Are thought experiments "disturbing"? The case of armchair physics

Samuel Schindler and Pierre Saint-Germier

Aarhus University

This paper is forthcoming in *Philosophical Studies*. Please cite the published version.

Abstract

Proponents of the "negative program" in experimental philosophy have argued that judgements in philosophical cases, also known as case judgements, are unreliable and that the method of cases should be either strongly constrained or even given up. Here we put one of the main proponent's account of why philosophical cases may cause the unreliability of case judgements to the test. We conducted our test with thought experiments from physics, which exhibit the exact same supposedly "disturbing characteristics" of philosophical cases.

1 Introduction

The use of judgements in so-called 'cases' or thought experiments in philosophical theorizing has recently come under severe attack: there is now a host of studies showing that the 'case judgements' by the folk and even by professional philosophers vary with extraneous demographic and presentation variables.[1] Proponents of the 'negative program' of experimental philosophy argue that the method of cases ought to be strongly constrained or even given up entirely because case judgments are not considered reliable evidence for philosophical theorizing (Alexander and Weinberg 2007, Knobe and Nichols 2017, Machery 2017).

In what probably constitutes the most extensive and systematic treatment within the negative program, Machery (2017) has claimed that philosophical cases exhibit three types of "disturbing characteristics" which, he believes, are likely to cause the unreliability of case judgments: cases are (i) unusual, (ii) pull apart properties that usually go together,

[1] Those results are conveniently summarized in chapter 2 of Machery (2017).

and (iii) contain superficial content irrelevant to the philosophical point in question.[2] Cases are unusual, according to Machery, "if and only if we encounter it infrequently or if we rarely read texts about it" (113). Trolley cases, for example, are unusual in that subjects are asked to decide over other lives; something that most subjects would not have had to consider at any time during their normal lives. Philosophical cases, according to Machery, also pull apart properties that go together in everyday life. For example, in the footbridge case, subjects are asked to engage in physical violence in order to do more good than harm, although violence usually causes more harm than good (116). In Gettier cases, subjects cannot rely on their usual strategies for identifying knowledge, as such cases pull apart properties epistemologists have identified as crucial for knowledge attributions (e.g. safety and adherence). Lastly, philosophical cases are often described in "vivid terms", contain "many irrelevant narrative elements", and are presented in a "tendentious manner" (119). Machery believes that any of these three characteristics can individually cause the unreliability of case judgements (112).

Machery doesn't claim to be in possession of any direct evidence for the three disturbing characteristics *in fact* causing the unreliability of case judgements. For example, he is not able to provide any evidence that the pulling apart of properties in trolley *actually* causes subjects' susceptibility to order effects. Yet, Machery is fairly confident that the characteristics he identifies to be "likely culprits" on the basis of his having provided "*good explanations* of why the judgements examined by experimental philosophers are influenced by demographic and presentation variables" (112; added emphasis).

How could one go about testing Machery's claims about the characteristics of cases being disturbing? It is good practice in science to test hypotheses not only on the basis of the evidence which they were constructed for, but also with evidence that lies outside of their original domain of application. This avoids the suspicion of the hypothesis being gerrymandered to the evidence at hand and to actually get at something real (Worrall 2002). Similarly, in philosophizing it is good practice to probe the claims of one's interlocutor by testing them against examples that satisfy the conditions set out by one's interlocutor but which were not considered and which might possess features which are

[2] Machery emphasizes that not every case must exhibit *all* of these three features in order to cause the unreliability of case judgements, nor does he claim that any of the characteristics necessitate unreliability. He only believes that they make unreliability *more likely* (112).

capable of challenging one's interlocutor's claims. In this paper we apply this strategy to probe Machery's claims about the disturbing characteristics of cases. The example which we think satisfies at these characteristics are *thought experiments in physics*.

Thought experiments in physics, too, are "unusual" in Machery's sense: they usually involve quite bizarre situations which we would not normally encounter. We, for example, do not frequently encounter situations in which we are asked to make a judgement about whether or not a cat trapped in a box in which it has a 50% chance to die, is either dead or alive upon opening the box. Thought experiments in physics also pull apart properties that usually go together: in the so-called Einstein's elevator thought experiment, one is asked to consider the trajectory of falling objects in a gravitational field-free space (usually we are surrounded by the gravitational field of the earth). And thought experiments in physics are also rich in superficial content which is irrelevant to the point under consideration. For example, in Galileo's thought experiment of two connected falling objects, the precise weight of the two objects is irrelevant to the point under consideration (namely) – so long as one body is suitably heavier than the other.[3]

In this paper we therefore tested whether thought experiments in physics can be said to affect the reliability of judgments in a way that Machery claims thought experiments in philosophy affect the reliability of judgments. We found no evidence for this claim.

In Section 2 of this paper we present the methods used in our test of the hypothesis that physicists make more reliable case judgments than non-physicists. In Section 3, we present the results of our study. In Section 4 we address several objections against our results and interpretation. In Section 5 we draw our conclusion.

2 Method

What does it mean for a case judgement to be reliable? Machery provides the following definition:

> *T*, a psychological process outputting judgments, is reliable in environment *E* if and only if in *E* either *T* has the disposition to produce a large proportion of *true judgments* or, if *T* is an inferential process, *T* has the disposition to produce a large proportion of *true judgments if*

[3] We will discuss these and other thought experiments in detail later in the paper. See also Appendix 2 for a detailed description of the thought experiments we used.

> *its inputs are true*. Reliability here is a dispositional property, and a process used only once can still be assessed for reliability. (Machery 2017, 96; added emphasis)[4]

By implication, a *judgement* (generated by the relevant psychological process) is reliable if it is likely to be correct.[5] For Machery, the method of cases constitutes an environment in which this condition is not satisfied and in which the generated judgements are unreliable. For Machery's account about the disturbing characteristics to be correct, the following hypothesis would have to be rejected:

(H1) Subjects are likely to make reliable judgements in thought experiments in physics.

In order to test hypothesis 1 we decided to use subjects which seem prima facie most competent in making judgments in thought experiments in physics, namely physicists.

How does one know whether a given judgement in a thought experiments in reliable? Machery (2017) unfortunately provides little guidance here; he only focuses on conditions under which case judgements are off. But they must be off with regard to *some* standard. Horvath and Wiegmann (2015) provide a solution: they propose to use the textbook consensus about case judgments as a *defeasible* standard for gauging the reliability of the case judgements of our subjects. For thought experiments in physics, this standard is readily available (e.g. Brown and Fehige 2011). We shall use it here in the way suggested by Horvath and Wiegmann.

Machery's account allows for another hypothesis. Machery believes that thought experiments are not only outside the "proper domain" of reliable judgements (112), but he also believes that philosophers have no special expertise in making case judgements. In other words, Machery seems to think that thought experiments are just as disturbing to the folk as they are to philosophers.

The question of whether or not philosophers have an expertise in making case judgments has played an important role in debates between experimental philosophers and proponents of the traditional armchair method (Machery 2011, Williamson 2011). If philosophers do possess such expertise, then there would seem to be a case for not using

[4] Machery in the same place discusses other proposals for how to assess the epistemic credibility of case judgements (hopelessness and calibration), but provides arguments for why reliability is the best measure.
[5] Machery himself uses the predicate 'reliable'/'unreliable' for judgments throughout his book (Machery 2017). At the same time, he doesn't have much to say about the psychological processes generating the judgments and instead focuses on the characteristics of the environments in which, he believes, judgments become unreliable. We will turn to those in Section 3.2.

lay subjects but rather philosophers themselves when doing experiments on case judgements in philosophy. We tested whether the analogous idea would hold in thought experiments in physics:

(H2) Physicists are more likely to make reliable judgements in thought experiments in physics than non-physicists.

Of course, the reasons why physicists might be better than non-physicists could be different ones from the ones that could be brought to bear in the comparison between philosophers and non-philosophers. We will discuss this issue later on (Section 4).

2.1 Materials

We designed a set of six tasks. Each task consisted of a description of a thought experiment in physics, Phy-TE for short. Each task was also accompanied by a figure representing the Phy-TE, a comprehension question about the text, and a statement describing either the standard judgement about the thought experiment, or its negation. We chose six classical Phy-TE for our tasks, which are well known and discussed in the philosophy of science literature (Brown and Fehige 2011): Stevin's chain, Schroedinger's cat, Galileo's tower, Galileo's ship, and Newton's cannon. These thought experiments are listed and explained in detail in Appendix 2.

The thought experiments we used seem to exhibit several of the features singled out in philosophical cases as causes for the unreliability of case judgments. Einstein's elevator puts the reader into a highly 'unusual' situation of being dragged by a space ship through in an opaque container outer space. It's also highly unusual for cats, for example, to be trapped in a box in which they are at risk of dying from the poisonous contents of a broken flask.

Just like thought experiments in philosophy, many Phy-TE also pull apart properties that usually go together. For example, quantum properties are usually associated with micro-objects like electrons. Schroedinger's cat asks us to consider a scenario in which quantum properties are assigned to the macro-object of a cat. Or consider Newton's cannon: usually projectiles never have the speed that would be required for shooting an object into an orbit around the earth. Even thought experiments like Galileo's tower can be said to be unusual: normally, unsupported bodies of different shape and size fall to the

ground at different speeds (contrary to what Galileo wanted to show) because their air resistance is different.

Many Phy-TE also offer lots of 'irrelevant narrative elements' in what Machery calls 'superficial content' of a thought experiment. For example, it is entirely irrelevant whether the cat in the box dies because a Geiger counter or some other device triggers the release of a toxin, or whether the toxin is contained in a flask that breaks. Likewise, it is quite irrelevant to Newton's cannon whether the device used to shoot the object into an orbit around the earth is an (unrealistically powerful) cannon or an (unrealistically powerful) rocket launcher. It also doesn't matter what the weight of the bodies used in Galileo's tower thought experiment, so long one is significantly heavier than the other.

2.2 Procedure

The following procedure was used to test both H1 and H2. Participants were asked to carefully consider the thought experiment and answer two questions. The first question was a relatively simple comprehension question that asked participants to finish a statement about a relevant element of the scenario and was designed to probe whether they have understood the text. Only participants who answered the comprehension question correctly were included in our analysis. The comprehension questions were designed in such a way that they would not guide the participants towards the correct answers to the second question (see Appendix 2).

The second question asked participants to what extent they agreed or disagreed with a statement expressing a judgment in the hypothetical scenario. In half of the tasks participants were asked to evaluate statements expressing a correct judgement, and in the other half of the tasks participants were asked to evaluate the statement expressing the negation of a correct judgement. The participants were asked to answer by indicating their level of agreement/disagreement on a five-point Likert scale: 1 = Strongly disagree; 2 = Somewhat disagree; 3 = Neither agree nor disagree; 4 = Somewhat agree; 5 = Strongly agree. We measured those scores against the standard judgements reported in the literature on Phy-TE (Sklar 1992, Brown and Fehige 2011). For example, in the hypothetical scenario based on the Schrodinger's cat thought experiment, subjects were asked to indicate to what extent they agreed/disagreed with the statement "Before the box is

opened the dog is either dead or alive".[6] The six tasks were presented randomly and were followed by a short questionnaire gathering (see next section). For details concerning tasks and their design see Appendix 2.

2.3 Participants

In our first experiment testing hypothesis 1, we recruited physicists (n = 57) via mailing lists for PhD students, postdocs and faculty members in various physics departments in <blinded for peer review>. We required that our subjects had a PhD degree or were studying towards one. The mean age of participants in the physics group was 33 (SD = 10.2) of which 11 (19%) were female.[7] Out of all 57 participants in this group, 32 (56.1%) reported having a PhD and 25 (43.9%) being enrolled in a PhD programme in physics (see Appendix 1 for specialisations).

For testing hypothesis 2 in our second experiment, we recruited non-physicists (n = 60) via mailing lists for PhD students, postdocs and faculty members in various departments in science, social science, and humanities and excluding physics departments in <blinded for peer review> (see Appendix 1 for details). In order to rule out that a possible difference in task performance might be caused by different levels of education, we required a PhD or that subjects were studying towards one. The biggest subgroup of our subjects were political scientists (n=19). The mean age of participants in the control group was 34 (SD = 6.72). The control group consisted of 60 subjects of which 27 (47.3%) were female, 33 (57.9%) reported holding a PhD and 24 (42.1%) being enrolled in a PhD programme. The details regarding their education and areas of study are presented in Appendix 1. A background section located at the end of the questionnaire was used to collect information on participants' education, age, gender, level of English, and area of specialisation.[8]

---

[6] In our experiment, we assumed that physicists would reject this judgement. See Section 3 and Appendix 2 for further details.

[7] This reflects the unfortunate underrepresentation of women in physics (Sax et al. 2016)

[8] In total, 166 participants responded to our call and submitted their responses via the *Qualtrics* platform. Each participant had the option to leave their email address on an external Google Forms website in order to enter a lottery for 5 amazon.com vouchers of each $25 or to receive information about the results of the study. We excluded partly incomplete questionnaires (n = 29), subjects who did not satisfy our minimal criteria for education, i.e. not being enrolled in a PhD programme (n = 13), subjects who did not have at least an intermediate level of English (n=1), and subjects who did not answer the comprehension questions correctly (n=6).

3 Results and discussion

In what follows we will first discuss the results of our experiments pertaining to H1 and then the results pertaining to H2. For both results we distinguished between a CORR (=correct) and a HCON (=highly confident) measure, whereby HCON is a count of strong agreements with the textbook consensus (or strong disagreements, when we presented statements negating the textbook consensus) and CORR is a count of *both* the "strongly agree/disagree" and the "somewhat agree/disagree" responses.

3.1 Results pertaining to H1

Out of our 6 tasks, physicists answered 4.30 tasks correctly on average in the CORR condition (SD=1.12) and 3.44 tasks in the HCON condition (SD=1.23). In other words, although physicists did fairly well, they clearly didn't do perfectly. Could the "disturbing" characteristics of thought experiment explain the imperfect performance of physicists? In order to answer the question, we analyzed the performance of physicists for each task (see Table 1). In the next section we discuss whether these results can be explained by Machery's idea of disturbing characteristics of thought experiments.

	% CORR	% HCON
Stevin's chain	68.4	56.1
Einstein's elevator	59.6	50.9
Schroedinger's cat	26.3	14.0
Newton's cannon	82.3	57.9
Galileo's ship	94.7	70.2
Galileo's tower	98.2	94.7

Table 1: percentages of correct responses in both the CORR and the HCON count for physicists.

3.2 Can disturbing characteristics explain physicists' varied performance?

As mentioned in the introduction of this paper, Machery believes that thought experiments are disturbing because they are (i) 'unusual', (ii) pull apart properties that usually go together and (iii) contain superficial content which is irrelevant to the point under consideration. We also mentioned that thought experiments in physics fulfil these characteristics. Schroedinger's cat is not a scenario many people usually come across. We usually do not consider scenarios in which we are not subject to the influence of the

gravitational force. This is also because these scenarios pull apart properties that usually go together (such as our daily experience of a massive gravitational field exerted by the earth and us dropping an object and observing its trajectory). And as mentioned, thought experiments in physics also contain a lot of irrelevant detail: it matters little whether we consider a cat or a dog in Schroedinger's thought experiment, whether the balls in Strevin's chain are made of metal or marble, whether the two connected objects in Galileo's thought experiment together weigh 10kg or a 100kg.

Contrary what Machery claims about thought experiments in philosophy, the "disturbing characteristics" of thought experiments in our sample cannot explain why some judgements may be more reliable than others. Consider for example Stevin's chain and Newton's cannon. Whereas the former task seems to describe a perfectly normal scenario, the latter makes some rather outlandish assumptions about the power of cannons. In Newton's cannon the properties of being an terrestrial projectile and having realistic terrestrial speed are pulled apart. In contrast, in Stevin's chain it is hard to see which properties are supposed to come apart that usually go together. So if anything, Newton's cannon should be more disturbing than Stevin's chain. And yet, physicists did much better in Newton's cannon than they did in Stevin's chain.

It also appears that different characteristics of thought experiments can pull in different directions. For example, Schroedinger's cat is unusual in that cats are usually not situated like that. On the other hand, it may be fairly familiar to physicists. So is it unusual or not? Lastly, with regard to the third of Machery's criteria, all of our tasks, by virtue of being thought experiments, contain 'irrelevant narrative details'. It's not easy to see how they would be more disturbing in some tasks rather than others.

In sum, the characteristics of thought experiments that Machery has identified as disturbing cannot account for the imperfect and varied performance of physicists. It thus seems unlikely that they play the negative role that Machery has identified for them.

3.3 Results pertaining to H2

In order to test our hypothesis that physics experts are more likely to judge Phy-TEs correctly than non-physicists, we conducted the same test that we used for the physicists with non-physicists. We then determined with a t-test whether the difference in the means of the number of correct judgements by physics experts and the non-physicists was

significant for any of these two counts. For both counts, our H2 was confirmed: physicists on average judged about one more task correctly than the non-physicists in the CORR count and about one-and-a-half task more in the HCON count (see Table 2). The effect sizes for the CORR and HCON count are "large" and "very large", respectively, on the scale suggested by Hyde (2005) and endorsed by Machery (2017).[9]

	Physicists, mean	SD	Non-Physicists, mean	SD	*P*	t
CORR	4.30	1.12	3.37	1.24	.000	-4.22
HCON	3.44	1.23	2.02	1.40	.000	-5.89

Table 2: Average number of correctly answered tasks by physicists and non-physicists out of a total of six tasks.

Since physicists are clearly more reliable than non-physicists in making judgments in Phy-TE, it is not the case that thought experiments constitute environments in which it is unlikely that expertise can be had—contrary to what Machery has argued.

We also conducted negative binomial regression analysis to test for the influence of gender, age, response duration, exposure of controls to physics at university, and level of English (see Appendix 4). None of these factors was significant. That means that the difference we found between the performance of the physicists and the non-physicists cannot be accounted for by at least one of the extraneous factors which have been reported to influence case judgements in the experimental philosophy literature, namely gender (Buckwalter and Stich 2014) (but see Starmans and Friedman 2012, Adleberg et al. 2015, Seyedsayamdost 2015).

Finally, we also analyzed our lay subjects' performance in each of the six tasks: Table 5 lists the percentages of correct judgements for each thought experiment for both CORR and HCON, for both physicists and non-physicists, for comparison. It also lists whether the difference between expert and layperson judgements was significant, which we determined with a chi-squared test (see Table 9 in Appendix 3 for further details).

	% correct in CORR		*p*	% correct in HCON		*p*
	Physicists	Non-physicists		Physicists	Non-physicists	

[9] Cohen's *d* is 0.79 for CORR and 1.08 for HCON.

Stevin's chain	68.4	59.6	ns	56.1	38.6	ns
Einstein's elevator	59.6	54.4	ns	50.9	31.6	*
Schroedinger's cat	26.3	10.3	*	14.0	7.0	ns
Newton's cannon	82.3	68.4	ns	57.9	24.6	***
Galileo's ship	94.7	66.6	***	70.2	42.1	**
Galileo's tower	98.2	77.2	**	94.7	57.9	***

Table 3: percentages of correct responses in both the CORR and the HCON count for both physicists and non-physicists. Statistically significant differences from our χ^2 test at the 95% level are marked with a star (*), at the 99% level (i.e., $p < 0.01$) with a double-star (**) and at the 99.9% level (i.e., $p < 0.001$) with a triple-star (***). See Table 9 in Appendix 3 for details regarding the tests.

Again, as mentioned in the previous section, Stevin's chain should not be very disturbing to subjects (because it does not seem to pull apart properties that usually go together), non-physicists got it wrong more often than Newton's cannon, for example, which seems much more disturbing by Machery's standards. It's also worth mentioning that Schroedinger's cat again stands out from the other tasks in that most subjects in both groups answered contrary to our expectation. This calls for an explanation.

3.4 The oddball of Schroedinger's cat

Before we will try to elucidate the oddball of Schroedinger's cat, let us first provide some further background on the original motivation for this thought experiment, which shaped our expectation of how subjects would respond to this thought experiment.

Schrödinger used his thought experiment to challenge Bohr and Heisenberg's Copenhagen interpretation of quantum mechanics. Schroedinger's reasoning was that if the Copenhagen interpretation were correct, then the cat in the box (in our case: a dog) should be in a state of superposition before the opening of the box (the "measurement") causes a collapse of the wavefunction. However, since we would under normal circumstances judge that the cat/dog does have a definite state before we open the box, the Copenhagen interpretation must be false. Since the Copenhagen interpretation is indeed by far the most accepted interpretation amongst physicists[10], we expected physicists to 'bite the bullet' and accept that the cat/dog actually is in a state of superposition.[11] It turned out, however, that although physicists judged this way significantly more often

[10] For a survey see Schlosshauer et al. (2013).

[11] We were influenced by a standard textbook in the philosophy of physics (Sklar 1992, 184).

than the folk, most physicists instead made a judgment in accordance with the more 'common sensical' assessment. They therefore presumably rejected that the Copenhagen interpretation (contrary to what Schroedinger thought) *entails* the judgement that the cat/dog must be in a state of superposition. From our pilots, we gathered that physicists did this on the basis of making a distinction between micro-objects such as electrons and macro-objects, such as cats and dogs. We take it that such a distinction can be made within the Copenhagen interpretation on the basis of Bohr's correspondence principle, which says (roughly) that the predictions of quantum mechanics approximate those of classical mechanics when it comes to classical objects (such as cats and dogs).

Another reason why physicists did not answer in agreement with our expectations may have to do with the way we phrased our question. The sentence we presented them was**:** "Before the box is opened the dog is either dead or alive." Our expectation was that physicists would judge this statement incorrect, because we took the Copenhagen interpretation to imply that the dog is neither dead nor alive, but rather in a state of superposition. But possibly, physicists parsed the exclusive XOR as a simple OR.

We should note that even when reversing the scales for our Schroedinger cat task so that (strong) agreement with the presented judgement is assumed to be correct (instead of incorrect as in our original analysis), our hypothesis would remain confirmed (see Table 4). The effect sizes for the CORR and HCON count are "moderate" and "large", respectively, on the scale embraced by Machery (2017, 46).[12]

	Physicists, mean	SD	Non-Physicists, mean	SD	*P*	t
CORR	4.60	.979	4	1.363	.003	-2.68
HCON	3.86	1.187	2.68	1.549	.000	-4.55

Table 4: Average of correctly answered questions out of a total of six tasks, with the scale for the Schroedinger cat reversed ($p < .001$ for HCON and $p < .005$ for CORR).

At the same time, it should be pointed out that this scale reversal would result in the non-physicists answering correctly more often than the physicists (see Table 5). This would make the Schroedinger cat the only of our six tasks in which this would be the case.

[12] Cohen's *d* is 0.51 for CORR and 0.86 for HCON.

	% correct in CORR		% correct in HCON	
	Physicists	Non-physicists	Physicists	Non-physicists
Schroedinger's cat (reversed scales)	66.7	82.5	56.1	73.7

Table 5: Percentages of correct answers for Schroedinger's cat with reversed scales in both CORR and HCON. CORR: X^2 (1, 114) =3.746 (p=.053); HCON: X^2 (1, 114) =3.851 (p=.050).

One may ask again: can the disturbing characteristics of Schroedinger's cat explain the result? It looks as though most physicists and non-physicists judged the highly unusual scenario of the thought experiment in a way that one would usually judge a midsized object like a cat. Thus, if the characteristics of thought experiments are disturbing at all, then they don't seem to be very disturbing in this case.

4 Objections

In this section we consider four objections to our results and our interpretation: (i) that we did not create the right circumstances to test the adverse effects of thought experiments, (ii) that physicists did fairly well in our tasks (and better than non-physicists) simply because they recalled the tasks, and (iii) that physicists' performance is entirely explained by them being in possession of the correct physical principles.

4.1 Objection 1: not the right contexts?

An objection one may raise against the interpretation of our results may be that we didn't create the right contexts for bringing out the adverse effects of thought experiments in judgments. In particular, we didn't produce contexts in which we could detect any presentation effects, such as order of presentation, and we did not test subjects different backgrounds, e.g., from different cultural backgrounds.

There are three things we have to say about this objection. First, Machery's claim about thought experiments being not the proper domain of reliable judgments is an unqualified claim, i.e., Machery does not claim that judgements in thought experiments are unreliable only in circumstances in which one tests for the effects of presentation or demographic variables. On the contrary, Machery views thought experiments with disturbing characteristics as *intrinsically* adverse environments for reliable judgments. He

just thinks that experiments investigating the effects of presentation and demographic variables create contexts in which those adverse effects come out most clearly.

Second, we take it that the dialectical situation is such that it is for the armchair critics, and not for us, to demonstrate the effects of extraneous variables on judgements in thought experiments also in physics. Third, even if armchair critics would accept their burden of proof and seek to demonstrate that judgements in Phy-TE are influenced by extraneous variables in the way case judgements of the folk allegedly are, we remain confident. In order for armchair critics to show that judgments in thought experiments with disturbing characteristics are outside the proper domain of *any* subjects, physicists would have be shown to vulnerable to the effects of extraneous variables to such an extent that the difference in reliability between the folk and the experts is nullified.

4.2 Objection 2: just recall?

In response to our results, one could speculate that physicists did fairly well (and better than the folk) simply because they recalled the correct solutions that they've been drilled into. The difference that we found between the folk and the physicists could therefore *not* be interpreted as a sign for the higher level of expertise of physicists in making judgements in thought experiments (contrary to what we suggested here).

There are several reasons against this interpretation. First, our thought experiments are not a very salient part of physicists' training and don't occupy very prominent role in textbooks. If physicists' performance is best explained by their recall abilities, then their performance probably should have been worse than it actually turned out to be. Second, if physicists just recalled the correct textbook answers, then – everything being equal – their performance should be more uniform across the tasks, unless it could be shown that the thought experiments with worse performance occur less frequently in physicists' training than the thought experiments in which our subjects do better. It would be for the critics to support this implication. Finally, as already mentioned, physicists answered the Schroedinger cat thought experiment contrary to what one would expect based on the textbook consensus. This is a problem for the recall hypothesis.

4.3 Objection 3: better physical knowledge?

One may accept our results but deny that they have any implications for thought experiments in philosophy. The reason that physicists are more reliable than the folk, the

objection continues, can be fully explained by the fact that Phy-TE require the knowledge of correct physical principles, whereas that is not the case in philosophical cases. Our results indicate, however, that this is not the case.

First of all, physicists' performance was not perfect: there was a substantial number of physicists who did not answer according to our expectations based on the textbook consensus. So if their better performance was fully determined by their better knowledge of the relevant physical principles, their grasp of those principles wasn't as good as one might expect – particularly given that the tasks were not very demanding in that regard: most of the tasks required only a basic understanding of classical mechanics. Second, if performance really was fully determined by the possession of correct principles, one should expect it to be harder for non-physicists to get right tasks that require more knowledge of physical principles than tasks which require less knowledge of physical principles. Three of our six tasks require some very basic knowledge of classical physics (Stevin's chain, Galileo's ship, Newton's cannon). Although the difference between expert and lay judgements is statistically significant in some tasks (Galileo's ship in CORR and HCON; Newton's cannon in HCON), there is no difference in others (Stevin's chain in both CORR and HCON; Newton's cannon in CORR). On the other hand, tasks in which physicists should have had no major advantage, non-physicists did much poorer than physicists. In particular, Galileo's tower, although an application of classical physics at first sight, is in fact simply a reductio ad absurdum; no knowledge of physics is required. Still, there is a marked difference between the judgements of physicists and non-physicists in both of our count categories.

Turn now to the two tasks which arguably require more advanced knowledge in physics than the other four tasks, namely Schroedinger's cat and Einstein's elevator. In the latter task, there was no significant difference in the CORR count, but there is a significant difference in the HCON count at the 95% level. In the former task, there was a significant difference in the CORR (also at the at the 95% level) but not in the HCON count. Thus, even though the difference between expert and lay judgements should be most pronounced if it were true that the difference is to be explained in terms of the advanced knowledge of physicists, it is either not as pronounced as in other tasks (for example as in Galileo's tower, which is significant at the 99% level) or it doesn't exist in the first place.

In sum, knowledge of the correct physical principles does not fully determine subjects' judgements in Phy-TE. In particular, physicists did not do much better than non-physicists in tasks that would seem to require more physical knowledge than other tasks. Phy-TE and philosophical cases may be more similar than the objection would have it.

5 Conclusion

Our first experiment shows that physicists are reliable when making judgements in thought experiments in physics; they don't seem to be disturbed by the characteristics identified by Machery as intrinsically problematic. This, at least to some extent, undermines Machery's far-reaching skepticism of the method of cases.

Our second experiment shows that physicists are more reliable than non-physicists when it comes to making judgements in thought experiments in physics. This undermines Machery's idea that thought experiments constitute environments in which no expertise is possible. Furthermore, we believe that the results of our second experiment lend support to what has come to be known as the *expertise defense* in philosophy: trained experts make more reliable judgments than the folk. This idea has been used to argue that empirical results demonstrating the unreliability of case judgements by the folk have no bearing on whether or not case judgments by philosophers can be used as (defeasible) evidence in philosophical theorizing (Hales 2006, Ludwig 2007, Horvath 2010, Devitt 2011, Williamson 2011).[13] The expertise defense is usually motivated by an analogy to judgments by philosophers in thought experiments and expert judgments in mathematics, physics, and linguistics, but the analogy has been criticized (Nado 2014, 2015). It may well be that an analogy between philosophers' case judgements and physicists' judgements in thought experiments in physics fairs much better. But this remains to be argued in detail elsewhere.

References

Adleberg, T., M. Thompson, and E. Nahmias. 2015. Do men and women have different philosophical intuitions? Further data. *Philosophical Psychology*, **28** (5): 615-641.

[13] Although there is some evidence that professional philosophers are subject to the influence of extraneous effects, much of the negative program's case (in particular outside the moral realm) rests on studies with the folk (see Machery 2017). So the expertise defense is still a live option.

Alexander, J. and J.M. Weinberg. 2007. Analytic epistemology and experimental philosophy. *Philosophy Compass*, **2** (1): 56-80.
Brown, J.R. and Y.J.H. Fehige. 2011. Thought experiments. *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, http://plato.stanford.edu/archives/fall2011/entries/thought-experiment/.
Buckwalter, W. and S. Stich. 2014. Gender and Philosophical Intuition. In *Experimental Philosophy*, Joshua Knobe and Shaun Nichols (eds.), Oxford: Oxford University Press.
Devitt, M. 2011. Experimental semantics. *Philosophy and Phenomenological Research*, **82** (2): 418-435.
Hales, S.D. 2006. *Relativism and the Foundations of Philosophy*. Cambridge, MA: MIT Press.
Horvath, J. 2010. How (not) to react to experimental philosophy. *Philosophical Psychology*, **23** (4): 447-480.
Horvath, J. and A. Wiegmann. 2015. Intuitive expertise and intuitions about knowledge. *Philosophical Studies*: 1-26.
Hyde, J.S. 2005. The gender similarities hypothesis. *American psychologist*, **60** (6): 581.
Knobe, J. and S. Nichols. 2017. Experimental Philosophy. *The Stanford Encyclopedia of Philosophy (Winter 2017 Edition)*, edited by Edward N. Zalta, https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/.
Ludwig, K. 2007. The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy*, **31** (1): 128-159.
Machery, E. 2011. Thought experiments and philosophical knowledge. *Metaphilosophy*, **42** (3): 191-214.
———. 2017. *Philosophy within its proper bounds*: Oxford University Press.
Nado, J. 2014. Philosophical Expertise. *Philosophy Compass*, **9** (9): 631-641.
———. 2015. Philosophical expertise and scientific expertise. *Philosophical Psychology*, **28** (7): 1026-1044.
Sax, L.J., K.J. Lehman, R.S. Barthelemy, and G. Lim. 2016. Women in physics: A comparison to science, technology, engineering, and math education over four decades. *Physical Review Physics Education Research*, **12** (2): 020108.
Schlosshauer, M., J. Kofler, and A. Zeilinger. 2013. A snapshot of foundational attitudes toward quantum mechanics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, **44** (3): 222-230.
Seyedsayamdost, H. 2015. On gender and philosophical intuition: Failure of replication and other negative results. *Philosophical Psychology*, **28** (5): 642-673.
Sklar, L. 1992. *Philosophy of physics*: Dimensions of Philosophy S.
Starmans, C. and O. Friedman. 2012. The folk conception of knowledge. *Cognition*, **124** (3): 272-283.
Williamson, T. 2011. Philosophical expertise and the burden of proof. *Metaphilosophy*, **42** (3): 215-229.

Worrall, J. 2002. New evidence for old. In *In the Scope of Logic, Methodology and Philosophy of Science*, Peter Gardenfors (ed.), Dordrecht: Kluwer, 191-209.

Appendix 1

	PHYSICS	CONTROL	*p*
Background information	*N* = 57	*N* = 57	
Age (in years)	33 (10.02)	34 (6.72)	ns
Gender (female)	11 (19%)	27 (47.3%)	*
Holds a PhD	32 (56.1%)	33 (57.9%)	ns
Enrolled in a PhD programme	25 (43.9%)	24 (42.1%)	ns

Table 6: Background information about all participants and their education (p=.001).

Holds a PhD in:	*N* = 33
Anthropology	2
Chemistry	2
Economics	1
Engineering	1
History	4
Languages	1
Literature	1
Mathematics	1
Medicine	2
Musicology	1
Political science	11
Psychology	1
Philosophy	3
Sociology	2
Enrolled in a PhD programme	N = 24
Business & managements	2
Computer Science	2
History	1
Linguistics	2

Literature	1
Mathematics	4
Medicine	1
Political science	8
Psychology	2

Table 7: Information about the education of participants from the CONTROL group.

Specialization in PHYSICS group	*Number of participants who chose the answer*
Astrophysics and cosmology	9
Atomic- molecular- and optical physics	24
Biophysics	1
Solid-state- and materials physics	11
Sub-atomic physics	8
Nano physics	8
Statistical physics	5
Other*	11

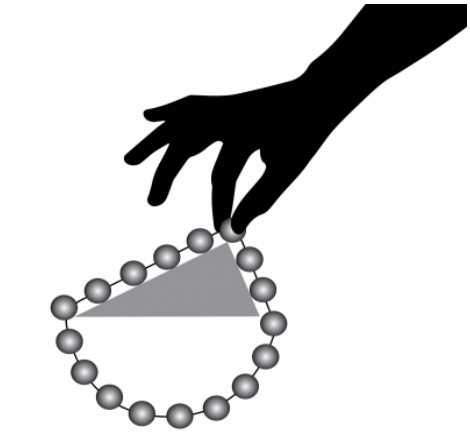
Table 8: Areas of specialisation of physicists via self-identification (multiple answers allows). Under "Other" subjects stated: Particle physics (3); Quantum physics (4); Fluid mechanics (1); Condensed matter physics (1); Nuclear physics (1) Applied physics and methodology (1).

Appendix 2

In our six tasks, we asked subjects to consider a scenario (S), answer a comprehension question (CS), and say whether they would agree with the judgements offered (J).

Stevin's chain

S: "Imagine that somebody put a chain with evenly spaced metal balls with the same size and weight on top of an inclined frictionless plane."



CS: "The inclined plane in the above scenario is …"

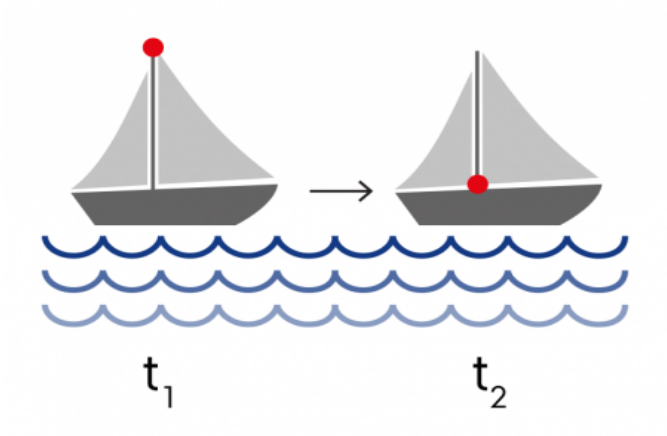
J: "Once the chain is released it will move sideways." [This is incorrect.]

J is the negation of the judgment elicited in a famous thought experiment by Simon Stevin. With this thought experiment Stevin wanted to demonstrate the plausibility of his claim that for inclined planes with the same height, the force needed to keep weights in their position on those planes varies inversely with the planes' lengths. More specifically, in the depicted scenario S Stevin used a pair of planes of which one was double the length of the other and the weights placed on the longer plane were double the amount of weights on the shorter plane. According to his law, the weights on those two planes (which are connected to each other) should balance each other out. In order to drive home the point, Stevin connected the weights on those two planes with a chain of further weights (seen at the bottom of the figure). Now, if one were to deny Stevin's 'law' and approve of the statement that the entire chain moves to the right or to the left (as in J), it's not clear how one could deny that the chain *keeps* moving either to the right or left. After all, the chain is uniform (equal weights, equal distances between the weights). But since this would constitute a perpetual motion, which is ruled out by the 2nd law of thermodynamics. Ernst Mach, who discussed this task in his *The Science of Mechanics*, ruled this out on an "instinctive basis".

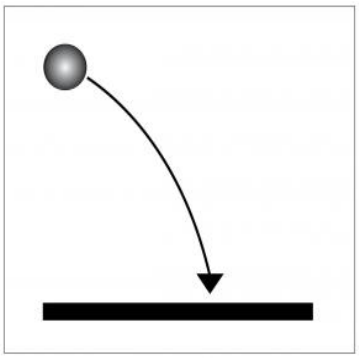
Galileo's ships

S: "Imagine yourself standing at the coast and observing a ship moving with constant speed. The picture shows a snapshot of the ship's movement at two points in time: t1 and

t2. At t1, a cannon ball is dropped from the top of the mast of the ship and at t2 the cannon ball has reached its final position:"



CS: "As the observer you are located on …"
J: "When seen from the coast, the trajectory of the ball moving from t1 to t2 is as in the following picture:" [This is correct]



Galileo used this thought experiment in his *Dialogue Concerning the Two Chief Systems of the World* to persuade those believing in the geocentric system that a moving earth would not necessarily pose any problems for terrestrial physics (people were concerned that a moving earth would imply objects on earth flying through the air). The object falling from the top of the mast to its bottom on a moving ship illustrates that the trajectory of falling objects may appear straight when it in fact decomposes into straight and rectilinear motion (as in our second picture). Galileo's ship also demonstrates what has come to be known as Galilean relativity: the classical laws of physics are the same in all inertial frames (and two inertial frames can be transformed into each other via Galilean transformations).

Newton's cannonball

S: "Imagine shooting a cannonball from a high elevation on earth into the distance. On the picture, you see the trajectories of a cannonball shot with (relatively) low speed, A, and with a higher speed, B. Cannon balls following A and B will land back on earth."

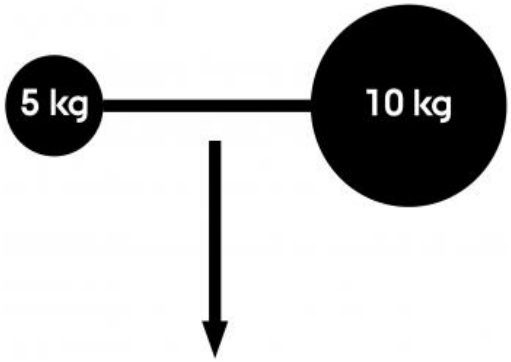
CS: "The cannonball following trajectory B will land on …"

J: "Trajectory C is possible." [This is correct]

Newton used this thought experiment in the *The System of the World* to show that the orbital motion of the moon (and the planets around the sun) is accounted for by the same forces that act on earth (namely an inertial and a gravitational one).

Galileo's tower

S: "Imagine you connect a steel ball of 10kg and a steel ball of 5kg with a tight chain and drop the combined object from a high elevation in a vacuum. How does one determine the speed of fall of the combined object? One proposal is to average the speed of the two objects (when they fall separately): since 5kg falls slower than 10kg, the combined object will fall slower than the 10kg ball. Another proposal is to add the weights: and since 15kg > 10kg, the combined object will fall quicker. Yet another proposal is that the combined object falls just as fast as the 10kg ball on its own, since the weight makes no difference to the speed of fall."



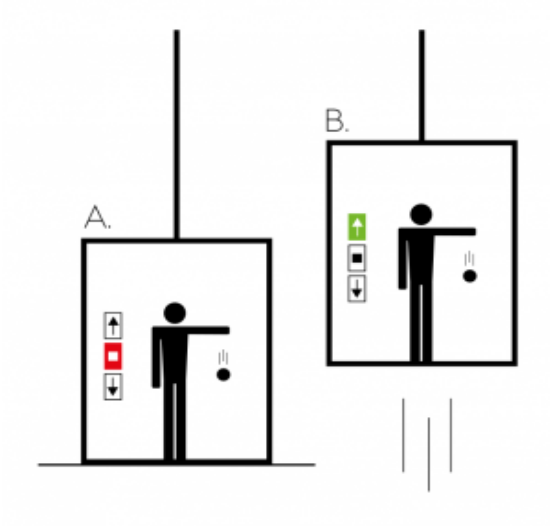
CS: "The combined object weighs …. kg."

J: "The combined object will land just as fast on the ground as the 10 kg steel ball alone". [This is correct]

This is another thought experiment by Galileo, expounded in his *Dialogues concerning two new sciences*. Galileo used this thought experiment to demonstrate an internal contradiction in Aristotle's physics, according to which heavier bodies fall quicker to the ground than lighter ones: in situations such as the one described, Aristotelian physics implies a contradiction, namely that both the combined object falls quicker *and* slower than the heavy object alone. On the basis of this thought experiment (and other evidence), Galileo argued not only that Aristotelian physics is false, but also that all bodies fall at the same rate (which he could not demonstrate at the time, as he had no means for producing vacuums).

Einstein's Elevator

S: "Consider a person in the scenarios A and B. In A, the person is standing inside an elevator that sits on the ground level. In B, the person is inside an elevator that is dragged through empty space somewhere in the universe with uniform acceleration (i.e., the speed increases constantly). In neither A or B can the person see what's going on outside the elevator. In B, the person does not feel that the elevator is being dragged: the elevator appears perfectly stable to her. Suppose that the person wants to find out whether she is in A or B by dropping a ball to the floor."



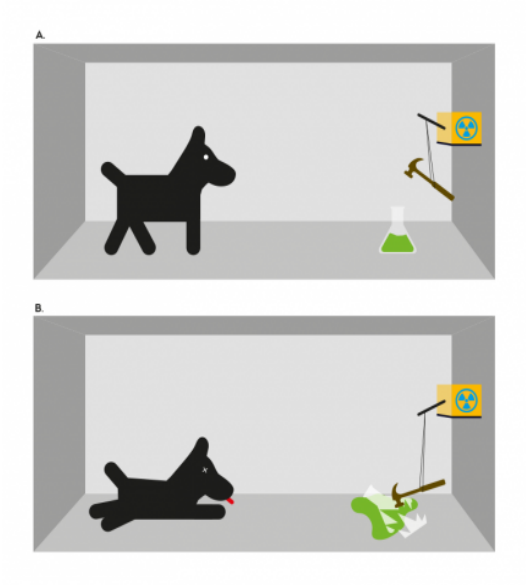
CS: "In B, the elevator is dragged through…"

J: "The person can determine whether she is in A or B by the manner in which the ball drops to the floor." [This is incorrect]

Einstein (and Infeld) used this thought experiment to illustrate the equivalence between inertial and gravitational forces, which underlies the general theory of relativity. The trajectories of the balls will only then be indistinguishable in the two scenarios if the acceleration equals the strength of gravity on the surface of the earth. This is suggested in the thought experiment by the person in the elevator "not feeling" any drag.

Schrödinger's cat

S: "Imagine a dog trapped in an opaque box. There is a very small amount of radioactive substance in the box: there is a probability of 50% that one atom of that substance decays within one hour. Whenever one atom of this substance decays, a Geiger counter will detect this atom and trigger the destruction of a flask containing a highly toxic substance. As soon as the flask breaks, the dog dies instantly. Suppose that the dog is kept in the box for one hour before the box is opened."



CS: "If one atom decays, then the dog will …"

J: "Before the box is opened the dog is either dead or alive." [Our expectation was that physicists should judge this as incorrect]

Erwin Schrödinger used this thought experiment to challenge Bohr and Heisenberg's Copenhagen interpretation of quantum mechanics.

The wave function of quantum mechanics describes the system in terms of probabilities. According to the Copenhagen interpretation, the probability of the state of a physical system at some point in time describes the actual system and is not just an expression of our own ignorance. The system is also said to be in a "superposition" of states. When we measure the system is said to "collapse" and the system adopts a definite state. Which state the system actually adopts upon measurement, however, cannot be determined within quantum mechanics.

Schroedinger's reasoning was that if the Copenhagen interpretation were correct, then the cat in the box (in our case: a dog) should be in a state of superposition before the opening of the box (the "measurement") causes a collapse of the wavefunction. However, since we would normally judge that the cat/dog does have a definite state before we open the box, the Copenhagen interpretation must be false. As explained in the main text, there are legitimate ways of avoiding this conclusion. In our analysis we presumed that the correct response in this task would be the denial of J (but see above).

Appendix 3

Question-by-question chi-square tests in the HCON count for correctly answered questions by physicists vs. non-physicists.

	CORR		HCON	
	X^2	*p*	X^2	*p*
Stevin's chain	X^2 (1, 114) =0.9522	.329	X^2 (1, 114) =3.5185	.061
Einstein's elevator	X^2 (1, 114) =0.3221	.570	X^2 (1, 114) =4.3804	.036*
Schroedinger's cat	X^2 (1, 114) =4.7281	.030*	X^2 (1, 114) =1.4902	.222
Newton's cannon	X^2 (1, 114) =3.0299	.082	X^2 (1, 114) =12.0689	.000***
Galileo's ship	X^2 (1, 114) =14.4190	.000***	X^2 (1, 114) =9.1200	.003**
Galileo's tower	X^2 (1, 114) =11.75257	.001**	X^2 (1, 114) =21.4023	.000***

Table 9: Chi-square tests in the HCON and CORR count comparing physicists and non-physicists for each of our thought experiments. Statistically significant differences at the 95% level (i.e., $p < 0.05$) are marked with a star (*) at the 99% level (i.e., $p < 0.01$) with a double-star (**) and at the 99.9% level (i.e., $p < 0.001$) with a triple-star (***).

Appendix 4

Dependent Variable: Total Strictly Correct			
Variable	Model 1	Model 2	Model 3
Physics (dichotomous)	0.55*** (0.12)	0.49* (0.13)	0.41** (0.15)
Marginal Effect of Physics (How many more questions do physicists get right on average)	1.45***	1.34***	1.52** if female 1.22** if non-female
Physics * Female			0.27 (0.28)
Exposure to physics at university (but no bachelor degree) (dichotomous)		0.19 (0.25)	0.16 (0.25)
Age		0.004 (0.006)	0.005 (0.006)
Female		-0.31* (0.14)	-0.44* (0.20)
Natural log of duration (seconds)		-0.01 (0.07)	-0.004 (0.07)
Degree of English proficiency (1=intermediate, 2=advanced, 3=native)		-0.12 (0.12)	-0.12 (0.13)
Constant	0.670*** (0.09)	0.97*** (0.59)	0.99 (0.59)
Log Likelihood	-198	-194	-194
Ln_Alpha	-27.09	-27.09	-27.09
Alpha	1.71 e-12	1.71 e-12	1.71 e-12

Table 10: Negative binomial regression analysis for the HCON count. According to Model 1 the marginal effect of having a PhD degree in physics (or studying towards one) is 1.45. Thus, having a physics degree is predicted to increase the number of questions answered correctly by 1.45. Model 2 includes control variables for gender, exposure of controls to physics at university, age, duration of task performance, and level of English. Model 3 interacts the factor 'women' with physicists and non-physicists. It predicts that women with a PhD degree in physics (or ones studying towards one) judge 1.52 tasks more correctly than women in the control group. *denotes $p < 0.05$ (95% statistically significant); ** denotes $p < 0.01$ (99% statistically significant); *** denotes $p < 0.001$ (99.9% statistically significant). Numbers in the table represent the difference in the logs of expected counts in the response variable for a one-unit change in the independent variable.

Dependent Variable: Total Lax Correct			
Variable	Model 1	Model 2	Model 3
Physics (dichotomous)	0.24** (0.10)	0.23* (0.11)	0.22 (0.13)
Marginal Effect of Physics (How many more questions do physicists get right on average)	0.93*	0.89*	0.79 if female 0.91 if non-female
Physics * Female			0.01 (0.23)
Exposure to physics at university (but no bachelor degree) (dichotomous)		0.23 (0.19)	0.23 (0.19)
Age		-0.002 (0.006)	-0.002 (0.006)
Female		-0.18 (0.11)	-0.19 (0.15)
Natural log of duration (seconds)		0.04 (0.06)	0.04 (0.06)
Degree of English proficiency (1=intermediate, 2=advanced, 3=native)		-0.03 (0.10)	-0.03 (0.10)
Constant	0.21*** (0.07)	1.14* (0.49)	1.14* (0.49)
Log Likelihood	-203	-201	-201
Ln_Alpha	-21.47	-21.47	-21.47
Alpha	4.75 e-10	4.75 e-10	4.75 e-10

Table 11: Negative binomial regression analysis for the CORR count. According to Model 1 the marginal effect of having a PhD degree in physics (or studying towards one) is 0.93. Model 2 includes control variables for gender, exposure of controls to physics at university, age, duration of task performance, and level of English. Model 3 interacts the factor 'women' with physicists and non-physicists. It predicts that women with a PhD degree in physics (or ones studying towards one) judge .89 tasks more correctly than women in the control group. *denotes $p < 0.05$ (95% statistically significant); ** denotes $p < 0.01$ (99% statistically significant); *** denotes $p < 0.001$ (99.9% statistically significant). Numbers in the table represent the difference in the logs of counts in the response variable for a one-unit change in the independent variable.