowledge. See Starmans and Friedman (2012). I renamed the cases: the *Non-truth Case* is their *False Belief* condition, and the *Knowledge Case* is their *Control* condition. Vignettes were drawn from Experiment 1A in their study, with *banknote* instead of *dollar bill* to adjust to my UK sample.

⁷ See Starmans and Friedman (2012) for a more detailed discussion of these conditions, and references to the relevant literature on the reasons backing the consensus.

participants were asked to rate their confidence in their answer to the “know/believe” question.

Before proceeding, let me report the results of the baseline treatment. In the Gettier Case, 72% of the participants attribute knowledge to Peter, which is the same percentage reported by Starmans and Friedman (2012) ($Z = -0.02$, $p = .981$, all two-tailed difference in proportions tests). I didn’t fully reproduce their results in the Non-truth Case (False Belief condition) ($Z = -3.98$, $p < .001$), where I found knowledge attribution in 44% of the participants, as opposed to 11% in Starmans and Friedman’s sample. Nor did I replicate the Knowledge Case (Control condition) ($Z = 2.07$, $p < .05$). I there found knowledge attribution in 74% of the participants, as opposed to 88% in Starmans and Friedman’s sample. Various factors may account for these differences: I used a UK sample, where Starmans and Friedman have a US sample; I use Prolific, they MTurk; and my sample is less highly educated. These replication failures do not of course affect the validity of carrying out a randomized controlled trial, because we are there only interested in the relative differences between the treatments in otherwise similar samples, and not in differences with other experiments with different samples.

3.2 Model

Knowledge attribution error is measured in two ways in the literature. One measure uses a binary variable $Corr_u$, which takes the value 1 when the answer is correct, and 0 if false. The other measure incorporates the subject’s level of confidence in the answer. Confidence was measured on a fully anchored 7-point Likert scale. I rescaled to a scale ranging from 1 to 10 in conformance with the literature.⁸ The variable $Corr_w$ then is equal to the rescaled confidence level if the answer is correct, and equal to minus the rescaled confidence level if the answer is incorrect. So in line with Starmans and Friedman (2012) I calculated the *weighted knowledge attribution* (or *ascription*) (WKA) of each participant. WKA is a number ranging from -10 to 10 , resulting from multiplying the confidence rating of a participant with their answer to the “know/believe” question (where attribution of belief receives the value -1 and knowledge 1). So $Corr_w$ is equal to WKA if “know” is the correct answer, and equal to minus WKA (i.e., $-1 \times \text{WKA}$) if “believe” is the correct answer.

The following regression model estimates the impact of the experimental techniques (financial incentives, learning, group consultation, experimental control) on correct knowledge attribution rates:

$$Y_i = \alpha + \beta X_i + \gamma_s Z_s + \varepsilon_{i,s}.$$

Here Y is the variable under study (capturing binary or weighted correct knowledge attribution), X a vector of baseline covariates, ε the error term, and γ the parameter of interest that captures the effect of the experimental techniques (Z), respectively,

⁸ I used a 7-point Likert scale in the survey, as it has some methodological advantages over the 10-point scale that Starmans and Friedman (2012) use: it offers a mid-point, and is not too fine-grained to obscure interpretation among participants (Lissitz and Green 1975).

on the outcome under study. Controls include age, income, and wealth, and dummy variables for gender, ethnicity, religion, and education. In the group consultation experiment, I conducted a difference in proportions test in addition to linear regression, following the economics literature.⁹

4 Incentives

Psychologists rarely use incentives, generally on the grounds that typical experimental subjects are “cooperative and intrinsically motivated to perform well” (Camerer 1995, 599), or also, as Thaler (1987) dryly points out, because their research budgets tend to be much smaller than those in economics departments. Experimental philosophers borrow their methods mainly from psychology (and face similar budget constraints), so it is unsurprising that they rarely use incentives.

But incentives bring advantages: they motivate participants to take the task more seriously, and thereby increase cognitive effort and reflection. Not always, of course, for incentives may lead to “choking under pressure” (Baumeister 1984); nor do incentives increase performance in tasks requiring general knowledge (“What is the capital of Honduras?”) (Yaniv and Schul 1997). But when doing well only requires thinking things through carefully, incentives usually boost success.¹⁰

Economists are primarily interested in tasks that require at least a bit of reflection, and that is why the use of incentives is the norm in economics. What about the tasks that experimental philosophers offer to their subjects? I believe that most philosophers would think that answering them correctly is more a reflection task than a general knowledge task.¹¹ Drawing from an analogy between philosophy and other fields is initially plausible, despite some criticism that has been advanced against such moves.¹² It is at least worth working from this analogy, because it leads very clearly to a further, and empirically testable, convergence hypothesis: *if*

⁹ Following what is common among economists I use and report linear regressions with binary outcome variables because of the natural interpretation of the resulting coefficients. For all three studies reported in this paper, logistic regressions do not change the findings.

¹⁰ Course credits and candy bars may also figure as incentives, but I follow economists here and solely include financial incentives. All experimental subjects use and want money, and know how to compare money; moreover, their demand for money is not easily exhausted over the course of an experiment (unlike demand for candy bars). Hence subjects are likely motivated to maximize their monetary payoff, which will lead them to pay sufficient attention to the details of the tasks they face. Lack of attention among laypeople is a recurring theme in the literature on the expertise defense. But you don’t need to accept the expertise defense to find it plausible that for laypeople judgments to be relevant, they should satisfy some minimum standard of reflection and attention (Weinberg et al. 2001, 2010). Also see Jackson (2011) on the insignificance of unreflective judgments in Linda the Bank Teller and knowledge attribution tasks.

¹¹ This is underscored by the fact that the analogy between probability and utility theoretic tasks and philosophy tasks has figured in various replies to the experimental challenge. See footnote 3.

¹² The criticism is particularly directed at the role the purported analogy between philosophy and other fields (mathematics, physics, law, etc.) plays in the expertise defense. The analogy has a slightly different function in the present paper: it primarily serves to put the relevance of particular experimental findings about laypeople judgments into relief by comparing philosophy with other disciplines.

standard experimental tools are used that bring down error rates in laypeople judgments about such tasks as Linda the Bank Teller, *then*, with such techniques, laypeople judgments about philosophy tasks should converge to those of expert philosophers as well. This shouldn't be expected to be a massive change. It may well be smaller than the almost 50% reduction of error we see in Linda the Bank Teller—and I totally admit that not all experiments in economics are so successful! But it would really be exceedingly odd if no statistically significant reduction in error rates happened at all.

These reflections led me to design an experiment in which participants were exposed to an incentivized knowledge attribution task, to be compared with a non-incentivized baseline.¹³ Participants were paid a bonus upon giving the correct answer to a knowledge attribution question. I decided to set the incentive at £5, offering about 50% more than the \$4 that brought down the error rate in the Linda the Bank Teller task, in a well-known experiment in economics conducted by Charness et al. (2010).¹⁴ That is, I made it clear upfront that participants were to gain £5 if they answered “correctly” a specifically designated question, and that they would receive this on top of the appearance fee of £0.50 (and I made sure to actually pay them). I also made clear that this was not a general knowledge type of task, by telling them: “You can find the correct answer to the question through careful thinking about the concepts involved.”

The experimental philosopher might marshal an objection to the effect that it is problematic, if not plainly unacceptable, to pay participants £5 for giving the textbook answer to a question about knowledge attribution.¹⁵ I agree that in a way the incentive structure is biased towards a particular conception of knowledge, and that I could have chosen an alternative payment scheme. I'm also aware of the fact that the wording suggests that participants should exclude the possibility that both answers may be correct or defensible, and this is indeed more inspired by positions held in the camp of the expertise defense than by those of experimental philosophers.¹⁶

Yet, my response is that, first, I borrow this approach from economics. If an experiment pays out a monetary reward if a participant concurs with expected utility theory, economists who reject that theory may object on similar grounds. These economists are a minority, but they exist: Simon's (1948) *satisficing theory* offers a compelling alternative to expected utility theory. Indeed, even in the case of Linda the Bank Teller there may be reasons to doubt the assumption that there is only one correct answer, as subjects may differ in the weights they assign to the semantic and

¹³ Incentivization is one treatment in the first RCT experiment; the two other treatments involve (short and long) learning, discussed in Sects. 5 and 6.

¹⁴ Their experiment was conducted in 2008. \$4 in 2008 corresponds to about \$4.55 or £3.40 at the time of conducting the experiments reported in this paper.

¹⁵ Following Weinberg et al. (2010, 343), e.g., you might say that I use “entrenched theory” as the benchmark in the experiments.

¹⁶ Ludwig (2007, 150), e.g., says about Kripke's Gödel/Schmidt case that the “way the thought experiment is set up, there is only one answer that is acceptable. It is not described in a way that allows any other correct response.”

pragmatic aspects of the question they are asked. Some might stress pragmatics, and think they have to determine what is maximally informative, while meeting minimal threshold justifiability (and then answer that Linda is a feminist bank teller). Others might stress semantics, and look for a maximally justifiable, if less informative, answer.¹⁷

A second way to respond is to observe that the goal of offering monetary incentives is *not* to express praise or blame for a particular answer. Incentives are introduced to increase the attention with which participants approach a question; that is, they are meant to motivate participants to think things through more carefully, and to reflect more carefully. I didn't of course inform participants about what the right answer was until the entire experiment was over. The only thing participants knew was that they would be paid £5 if they gave the correct answer to one specific question. So while I am sympathetic to the experimental philosopher's potential reluctance, I am confident the experimental design is justifiable.

Did incentives align the folk with the experts? Not at all. Regarding the results of this experiment I can be brief: offering incentives does not decrease knowledge attribution error rates (in any of the three cases: Gettier, Non-truth, Knowledge).¹⁸

So let us move on to learning.

5 Short learning

As we saw, Binmore (1994, 184) asked whether the findings at the core of the behavioral challenge to economic theory survive if subjects have had sufficient time to “familiarize themselves with all the wrinkles of the unusual situation in which the experimenter has placed them.” The answer is, in economics: not a lot survives of the behavioral challenge when participants are given the opportunity to learn prior to the experiment.

Binmore's line of reasoning foreshadowed the expertise defense. Ludwig, for instance, complains about the use of “untutored” experimental subjects (2007, 151), and maintains that subjects “may well give a response which does not answer to the purposes of the thought experiment” (2007, 138). When we philosophers conduct thought experiments, we activate our competence in deploying the concepts involved, Ludwig believes. Laypeople judgments, by contrast, will “not solely” be based in such competence (2007, 138), as laypeople may have difficulties “sorting out the various confusing factors that may be at work” when they are confronted with a scenario (2007, 149).¹⁹

This is just as the economists have it: laypeople responses are often “inadvertently triggered,” and so they have only “superficial resemblance” to the

¹⁷ I am indebted here to Williamson (2007, 96), where relevant references can be found.

¹⁸ Details on methods and results can be found in online supplementary materials.

¹⁹ Similar positions have been articulated by such authors as Hofmann (2010), Horvath (2010), Kauppinen (2007), Sosa (2007) and Williamson (2007, 2009, 2011).

problem at hand, to use Binmore's words, because laypeople are generally "inexperienced" and "unfamiliar" with the situation (1994, 184).

So what should we do? Here as above, I suggest that we confront the experimental challenge *empirically*, as the economists confront the behavioral challenge. What Binmore suggests here is to include in the experiment a *learning* phase in which subjects familiarize themselves with the tasks at hand, which consists of a sequence of practice tasks (possibly with feedback and/or explanations) preceding the *real* experiment.

There are good reasons to believe that, in the context of judgments about knowledge attribution, a phase of learning would make a larger contribution to reducing error rates than offering incentives. While the economics literature suggests that incentives increase attention and cognitive effort, one may wonder if attention and effort will be applied to the right kind of issues. It is initially plausible to think that what sets us apart from non-philosophers is that we have learned "how to apply general concepts to specific examples with careful attention to the *relevant* subtleties" (Williamson 2007, 191, my emphasis). If that is true, then incentives may misfire a bit (which may explain the results of the previous section), and genuine learning is then likely to do more.

This is what I did. I developed a short learning phase without incentives, with minimal feedback (discussed in this section), and one longer learning phase with textbook style explanations, where incentives were necessary to motivate participants to stay in the experiments for 20 minutes (discussed in the next section). My cue was not the Linda the Bank Teller experiment, but another behavioral challenge to economic orthodoxy, namely, experiments having to do with the base-rate fallacy. These experiments involve tasks requiring participants to calculate the conditional probability that you have a certain disease given that you know the outcome of a certain test. Kahneman and Tversky (1973), for instance, found that people tend to overlook the fact that to correctly determine this probability, they must also consider the unconditional probability of having that disease (the occurrence of that disease in the relevant population, or *base rate*). Most participants didn't do that.

If you consult the psychology literature, you may get the impression that Kahneman and Tversky's findings are quite stable. Yet most of these experiments are "snapshot studies" (Hertwig and Ortmann 2001, 387) as they confront participants with one case only. That is, these participants are "inexperienced," and "unfamiliar" with the "wrinkles" of the situation (Binmore 1994, 184).

What happens if participants are given the opportunity to familiarize themselves with these wrinkles? In a well-known experiment, Harrison (1994) reduced the base-rate fallacy by having participants go through a round of ten practice tasks, followed by feedback on performance. In a next round of ten tasks, they displayed the base-rate fallacy considerably less frequently.²⁰

²⁰ Harrison's (1994) paper is extremely rich in experimental detail, and covers a whole lot more than this. For expository purposes I ignore, among other things, that Harrison was particularly interested in testing Kahneman and Tversky's (1973) explanation of the base-rate fallacy in terms of the representative heuristic.

Psychologists might demur. Thaler (1987), for instance, has argued that when it comes to major life decisions, people don't have sufficiently many opportunities for learning (think of mortgage or retirement planning decisions), and that this diminishes the force of the economist's response to the behavioral challenge. But this depends on one's explanatory goals. Snapshot studies may well help you to explain household financial planning (which people do only infrequently), whereas genuine learning is perhaps more likely among professional investors (who work with finance all the time). I take no stance here, but only note that Thaler's objection does not have much force if you are interested in making laypeople subjects more reflective and more attentive to the relevant details of the cases they are confronted with.

In the *short learning* study I tested the effect of a simple type of learning. Participants were told that they would have to carry out five small tasks (each consisting of reading a small story and answering a question), plus a small survey. They were informed that after each of the first four tasks they would receive feedback on whether they answered the question correctly, and that this might help them to answer the last question correctly. The practice vignettes were drawn from earlier work in experimental philosophy.²¹ They were presented in random order. Two of the four cases had "really knows" as the textbook answer, and two "only believes." Feedback was of the form: "The correct answer is: Bob only believes that Jill drives an American car."

What about the results? They were not much different from the incentivized experiment. There is no improvement in the Gettier Case and the Non-truth Case, and only some improvement in the Knowledge Case ($p < .01$). What works in economics doesn't seem to work in philosophy.

6 Long learning

But perhaps this is too quick. Perhaps participants were not given sufficient time and opportunity to practice, and since an incentive structure was missing, they may not have felt the need to practice particularly intensely. Hence I presented another group of participants with an incentivized training regimen that is more in line with what undergraduate students may encounter in an introduction to philosophy/epistemology.

²¹ The idea that feedback on answers to thought experiments is among the mechanisms that generate philosophical expertise is widely shared by advocates of the expertise defense. Horvath (2010, 471), e.g., writes that "we have ample *prima facie* reason to believe that the relevant cognitive skills of professional philosophers have been exposed to enough clear and reliable feedback to constitute true expertise concerning intuitive evaluation of hypothetical cases." Ludwig (2007, 149) believes that training fosters "sensitivity to the structure of the concepts through reflective exercises with problems involving those concepts," and elsewhere suggests that to that end it may be useful to "run through a number of different scenarios" (2007, 155). Among experimental philosophers, Weinberg et al. (2010, 340) argue that the empirical literature on (general) expertise shows that the acquisition of expertise requires "repeated or salient successes and failures" based on clear and frequent feedback, but they question whether philosophy can ever provide such feedback, among other things because training of intuitions against "already-certified expert intuitions" would seem to be "non-starter" leading to a vicious regress.

First, they got eight instead of four practice cases with simple feedback (correct/incorrect), half of them with “really knows,” half of them with “only believes” as the epistemological consensus answer.

Secondly, after the first four of these eight practice cases, I inserted two cases (one “really knows,” the other “only believes”) with elaborate feedback (an explanation that philosophy students may be given in class), which means that after the entire learning phase, subjects will have gone through ten cases. A practical reason for this number is that I estimated that the total session would last around 20 minutes, which is about the maximum an online audience would be willing to invest. A theoretical reason was that, as far as I can see, introductory courses give the median student exposure to fewer vignettes (but of course to more theory). Moreover, ten is what Harrison (1994) uses in his work on the base-rate fallacy. As a result, this number is a defensible choice.²² All ten practice cases were selected from the experimental philosophy literature (the first four were identical to those in the short learning study).

The third element of the long learning study is that at various places during the learning phase, participants are recommended to take some time to reflect on the differences between knowledge and belief, and to connect their reflections with the practice cases. This is quite common in philosophy teaching, but was also introduced with the explicit aim of drawing subjects’ attention to potentially “overlooked aspects of the description of the scenario” (Williamson 2011, 226).²³

Fourthly, I provide participants with an incentive to pay attention to the practice session (and to mimic grading in a philosophy class), namely, a bonus of £5 for a correct answer to the final (and clearly designated) question they see (about a Gettier, Non-truth, or Knowledge Case).

Altogether, I believe this setup is maximally close to what happens in a philosophy undergraduate course, given the inherent constraints of online experiments. Moreover, the design is entirely in line with the methods used by economists answering to the behavioral challenge.

So, what do we find? There is no improvement in the Gettier Case. In the two other cases, the result was the reverse of that in the previous (non-incentivized, short) learning treatment: no improvement in the Knowledge Case, and some

²² Weinberg et al. claim that the set of cases that a typical professional philosopher will have gone through after completing undergraduate and graduate studies is “orders of magnitude” (2010, 342) smaller than, say, how often a chess player practices a given opening, and that “[p]hilosophy rarely if ever...provides the same ample degree of well-established cases to provide the requisite training regimen” (2010, 341). Williamson (2011) finds it more relevant to compare the number of times a philosophy student gets feedback with the number of times a law student gets similar feedback. Just as economists don’t turn subjects into probability or decision theorists when they rebut the behavioral challenge, for our purposes I don’t need to provide subjects with anything that comes close to a philosophy education.

²³ Together with the two vignettes in the middle of the learning phase on which subjects get more extensive feedback, this brings long learning somewhat in line with a suggestion voiced by Kauppinen (2007) that ideally eliciting judgments about philosophical thought experiments involves something of a dialogue between experimenter and subject, because that would increase the likelihood that the elicited judgments result from conceptual competence and not from irrelevant perturbing factors. Kauppinen (2007, 110) seems to support my convergence hypothesis here, as he believes that laypeople judgments in such dialogical experiments would “line up with each other.”

improvement in the Non-truth Case ($p < .01$). In sum, learning doesn't seem to align laypeople and expert knowledge attribution.

7 Group consultation

Besides incentives and learning, economists have more recently started using *group consultation* as a way to align laypeople and theory. The key illustrative reference point is again Linda the Bank Teller. Charness et al. (2010) examined what would happen if participants were given the opportunity to talk about the question in groups, before individually answering it. Participants were grouped in pairs or triples. They were presented with the case, and each pair or triple was asked to have a discussion about the right answer to the question. After this consultation phase, participants had to individually answer the question. The researchers found that when participants discuss Linda the Bank Teller in pairs, the individual error rates reduce from 58% to 48%, a drop of almost 20% (10 percentage points).²⁴

May we hope that group consultation has a similar effect on laypeople judgments concerning knowledge and belief? A mechanism that might explain why group consultation benefits individual judgments about probability is described as the application of the *truth-wins* norm in *eureka-type* problems (Lorge and Solomon 1955). These are problems in which, when people fail to find the solution themselves, they will immediately recognize the solution (*eureka*) if someone points it out to them—and so the *truth wins*. Mathematics may be a case in point: you may not be able to find a proof for some proposition yourself, but when someone gives you a proof, you see it's a proof of that proposition. Or more generally, if two people consult with the aim of solving a eureka-type problem, and each has a positive probability p to find the solution, the truth-wins norms suggests that consultation will increase the probability of both knowing the solution to $1 - (1 - p)^2$, which is then greater than p .

Most studies do not find improvement of the size predicted by the truth-wins norm, even in cases that are unproblematically eureka-type; yet most studies do find substantial improvement (Heath and Gonzalez 1995). The improvement that Charness et al. (2010) found in the Linda the Bank Teller case is less than what the truth-wins norm predicts, but still considerable.²⁵

It is initially plausible to see philosophical questions about knowledge as eureka-type questions, particularly for advocates of the expertise defense who maintain that Gettier's (1963) paper provided a refutation of the justified true belief theory of

²⁴ The baseline error rates (i.e., for individuals) that Charness et al. (2010) found were lower than in Tversky and Kahneman (1983).

²⁵ The *ex post* (empirical) individual probability of error in their experiment is .58, so success probability is .42. Using this as a measure of the individual probability of answering the question correctly, we should expect pairwise consultation to lead to a success rate of $1 - (1 - .42)^2 = .66$. The success rate (pairwise) they empirically find is $1 - .48 = .52$.

knowledge that was accepted “[a]lmost overnight” by the “vast majority of epistemologists throughout the analytic community” (Williamson 2005, 3).²⁶ Just as with proofs in mathematics, it may have been difficult to design a counterexample to the justified true belief theory; but once it was there, (almost) everyone realized it was one.

So I designed an experiment that stays as close as possible to the design of Charness et al. (2010), and conducted one baseline experiment without group consultation and one experiment with group consultation. Like these scholars, I used a student sample (economics and business students, not philosophy, University of Groningen, The Netherlands), and conducted the experiment in a large lecture hall with ample room for groups to have private conversations.

Regarding results, here, too, there is little doubt that the convergence hypothesis doesn’t gain any plausibility. Only in the Knowledge Case there is a tiny effect of small statistical significance, but in the unexpected direction, namely, *increasing* error. In the Gettier Case and the Non-truth Case nothing happens. What boosted success in the Linda the Bank Teller case (and helped confronting the behavioral challenge) does not improve knowledge attribution. So let’s turn to the last study.

8 Experimental control

This experiment is motivated by an observation due to Gigerenzer (2001). Gigerenzer reported results suggesting that German students perform better than American students in such tasks as Linda the Bank Teller and base-rate reasoning, and attributed the difference in performance to the way that experiments are conducted in Germany. In Gigerenzer’s lab, participants engage in face-to-face contact with the experimenter. They come to the lab individually, or in small groups, which allows them to focus on the task at hand without distraction. American experiments, by contrast, do not involve one-on-one contact between participant and experimenter. Participants do not really come to a lab, but rather are tested during classes, or in take-home experiments, where students get a booklet with assignments to be solved at home and to be returned to the experimenter afterwards (or students have to carry out the experiment online). According to Gigerenzer, it is this loss of *experimental control* that largely explains the difference between US and German samples: American experimental subjects are more likely to have faced distraction.

While take-home experiments may be rare in experimental philosophy, the majority of experiments do not involve the extensive experimental control that Gigerenzer’s lab offers. The first two studies reported in this paper are quite representative here: subjects were selected through an online portal (limited control), and participated in experiments in a lecture hall (a bit more control), and were presented with an online survey. There was fairly limited influence on the circumstances in which they answered the questions. Hence we should expect that

²⁶ Also see, e.g., Grundmann (2010), Horvath (2010), Jackson (2011) and Ludwig (2007).

with increased experimental control we see what we see in economics: convergence to orthodoxy.

So I recruited non-philosophy students to the Faculty of Economics & Business Research Lab of the University of Groningen, The Netherlands, and had them answer knowledge attribution questions in soundproofed, one-person cubicles, without access to Internet, and with a no-cellphone policy. But I faced a methodological issue here. It would be hard to justify comparing error rates in this sample with error rates in, say, the baselines of the first studies, or any other study. This is because these samples have distinct demographic characteristics. The samples used for the studies with incentives and learning are somewhat representative of the UK population. The experiment on group consultation has a mainly Dutch sample, even though it includes international students, but only undergraduate students. The lab experiment has a mainly Dutch sample, but includes international students at the undergraduate and graduate level. To circumvent these issues, I approached the question a bit more indirectly, and conducted another randomized controlled trial (RCT) to see if—with experimental control—incentives and learning effects are greater than in the first RCT (discussed in Sects. 4–6). This is an admittedly indirect measure, but I believe it's methodologically preferable. All the same, if we do find large differences between this study's baseline and the baselines of the previous two studies, this may give some evidence backing the convergence hypothesis that I ultimately set out to test.

So does experimental control finally bring answers closer to epistemological consensus? No: none of the treatments in this last study (incentives, short learning) had any statistically significant impact on error rates.

But perhaps there is some effect if we set aside the methodological scruples and compare baselines? Under experimental control, the baseline error rate in the Gettier Case is 63%. This compares positively with the error rate of 72% from the baseline of the first RCT, which is the baseline used to compare incentives, short and long learning. But as that baseline involves a UK national sample (whereas here we have students only, and mostly Dutch), it is more natural to compare the results with the baseline of the group consultation study, to which it is closer demographically. That baseline error rate was 52%, however. So even with a relaxed methodology, this study has not generated any support for the convergence hypothesis, however much I had expected this. Experimental control doesn't decrease error rates.

9 Conclusion

The idea of this paper was to see what remains of the experimental or restrictionist challenge to armchair philosophy if we adopt the machinery that economists have used to counter the behavioral challenge that arose out of research by such psychologists as Tversky and Kahneman. The economists showed that many of the anomalies detected by the psychologists become considerably smaller if participants obtain financial incentives, go through a learning period, work together in groups, or face more stringent forms of experimental control. The thought behind the project was that if similar effects could be demonstrated with regard to philosophy tasks, we

would be in a good position to describe the repercussions the experimental challenge should have for philosophy—just about the same as for economics, that is, not too much: *orthodox* economics is still alive and flourishing, and while the behavioral challenge has generated the field of *behavioral* economics, this new discipline is, after more than three decades, still fairly small.

But nothing of the sort. Rigorous testing shows that these techniques don't get us anywhere near the effect they have in economics. The experimental philosophy results seem to be much stronger than the behavioral economics results.

How philosophers should now deal with the experimental challenge is not my concern here. Yet I do wish to probe the force of the results by addressing a few objections that might be raised. I deal with most of these objections by conducting further robustness checks and/or providing more information on the samples.

To begin with, one might wonder how the samples compare with earlier results from the experimental philosophy literature when we consider the various control variables used in the regressions: gender, ethnicity, and so on. Examining the effects of ethnicity (culture) and socioeconomic status on knowledge attribution has been on the agenda of experimental philosophy from the very start (Weinberg et al. 2001). I don't find evidence for such effects.

Gender effects have also attracted quite a bit of attention. Buckwalter and Stich (2014), for instance, report that women tend to attribute knowledge more frequently on average in particular vignettes, but these results may not replicate (Machery et al. 2017; Nagel et al. 2013; Seyedsayamdost 2014). I find fairly considerable gender coefficients in the Gettier Case ($p < .05$ and $p < .01$), in two of the three treatments of the RCT, with women indeed more frequently than men attributing knowledge. In the other studies the effects are negligible.

Colaço et al. (2014) find some age effects. The age coefficients I find are only in a few cases significant, and always small.

To my knowledge, there are no studies that zoom in on religion or education, except that Weinberg et al. (2001) use educational achievement as a proxy for socioeconomic status. I find no evidence of a clear religion effect, but I do find more systematic and substantial education effects: subjects possessing a college degree are more likely to give the consensus answer in the incentivized and short learning treatments, both in the Gettier Case and the Non-truth Case, with significance ranging from the 5% to the 0.1% level. This effect is, however, fully absent in the Knowledge Case.

Finally, except for religion ($p < .05$) there is no evidence for interaction effects between treatment type and any control variable (not even if instead of the usual college education dummy we use a fine-grained variable with seven educational achievement brackets). So the way in which people respond to, say, incentives vis-à-vis the baseline is not influenced by their age, gender, and so forth, but only (marginally) by their religiosity.

All in all, considering the covariates I don't find effects that have not already been postulated in the literature before (except, perhaps, religion and education).

Further zooming in on education, a second issue to deal with arises out of remarks by some who have written in the spirit of the expertise defense to the effect that education might raise tendency to reflect. Devitt (2011, 426), for instance,

writes that “a totally uneducated person may reflect very little and hence have few if any intuitive judgments about her language.” In the regressions, I used a college education dummy, following standard practices, and as I said, there were some effects in some treatments and some vignettes. But perhaps a bigger and more systematic effect can be found if we use a dummy variable with a different cut-off point that distinguishes between “totally uneducated” subjects and all others?

And indeed, to some extent Devitt’s predictions are borne out by the data. If we call a subject *totally uneducated* if he or she has qualifications at level 1 or below (the lowest educational level in typical UK surveys), there is indeed a very substantial effect on correctness in the Gettier Case (but not in the Non-truth Case and Knowledge Case). However, as the baseline success rate for subjects with qualifications *above* level 1 (so the complement of the reference group) is only 30%, this result is unlikely to bring solace to the armchair philosopher: the “not totally uneducated” are still very far removed from the professional philosopher.²⁷

Third: Ludwig (2007, 151) makes the very interesting suggestion that after a training session we might want to select those subjects that are “best at responding on the basis of competence in the use of concepts,” and then present these subjects with scenarios of interest. To see if that works, let’s compare the best scoring subjects in the short and long learning phases, respectively, with all baseline subjects. In the short learning study, subjects that score best during the learning phase on average score *worse* on the final question than the average subject in the baseline, but the difference is not statistically significant. The best scoring subjects (ten out of ten questions correct) in the long learning study score *better* on the final question than the average subject in the baseline, but the difference is not significant. Iteratively relaxing the definition (so allowing for one error in the learning phase, or two, or three, and so on) in the short learning study does not yield significance. In the long learning study, however, it does. If we call a best scoring subject one who has answered nine or ten questions correctly, then 72% of the best scoring subjects answer the Gettier Case correctly, as opposed to 28% in the baseline, and this difference is highly significant ($Z = 3.36, p < .001$).²⁸

We shouldn’t be overly enthusiastic about this finding, however, since we may well have individuated here those subjects that would give the consensus answer to the Gettier Case anyway. Compare: if you look at those subjects that do well on ten practice questions preparing for the Linda the Bank Teller task, you may well find that these are subjects that come to the experiment with greater probabilistic

²⁷ Baseline success rate (constant) is .30. The coefficient is $-.24$ ($p < .05$), with regression here on the baseline (untreated) sample in the first RCT. Interpretation: success rate among those participants with level 1 or below is 6%; success rate among those with qualifications above level 1 is 30%; and the difference is of mild statistical significance. The difference becomes smaller, and no longer significant, if we relax the definition of the dummy. For the strictest definition of the dummy, however, the results seem fairly robust in that they persist if we consider the entire sample of the first RCT (baseline and the three treatments). Non-truth and Knowledge Case show no effects.

²⁸ We could also examine whether the number of correct answers during the learning phase predicts success on the final question: there is no effect in the short learning study, but in the long learning study, the correlation is positive and mildly significant ($p < .05$), in all three cases.

expertise already, so the success rate cannot then be attributed to the practice session.

To circumvent this, we could include a measure of reflection time; for indeed, experts who are forced to go through ten practice cases will rush through it, but laypeople who really learn from the practice will take their time. But this is of no avail: a simple regression showed that reflection time in the long learning study does not predict correctness in any of the three cases.

Fourth: Sosa (2009, 106) suggests that in order for there to be a philosophically relevant conflict between the intuitions of laypeople and those of philosophers concerning a given scenario, these intuitions should be “strong enough.” So one should perhaps expect that participants with strong intuitions are more likely to provide correct answers. We can test this by estimating the correlation between the strength of the intuition, as measured by the confidence question in the surveys, and correctness of the answer to the Gettier Case. But the results of that exercise have the wrong direction, and are not statistically significant, so here, too, my findings are sufficiently robust: they also hold for people with strong intuitions.²⁹

Fifth: one might think that a good predictor of correctness of answers to a question about a Gettier case is whether a subject has reflected sufficiently deeply on the case and question, or so the armchair philosopher may argue. Experiments conducted by Weinberg et al. (2012), among others, do not seem to offer much hope for a reflection defense of this type, but the authors acknowledge that the instruments they use to measure reflection may not be fully adequate. They use the Need for Cognition Test (Cacioppo and Petty 1982), and the Cognitive Reflection Test (Frederick 2005), but social science researchers use these tests (particularly the second one) more often as a proxy for intelligence (IQ) than as a genuine measure of reflection, which is conceptually distinct from intelligence. I included in the long learning study a few stages where subjects were explicitly recommended to reflect on the difference between knowledge and belief, which gives us a more direct measure of reflection: the time spent on these stages of the experiment. My findings here are, however, entirely in line with those of Weinberg et al. (2012): regression shows that people who reflect longer in the long learning study are not more likely to come up with the consensus answer in any of the three cases.³⁰

The upshot of the discussion so far is that further robustness checks weaken the force of a few objections that might be raised against my conclusions. Yet, while I’m confident that the results will robustly stand up to scrutiny, I’d like to mention a number of limitations of the study and indicate some avenues for future research.³¹

First: data always have specific demographic characteristics. For all we know, what I have found is valid only for UK citizens (incentives, learning), or only for

²⁹ Regressions on a dummy for very high confidence (confidence = 7), or on a dummy for high confidence (confidence = 5, 6, or 7) do not change this picture. On reporting confidence, also see Weinberg (2007), and on empirical work on confidence, see Wright (2010).

³⁰ One might object that I measure the quantity but not the quality of reflection, but it is hard to see how you could develop a measure of the quality of reflection that does not lead to circularity (endogeneity), that is, a measure that is independent of success in the tasks at hand.

³¹ Many thanks to an anonymous reviewer for pressing me on points 2–6 mentioned here.

students of economics and business (group consultation, experimental design). Such things hold for almost any study in experimental social science, but it is important to note that the external validity of our findings will only increase if similar studies are conducted with samples with different demographics.

Second: I only look at one Gettier Case, one Non-truth Case, and one Knowledge Case. We know from the experimental philosophy literature, however, that there is considerable variation in laypeople judgments across cases, and what I find here may, for all we know, be an artefact of the three vignettes. Therefore, future research should try and replicate with a different set of vignettes.

Third: the experimental challenge isn't only to do with knowledge attribution judgments, or with epistemology, but affects the whole of philosophy. I cannot rule out that incentives, learning, group consultation, or experimental control make laypeople converge to philosophers if they are asked to consider vignettes drawn from other branches of philosophy. In fact, as I take economics as my main source of inspiration, we should positively expect future research to show that in different contexts these experimental techniques have effects of different size and character: the economics results aren't homogeneous either.

Fourth: it may seem as though I think that the layperson–expert opposition is what has driven most of experimental philosophy, so I should stress that I don't think that, and realize that initially experimental philosophers were primarily concerned with deviations from philosophical consensus in particular segments of the laypeople population (culture, socioeconomic status, etc.). Analogous research does exist in psychology. It examines, to give one example, how culture or socioeconomic status influences perception. But such research has never played such a prominent role in the behavioral challenge to economics as Linda the Bank Teller and the other experiments cited in this paper. So I should acknowledge that setting aside the layperson–expert opposition would require a different experimental setup, and might lead to different results.

Fifth: while the design of the two learning treatments in the randomized controlled trial closely mirrors what is standard in economics experiments, it only very tentatively approximates full-blown professional philosophical training. This matters less if the aim is only to see whether the economist's tactics can be deployed against experimental philosophy, which was my initial plan; for indeed, the economists don't provide their subjects with anything near a genuine economics training either. But it matters more if you consider the relevance of my results to the expertise defense, because the expertise defender might retort that going through a handful Gettier cases doesn't make an expert. Even in the long learning treatment we cannot indeed exclude that subjects come with widely divergent background assumptions that influence the way they interpret the vignettes and the questions (Sosa 2009). I agree. But an advocate of the expertise defense should find it at least somewhat puzzling that incentives and the other experimental techniques that work in economics don't even make things a tiny bit better. So I acknowledge this point as a potential limitation of the study, and hope that more extensive empirical work, for instance using some of Turri's (2013) suggestions within the framework of a randomized controlled trial, will give us more definite answers.

Sixth: while the Linda the Bank Teller case has passed into a byword for behavioral economics, it shows only one of the many biases that have been documented in the literature, and we would therefore be grossly remiss if we didn't at least contemplate the possibility that knowledge attribution cases might be more like those biases that experimental techniques have a lesser hold over.

To begin with, clearly no experimental strategy will fully remove bias: even with incentives, still about a third of the subjects in Linda the Bank Teller go against probability theory. So my convergence hypothesis was that in knowledge attribution cases we would see a reduction of errors, not that they would fully vanish.

But indeed, the economics literature does show that the experimental techniques don't always work (Hertwig and Ortmann 2001). What seems to be beyond reasonable doubt is that incentives backfire when they create too much pressure on subjects, and that they may "crowd out" intrinsic motivation. It is also clear that group consultation is unlikely to do much work if subjects find it hard to ascertain their partner's credentials or otherwise to evaluate or corroborate their partner's contributions. Perhaps knowledge attribution tasks are more like Wason selection tasks and other logic tasks, where incentives don't seem to do as much as in Linda the Bank Teller. Or perhaps, as an anonymous reviewer kindly suggested to me, in order to understand why we find that these experimental techniques don't boost performance, we should look at tasks in which analytic answers are not so readily available as they are in logic and probability theory. Based on my reading of the extant economics literature, however, my impression is that a clear pattern has yet to emerge from the welter of documented cases, and that we cannot confidently answer these questions without more systematic further research. But it is certainly possible that future behavioral economics will develop cases that become just as central and paradigmatic as Linda the Bank Teller—and bear even greater similarity to philosophy tasks—but for which incentivization and the other techniques do *not* work, and this would probably weaken the force of the claims I defend in this paper.

Seventh: several authors have responded to the experimental challenge arguing that much published work in experimental philosophy (they think) doesn't meet the high standards of empirical research in the social sciences (Woolfolk 2013). This is strikingly similar to what economists say about psychologists. As Hertwig and Ortmann (2001) document, economists believe that psychologists use inferior experimental techniques. In particular, economists blame psychologists: for not giving participants clear scripts, for using snapshot studies instead of repeated trials involving learning and feedback, for the fact that psychologists don't implement performance-based monetary payments (incentives), and because psychologists often deceive participants about the purposes of the experiment. As a result, the economists think, their work is far more reliable than psychology.

I don't need to take a stance here, even if my methods are clearly drawn from economics rather than psychology. I carried out a randomized controlled trial, which is seen as the gold standard in economic impact analysis, and followed widely accepted practices in economics. I then mirrored the design of a frequently cited and influential paper in economics on the Linda the Bank Teller case, and here, too, used standard (and admittedly not very complex) econometrics. And the last study was conducted in highly controlled lab conditions. The statistical analysis is in line with

what is common in economics, as are the various *robustness checks* reported in the present section. Moreover, sample sizes are above what is generally expected in the field. But one can always do more, and I can only hope that others, too, will turn to economics for methodological inspiration—not to replace, but to accompany psychology.

It is time to conclude. I am aware of the potential limitations of the study, but despite these limitations, I feel there is little doubt that as it stands the consolation armchair philosophers might want to find in the economist's response to the behavioral challenge is spurious. My initial idea of saving the armchair by experiment has failed. This has made the experimental challenge all the more powerful.

Acknowledgements Warmest thanks are due to audiences in Norwich and Osnabrück, and to Mark Alfano, Tammo Bijmolt, Mark Curtis, Eugen Fischer, Joachim Horvath, Nikil Mukerji, Laetitia Mulder, Marijke Leliveld, Robert Lensink, Chiara Lisciandra, Shaun Nichols, Kevin Reuter, Melissa Vergara Fernández, Tom Wansbeek, Juliette de Wit, Raymond Zaal, Lieuwe Zijlstra, and two anonymous reviewers for this journal, and to Reinder Dallinga for lab assistance.

Funding Research for this paper was partly funded by the Dutch Research Council (NWO), Grant Number 360-20-310.

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alexander, J. (2016). Philosophical expertise. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 557–567). Chichester: Wiley.
- Alexander, J., & Weinberg, J. M. (2007). Analytic epistemology and experimental philosophy. *Philosophy Compass*, 2(1), 56–80.
- Bach, T. (2019). Defence of armchair expertise. *Theoria*, 85(5), 350–382.
- Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46(3), 610–620.
- Binmore, K. G. (1994). *Game theory and the social contract. Vol.1, Playing fair*. Cambridge, MA: MIT Press.
- Buckwalter, W. (2016). Intuition fail: Philosophical activity and the limits of expertise. *Philosophy and Phenomenological Research*, 92(2), 378–410.
- Buckwalter, W., & Stich, S. (2014). Gender and philosophical intuition. *Experimental Philosophy*, 2, 307–346.
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1), 116–131.

- Camerer, C. (1995). Individual decision making. In J. H. Kagel & A. E. Roth (Eds.), *Handbook of experimental economics* (pp. 587–703). Princeton: Princeton University Press.
- Charness, G., Karni, E., & Levin, D. (2010). On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda. *Games and Economic Behavior*, 68(2), 551–556.
- Clarke, S. (2013). Intuitions as evidence, philosophical expertise and the developmental challenge. *Philosophical Papers*, 42(2), 175–207.
- Colaço, D., Buckwalter, W., Stich, S., & Machery, E. (2014). Epistemic intuitions in fake-barn thought experiments. *Episteme*, 11(2), 199–212.
- Deutsch, M. (2009). Experimental philosophy and the theory of reference. *Mind & Language*, 24(4), 445–466.
- Devitt, M. (2011). Experimental semantics. *Philosophy and Phenomenological Research*, 82(2), 418–435.
- Drożdżowicz, A. (2018). Philosophical expertise beyond intuitions. *Philosophical Psychology*, 31(2), 253–277.
- Egler, M., & Ross, L. D. (2020). Philosophical expertise under the microscope. *Synthese*, 197, 1077–1098.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42.
- Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6), 121–123.
- Gigerenzer, G. (2001). Are we losing control? *Behavioral and Brain Sciences*, 24, 408–409.
- Grundmann, T. (2010). Some hope for intuitions: A reply to Weinberg. *Philosophical Psychology*, 23(4), 481–509.
- Hales, S. D. (2006). *Relativism and the foundations of philosophy*. Cambridge, MA: MIT Press.
- Hansson, S. O. (2020). Philosophical expertise. *Theoria*, 86(2), 139–144.
- Harrison, G. W. (1994). Expected utility theory and the experimentalists. *Empirical Economics*, 19, 223–253.
- Heath, C., & Gonzalez, R. (1995). Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. *Organizational Behavior and Human Decision Processes*, 61(3), 305–326.
- Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics a methodological challenge for psychologists. *Behavioral and Brain Sciences*, 24, 383–451.
- Hitchcock, C. (2012). Thought experiments, real experiments, and the expertise objection. *European Journal for Philosophy of Science*, 2(2), 205–218.
- Hofmann, F. (2010). Intuitions, concepts, and imagination. *Philosophical Psychology*, 23(4), 529–546.
- Horvath, J. (2010). How (not) to react to experimental philosophy. *Philosophical Psychology*, 23(4), 447–480.
- Horvath, J., & Wiegmann, A. (2016). Intuitive expertise and intuitions about knowledge. *Philosophical Studies*, 173(10), 2701–2726.
- Irikefe, P. O. (2020). A fresh look at the expertise reply to the variation problem. *Philosophical Psychology*, 33(6), 840–867.
- Jackson, F. (2011). On Gettier holdouts. *Mind & Language*, 26(4), 468–481.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.
- Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical Explorations*, 10(2), 95–118.
- Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, 60(1), 10–13.
- Lorge, I., & Solomon, H. (1955). Two models of group behavior in the solution of eureka-type problems. *Psychometrika*, 20(2), 139–148.
- Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy*, 31, 128–159.
- Ludwig, K. (2010). Intuitions and relativity. *Philosophical Psychology*, 23(4), 427–445.
- Machery, E. (2011). Thought experiments and philosophical knowledge. *Metaphilosophy*, 42(3), 191–214.
- Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2013). If folk intuitions vary, then what? *Philosophy and Phenomenological Research*, 86(3), 618–635.

- Machery, E., Stich, S., Rose, D., Alai, M., Angelucci, A., Berniūnas, R., et al. (2017). The Gettier intuition from South America to Asia. *Journal of Indian Council of Philosophical Research*, 34(3), 517–541.
- Mizrahi, M. (2015). Three arguments against the expertise defense. *Metaphilosophy*, 46(1), 52–64.
- Nado, J. (2014a). Philosophical expertise. *Philosophy Compass*, 9(9), 631–641.
- Nado, J. (2014b). Philosophical expertise and scientific expertise. *Philosophical Psychology*, 28(7), 1026–1044.
- Nagel, J., Juan, V. S., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition*, 129(3), 652–61.
- Ryberg, J. (2013). Moral intuitions and the expertise defence. *Analysis*, 73(1), 3–9.
- Schulz, E., Cokely, E. T., & Feltz, A. (2011). Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense. *Consciousness and Cognition*, 20(4), 1722–31.
- Schwitzgebel, E. (2009). Do ethicists steal more books? *Philosophical Psychology*, 22(6), 711–725.
- Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, 27(2), 135–153.
- Seyedsayamdost, H. (2014). On gender and philosophical intuition: Failure of replication and other negative results. *Philosophical Psychology*, 28(5), 642–673.
- Seyedsayamdost, H. (2019). Philosophical expertise and philosophical methodology: A clearer division and notes on the expertise debate. *Metaphilosophy*, 50(1–2), 110–129.
- Simon, H. A. (1948). *Administrative behaviour: A study of the decision making processes in Administrative Organisation*. New York: Macmillan.
- Simon, H. A., & Bartel, R. D. (1986). The failure of armchair economics. *Challenge*, 29(5), 18–25.
- Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, 132(1), 99–107.
- Sosa, E. (2009). A defense of the use of intuitions in philosophy. In D. Murphy & M. Bishop (Eds.), *Stich and his critics* (pp. 101–112). Chichester: Wiley.
- Starmans, C., & Friedman, O. (2012). The folk conception of knowledge. *Cognition*, 124(3), 272–83.
- Thaler, R. (1987). The psychology of choice and the assumptions of economics. In A. E. Roth (Ed.), *Laboratory experimentation in economics: Six points of view, Book section 4* (pp. 99–130). Cambridge: Cambridge University Press.
- Tobia, K., Buckwalter, W., & Stich, S. (2013). Moral intuitions: Are philosophers experts? *Philosophical Psychology*, 26(5), 629–638.
- Turri, J. (2013). A conspicuous art putting Gettier to the test. *Philosophers' Imprint*, 13(10), 1–16.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Science*, 90(4), 293–315.
- Vaesen, K., Peterson, M., & Van Bezooijen, B. (2013). The reliability of armchair intuitions. *Metaphilosophy*, 44(5), 559–578.
- Weinberg, J. M. (2007). How to challenge intuitions empirically without risking skepticism. *Midwest Studies In Philosophy*, 31, 318–343.
- Weinberg, J. M., Alexander, J., Gonnerman, C., & Reuter, S. (2012). Restrictionism and reflection challenge deflected, or simply redirected. *Monist*, 95(2), 200–222.
- Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters? *Philosophical Psychology*, 23(3), 331–355.
- Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, 29(1–2), 429–460.
- Williamson, T. (2005). Armchair philosophy, metaphysical modality and counterfactual thinking. *Proceedings of the Aristotelian Society*, 105(1), 1–23.
- Williamson, T. (2007). *The Philosophy of Philosophy*. Malden: Blackwell.
- Williamson, T. (2009). Replies to Ichikawa, Martin and Weinberg. *Philosophical Studies*, 145(3), 465–476.
- Williamson, T. (2011). Philosophical expertise and the burden of proof. *Metaphilosophy*, 42(3), 215–229.
- Williamson, T. (2016). Philosophical criticism of experimental philosophy. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 20–36). Chichester: Wiley.
- Woolfolk, R. (2013). Experimental philosophy: A methodological critique. *Metaphilosophy*, 44(1–2), 79–87.
- Wright, J. C. (2010). On intuitional stability: The clear, the strong, and the paradigmatic. *Cognition*, 115(3), 491–503.

Yaniv, I., & Schul, Y. (1997). Elimination and inclusion procedures in judgment. *Journal of Behavioral Decision Making*, 10(3), 211–220.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.