On Gender and Philosophical Intuition: Failure of Replication and Other Negative Results[1]

Hamid Seyedsayamdost
University of London

This draft was submitted to *Philosophical Psychology* (Taylor and Francis Publishing Company) on May 22, 2013 and is currently under review.

Abstract

In their paper titled *Gender and Philosophical Intuition*, Wesley Buckwalter & Stephen Stich argue that the intuitions of women and men differ significantly on various types of philosophical questions. Furthermore, men's intuitions, so the authors, are more in line with traditionally accepted solutions of classical problems. This inherent bias, so the argument, is one of the factors that leads more men than women to pursue degrees and careers in philosophy. These findings have received a considerable amount of attention and the paper is to appear in the second edition of *Experiment Philosophy* edited by Joshua Knobe & Shaun Nichols, which itself is an influential outlet. Given the exposure of these results, we attempted to replicate three of the classes of questions that Buckwalter & Stich review in their paper and for which they report significant differences. We failed to replicate the results using two different sources for data collection (one being identical to the original procedures). Given our results, we do not believe that the outcomes from Buckwalter & Stich (forthcoming) that we examined are robust. That is, men and women do not seem to differ significantly in their intuitive responses to these philosophical scenarios.

1. Introduction and Overview

In their paper titled *Gender and Philosophical Intuition*[2] Wesley Buckwalter & Stephen Stich approach the issue of gender disparity in the academic field of philosophy from a novel perspective. The authors argue that women's and men's intuitions differ in various areas of philosophy and more importantly that men's intuitions are more in line with commonly accepted solutions of classical philosophical problems. This inherent bias, so the authors, is one of the factors that leads more men than women to pursue degrees and careers in philosophy.

In supporting their claims, the authors review some recent findings in experimental philosophy (section 3 of their paper) and also present new data for four classical scenarios in which they report men and women to respond differently to survey questions (section 3.8). The thought experiments in this section (3.8) include the Brain in the Vat, Hilary Putnam's Twin Earth, John Searle's Chinese Room and the Plank of Carneades. These cases are of special interest to Buckwalter & Stich because these are cases that undergraduate students typically encounter early on in introductory philosophy classes. Hence, so the authors argue, if women's responses differ from commonly accepted solutions of philosophical problems, women could be discouraged from pursuing further philosophy courses. In addition to the scenarios of section 3.8, in section 3.2 Buckwalter & Stich present results on Compatibalism, Physicalism and Dualism cases where women and men are also reported to answer questions differently. We attempted direct replications of sections 3.2 and 3.8 of Buckwalter & Stich (forthcoming) and our results indicate that the outcomes reported by Buckwalter & Stich are not robust.

Furthermore, in section 3.1 Buckwalter & Stich report differences between men and women for two variations of a Gettier style scenario. We had collected data on four Gettier type scenarios for another study and analyzed the results to see how women and men answered the

questions. Although this is a conceptual replication, we believe the results to be relevant for this paper. Once again, our data showed no difference between the two groups of respondents.

Apart from the straight replication of the scenarios mentioned above, we also wanted to address what we believed to be a shortcoming in Buckwalter & Stich's choices of samples. For their statistical analyses Buckwalter & Stich restricted their data to respondents who had not taken any philosophy courses before. This is because the authors aimed to test unbiased responses. That is, responses that had not been influenced by previous study of the cases, which would have likely been 'male-centrist'. However, by filtering in this way Buckwalter & Stich tested samples of individuals who had no interest or perhaps possibility to pursue philosophy as a degree or career in the first place. Hence, the sample may not adequately represent the pool of students who set out for careers in philosophy.

To address this issue, we wanted to analyze individuals who had taken at least some philosophy courses but whose views had not been biased by previous study of the cases. We collected information on how many courses participants had taken and whether they had seen the scenarios before. In this way, we could evaluate the answers of respondents who had been interested enough to take at least some philosophy classes and may have pursued philosophy as a career but had not seen and not been familiar with these scenarios.[3] Here again, we failed to detect a difference between men and women.

We do not disagree with the general point Buckwalter & Stich (forthcoming) attempt to make. Intuitive responses to survey questions may differ among men and women for certain problems and this may lead more women or men to pursue certain fields and careers. However, if much rests on the results that Buckwalter & Stich present in section 3 of their paper, then the failure of replication weakens their argument. Buckwalter & Stich suggest that differences in intuitions may be one factor among many that influence career choices in philosophy and if our

results are robust, this factor plays a smaller role than the findings in Buckwalter & Stich (forthcoming) suggest.

Our main aim for this paper is to share our results with others, especially researchers who may want to build on the results of Buckwalter & Stich (forthcoming). It is because of this focus that we will keep the discussion on the role and importance of intuitions in philosophical endeavors to a minimum. Furthermore, others have provided a better overview and discussion on the issue of intuitions than we can present here.[4] Throughout this paper we will make frequent references to Buckwalter & Stich (forthcoming) and following our sections may be easier if readers have some familiarity with the original paper.

This paper is structured as follows. In the next section of this paper we will examine the classical philosophical scenarios presented in section 3.8 of Buckwalter & Stich (forthcoming). Specifically, in section 2.1 we will present the data for the replication of the experiments and in section 2.2 we will present the data on participants who had taken some philosophy classes but who indicated that they had not seen the cases before. In section 3 we will examine the scenarios for Compatibalism, Physicalism and Dualism that Buckwalter & Stich describe in their section 3.2. In section 3.1 we will present the replication and in section 3.2 we will examine the sample of respondents with some philosophy background but who were not familiar with the scenarios. In section 4 we will present the data for Gettier scenarios. In the final section we will provide some concluding remarks and a brief discussion on the possible reasons for why replication failed.

2.1: Brain in the Vat, Twin-Earth, Chinese Room and Plank of Carneades

For this section we collected data though two different sources. The first was through Amazon's Mechanical Turk (MT) following the methodology in Buckwalter & Stich (forthcoming). For

the second data set we ran surveys on SurveyMonkey (SM). We will describe the procedures of data collection for all data sets first and then present the results in the subsequent sub-section. This way we can compare the outcomes more readily.

Procedures and Methods of Data Collection

Mechanical Turk

We tried to follow Buckwalter & Stich's methodology as closely as possible. In the Human Intelligence Task (HIT) description respondents were given some brief information about what the task entailed, the approximate time needed to complete the task and some other information required by MT. Once participants accepted a task they were shown one of the four scenarios presented in section 3.8 of Buckwalter & Stich (forthcoming). The scenario was followed by a comprehension check question (the same that was asked in the original paper) and a question asking for a response on a seven-point scale. The one difference we made to Buckwalter & Stich's outline is the inclusion of another question asking whether respondents had seen the scenario before. We included this question for two reasons. First, as mentioned in the introduction we wanted to test participants with a background in philosophy but who were not familiar with these scenarios. Second, Buckwalter & Stich had run these same scenarios on MT and we wanted to be able to exclude respondents who may have had seen these cases in a run conducted by Buckwalter & Stich.

Following these three questions there was a brief demographic questionnaire where we asked about gender, age, education, number of philosophy courses taken, native language, ethnic background, level of religiosity and income in this order. Finally, we also had a section where participants could leave comments.

SurveyMonkey

Our second data set was collected through SurveyMonkey (SM). We collected data in two different runs conducted about six months apart. We believe the surveys to be similar enough that aggregating the data is unproblematic, however, we will also present the breakdown for each survey. The main difference between the two surveys was the number of scenarios presented to participants. In the first survey participants saw eight scenarios pseudo-randomized, whereas in the second data set participants only saw four scenarios. With the exception of one case, the questions were the same in both surveys just that in the shorter version the scenarios were split up into two different questionnaires. Each question in a survey was shown on a new page and the setup of the questions was the same as in MT. The demographic section was more comprehensive in the first SM survey. Survey invitations were sent out to the general population within the United States. For more information of participation details, see https://contribute.surveymonkey.com/how-it-works.

Results

Before we report our results, we will briefly present summaries of Buckwalter & Stich's outcomes in order to make comparison easier.

Brain in the Vat: Original Results

The first case that Buckwalter & Stich present in their section 3.8 is the Brain in the Vat scenario. The exact wording is as follows:

> George and Omar are roommates, and enjoy having late-night 'philosophical' discussions. One such night Omar argues, "At some point in time, like, the year 2300,

the medical and computer sciences will be able to simulate the real world very convincingly. They will be able to grow a brain without a body, and hook it up to a supercomputer in just the right way so that the brain has experiences exactly as if it were a real person walking around in a real world, talking to other people. The brain would believe it was a real person walking around in a real world, except that it would be wrong. Instead it's just stuck in a virtual world, with no actual legs to walk and with no other actual people to talk to. And here's the thing: how could you ever tell that it isn't really the year 2300 now, and that you're not really a virtual-reality brain? If you were a virtual-reality brain, after all, everything would look and feel exactly the same to you as it does now! George thinks for a minute, and then replies: "But, look, here are my legs". He points down to his legs. "If I were a virtual-reality brain, I wouldn't have any legs really, I'd only just be a disembodied brain. But I know I have legs, just look at them! So I must be a real person, and not a virtual-reality brain, because only real people have real legs. So I'll continue to believe that I'm not a virtual-reality brain."

George and Omar are actually real humans in the actual real world today, and so neither of them are virtual-reality brains, which means that George's belief is true.[5]

Following the scenario and a comprehension check question participants were presented with the sentence, "George knows that he is not a virtual-reality brain." Subsequently participants were asked to indicate their level of agreement/disagreement on a seven-point scale where the leftmost option was marked "Completely Disagree" the midpoint labeled "In Between" and the rightmost option marked "Completely Agree" (Completely Disagree = 1, In Between = 4, Completely Agree = 7).

Buckwalter & Stich report for N = 63 (Male = 24, Female = 39) a mean male score of 5.62 (SD = 1.97) and a female mean score of 6.72 (SD = 0.76). An independent-samples t-test comparing men and women yielded $t(61) = -3.12$ with $p < 0.01$ and $d = 0.81$[6]

Brain in the Vat: Replication Results

Mechanical Turk

For our data analysis we used the same filters as Buckwalter & Stich and excluded participants if they 1) answered the comprehension check question incorrectly, 2) finished the questionnaire in less than 30 seconds, 3) their native language was not English and 4) had taken some philosophy courses.

Our data for a sample of 114 individuals (58 Female and 56 Male) resulted in a mean score of 5.25 (SD = 2.24) for men and a mean score of 5.86 (SD = 1.85) for women. We conducted an independent samples t-test for men's and women's responses which yielded: $t(107) = -1.59$ (equal variance not assumed), $p = 0.115$.[7] Despite a sample that was close to twice as large as that of Buckwalter & Stich we did not detect a difference at the 10% level.

SurveyMonkey

The overall result for the Brain in the Vat scenario from our SurveyMonkey data is as follows. N = 100 (Male = 51, Female = 49). Male: Mean = 5.78, SD = 1.86. Female: Mean = 5.61, SD = 1.82. An independent-samples t-test comparing men and women yielded: $t(98) = 0.455$), $p = 0.650$.[8]

What stands out from the three data sets is the high value for women's mean response (6.72) in Buckwalter & Stich (forthcoming). Respondents typically have an aversion to selecting

the extreme points on Likert-type scales. Following is a visual presentation for the outcomes of the three procedures.[9]

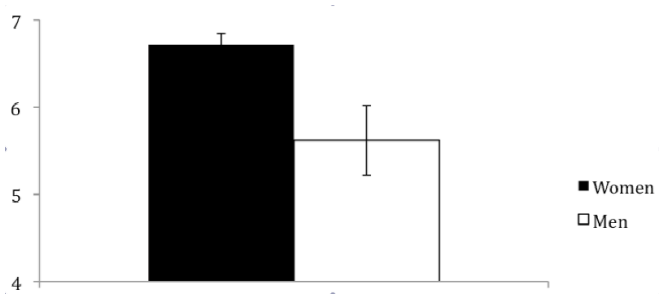

Fig. 1a: Brain in the Vat – Buckwalter & Stich (forthcoming)[10]

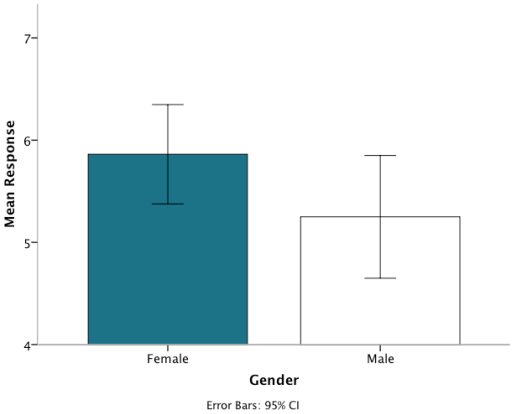

Fig. 1b: Brain in the Vat – Mechanical Turk

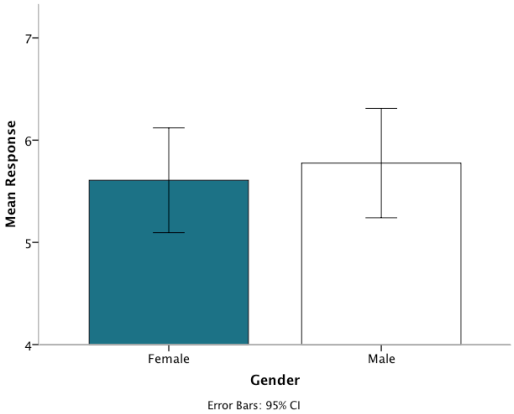

Fig. 1c: Brain in the Vat – SurveyMonkey

Twin Earth

Next, Buckwalter & Stich (forthcoming) present results for the Twin Earth scenario. The exact wording reads as follows:

> Suppose that elsewhere in the universe there is a planet called "Twin-Earth". Twin-Earth looks exactly like our Earth in virtually all respects. It is populated by twin equivalents to every person and thing here on our Earth, and even revolves around a star that appears to be exactly like our sun.

> Oscar grows up here on our Earth, while someone exactly like Oscar, who we can call "Twin-Oscar", lives on Twin-Earth. Oscar and Twin-Oscar both go through life having the same experiences, and both perceive their environment in exactly the same way. They look and act completely alike, and even experience the same emotions.
>
> In fact, there is only one difference between these two planets. The difference is that on Earth the stuff that fills the lakes and rivers and that people and animals drink is H2O, while on Twin Earth, the stuff that fills the lakes and rivers and that people and animals drink is another chemical compound, XYZ, that to the naked eye looks completely indistinguishable from the H2O on Earth. H2O and XYZ also taste exactly the same, and both have the ability to quench thirst and to sustain life.
>
> However, Oscar and Twin-Oscar both live before the development of modern science, and they have no idea about chemistry or molecular composition. When they go for a swim, both Oscar and Twin-Oscar point to the liquid in the lake and call it "water" even though on Earth that liquid is made up of H2O, and on Twin-Earth it is made up of XYZ.[11]

After reading the scenario and answering a comprehension check question, participants were asked the following question:

> When Oscar and Twin-Oscar say "water" do they mean the same thing, or different things?

Participants then entered their response on a seven-point scale where the leftmost option was marked "they mean different things", the midpoint labeled "in between" and the

rightmost option marked “they mean the same thing” (they mean different things = 1, in between = 4, they mean the same thing = 7).

Twin Earth: Original Results:

The outcome reported by Buckwalter & Stich is the following: N = 84 (Male = 35, Female = 49). Male: Mean = 5.63, SD = 2.21. Female: Mean = 4.49, SD = 2.42. Independent-samples t-test: $t(82) = 2.21$, $p < 0.05$. $d = 0.49$

Twin Earth: Replication Results

Mechanical Turk

In our MT sample there was no significant difference among men and women, and in fact women had a higher average mean than men. We used the same criteria as in the Brain in the Vat case to exclude participants from analysis: N = 117 (Male = 65, Female = 52). Male: Mean = 5.22, SD = 2.35. Female: Mean = 5.46, SD = 2.11. Independent-samples t-test: $t(115) = -0.589$, $p = 0.557$.

SurveyMonkey

The sample we collected through SurveyMonkey also did not yield a significant difference among women and men on the standard cut off points: N = 85 (Male = 40, Female = 45). Male: Mean = 5.88, SD = 2.07. Female: Mean = 5.22, SD = 2.57. Independent-samples t-test: $t(82) = 1.30$ (equal variances not assumed), $p = 0.20$. Below is a graphical presentation for the outcomes of the Twin Earth procedures.

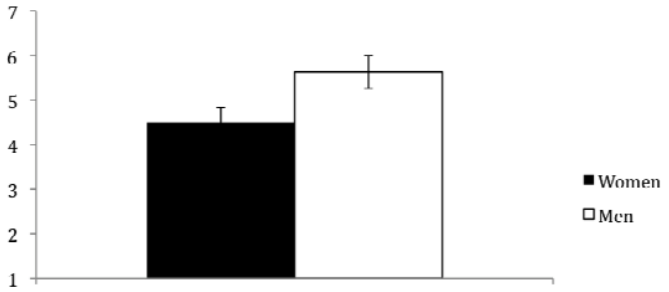

Fig. 2a: Twin Earth – Buckwalter & Stich (forthcoming)

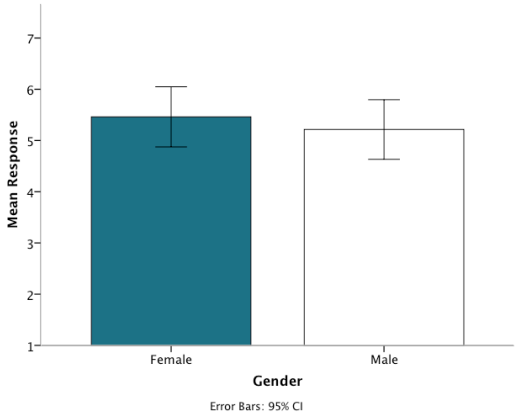

Fig. 2b: Twin Earth – Mechanical Turk

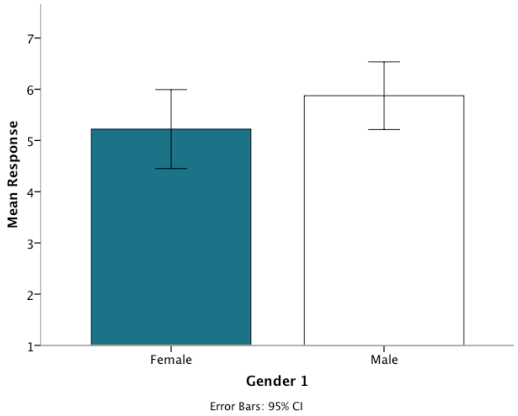

Fig. 2c: Twin Earth – SurveyMonkey

Chinese Room

The Chinese Room scenario was presented to individuals in the following way:

> Jenny is a native English speaker who can only speak English. She is locked in a room full of boxes of Chinese symbols, together with an instruction manual written in English for manipulating the symbols. People from outside the room send in notes on pieces of paper with Chinese symbols written on them, which unknown to Jenny, are questions in Chinese. Jenny's job is to look through her manual until she finds the symbols that look exactly like the ones written on the pieces of paper. When she finds that string of symbols, the manual will tell her

> what new string of symbols to write down, and send to the people outside the room.
>
> By following the instructions in the manual, Jenny is able to give the correct answers to the questions. The system consisting of Jenny and the instruction manual that she is using can be thought of as an unusual sort of computer. Jenny gets so good at following the instructions in the manual, that from the point of view of any one outside the room who speaks Chinese, her responses are absolutely indistinguishable from those of Chinese speakers.[12]

After reading the scenario and answering a comprehension check question, participants saw the statement

> The computational system consisting of Jenny and her instruction manual understands the Chinese written on the notes.

Respondents were then asked to indicate their level of agreement/disagreement on a seven-point scale identical to the one displayed in the Brain in the Vat scenario where the leftmost choice was labeled “Completely Disagree” the midpoint was marked “In Between” and the rightmost option was labeled “Completely Agree” (Completely Disagree = 1, In Between = 4, Completely Agree = 7).

Chinese Room: Original Results

Buckwalter & Stich report for N = 110 (Male = 37, Female = 73) Male: Mean = 4.95, SD = 2.07. Female: Mean = 5.64, SD = 1.35 (d = 0.42). Independent-samples t-test: $t(108) = -2.13$, $p < 0.05$.

Chinese Room: Replication Results

Mechanical Turk

There was no difference in our MT sample for the Chinese Room thought experiment. In fact both group means were identical at 3.31. The details are as follows: N = 103 (Male = 48, Female = 55). Male: Mean = 3.31, SD = 2.19. Female: Mean = 3.31, SD = 2.02. Independent Samples t-test: $t(101) = 0.008$, $p = 0.993$.

SurveyMonkey

There was no significant difference in our SurveyMonkey sample either: N = 80 (Male = 35, Female = 45). Male: Mean = 3.66, SD = 2.59. Female: Mean = 3.82, SD = 2.38. Independent Samples t-test: $t(78) = -0.296$, $p = 0.768$. For a graphical presentation of the outcomes, see below.

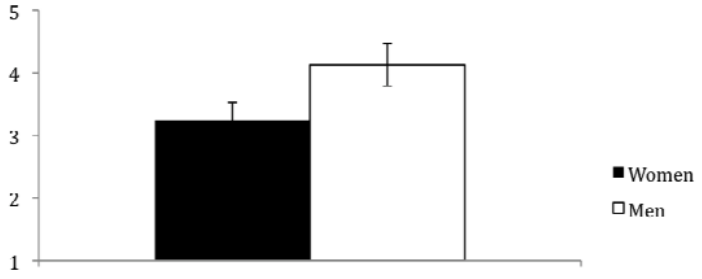

Fig. 3a: Chinese Room – Buckwalter & Stich (forthcoming)

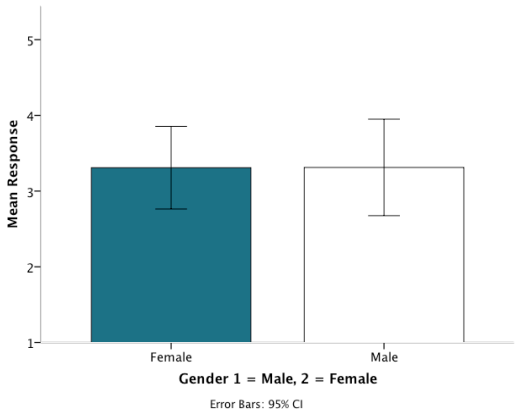

Fig. 3b: Chinese Room – Mechanical Turk

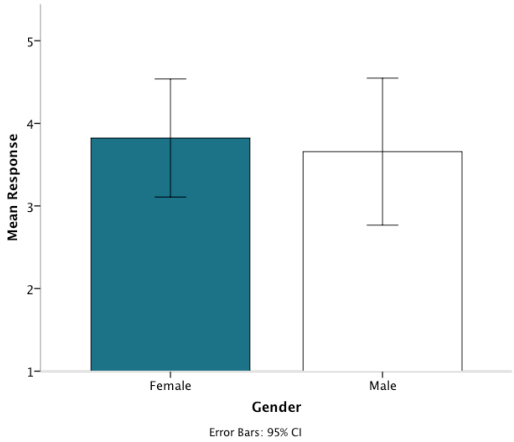

Fig. 3c: Chinese Room – SurveyMonkey

Plank of Carneades

The Plank of Carneades scenario participants were asked to consider was as follows:

> There are two shipwrecked sailors, Jamie and Ricki. They both see a small plank that can only support one of them and both of them swim desperately towards it. Jamie gets to the plank first. Ricki, who is stronger and is going to drown, pushes Jamie off and away from the plank and, thus, ultimately, causes Jamie to drown. Ricki gets on the plank and is later saved by a rescue team.[13]

After responding to a comprehension question participants were asked, "How morally blameworthy is Ricki for what he did?"

Participants answered on a seven-item scale, with the leftmost anchor labeled "not at all blameworthy" the midpoint labeled "in between" and the rightmost anchor labeled "extremely blameworthy" (not at all blameworthy = 1, in between = 4, extremely blameworthy = 7).

Plank of Careades: Original Results

Buckwalter & Stich report for N = 110 (Male = 37, Female = 73). Male: Mean = 4.95, SD = 2.07. Female: Mean = 5.64, SD = 1.35 (d = 0.42). Independent Samples t-test: $t(108) = -2.13$, $p < 0.05$.

Plank of Carneades: Replication Results

Mechanical Turk

Our MT data yielded no significant difference for N = 156 (Male = 70, Female = 86). Male: Mean = 5.20, SD = 1.55. Female: Mean = 5.51, SD = 1.44. Independent Samples t-test: $t(154) = -1.302$, $p = 0.195$.

SurveyMonkey Data:

Similarly with the SurveyMonkey data, our sample showed no significant difference: N = 98 (Male = 48, Female = 50). Male: Mean = 5.85, SD = 1.46. Female: Mean = 5.62, SD = 1.71. Independent Samples t-test: $t(96) = 0.727$, $p = 0.469$. For a graphical presentation, see below.

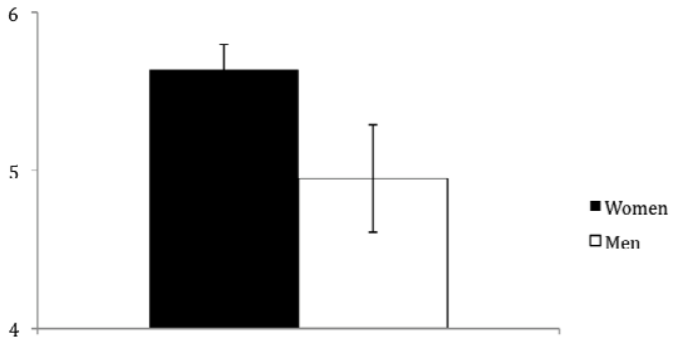

Fig. 4a: Plank of Carneades – Buckwalter & Stich (forthcoming)

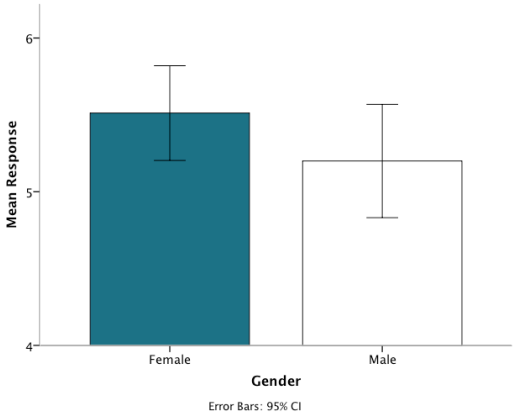

Fig. 4b: Plank of Carneades – Mechanical Turk

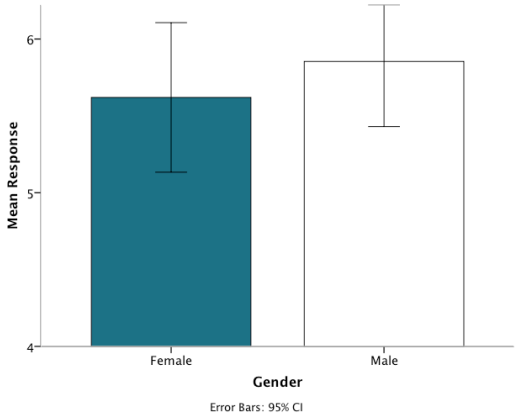

Fig. 4c: Plank of Carneades – SurveyMonkey

Given that we had collected data on whether respondents had seen the scenarios before, we also carried out statistical analyses excluding participants who had seen the scenarios prior to participating in our surveys. An independent-samples t-test for the two groups yielded a significant difference for the Brain in the Vat scenario only. The other scenarios remained non-significant. For the details of the tests, see Appendix A.

2.2: Some Philosophy Background but Not Seen Cases Before

As mentioned in the introduction we believe that the respondents Buckwalter & Stich selected for their analysis is not quite adequate. The reason is that anyone who had taken at least one or more philosophy courses was excluded from analysis. This leaves a

sample of respondents who never had an interest or perhaps possibility to pursue philosophy in an academic setting.

In the context of Buckwalter & Stich's discussion on who chooses to pursue philosophy as a degree or career, we thought it useful to examine those respondents who had taken some philosophy classes but indicated that they had not seen the scenarios. This way we wanted to attain a sample of individuals who had been interested to pursue philosophy as a career but who were unbiased by previous (possibly 'male-centrist') discussions of the cases.

To summarize, the criteria that had to be met for participants to be included in the analysis here were 1) comprehension check was answered correctly, 2) time spent to complete the task was not less than 30 seconds, 3) native language was English, 4) indicated that they had not seen the scenarios before and 5) indicated number of classes were between one and three. In specific, in the demographic section of the surveys we asked how many philosophy courses respondents had taken and the answer choices provided were '0', '1 to 3', '4 to 6' and '> 6'. This was the same for all surveys with the exception of the Chinese Room scenario where we asked whether participants had taken any philosophy courses and the answer choices were 'Yes' and 'No'. Respondents had to have chosen 'Yes' (in addition to fulfilling the other criteria) to be included in the analysis provided below.

For this group of respondents again, there were no statistically significant differences between women and men. The data in this section is drawn from the Mechanical Turk data sets. The samples for the SurveyMonkey data were relatively small after filtering in this way. None of the scenarios yielded a significant difference

and hence we will not present the outcomes here. We will present the summary of the outcomes and graphs for the Mechanical Turk data next.

Brain in the Vat: One to Three Philosophy Courses (MT)

N = 126 (Male = 85, Female = 41). Male: Mean = 4.95, SD = 2.37. Female: Mean = 5.68, SD = 1.82. Independent Samples t-test: $t(124) = -1.74$, $p = 0.085$.

Twin-Earth: One to Three Philosophy Courses (MT)

N = 88 (Male = 57, Female = 31). Male: Mean = 5.23, SD = 2.13. Female: Mean = 5.29, SD = 1.99. Independent Samples t-test: $t(86) = -0.134$, $p = 0.894$.

Chinese Room: More than One Philosophy Course (MT)

N = 77 (Male = 32, Female = 45). Male: Mean = 3.22, SD = 1.996. Female: Mean = 3.33, SD = 1.784. Independent Samples t-test: $t(75) = -0.264$, $p = 0.792$.

Plank of Carneades: One to three Philosophy Courses (MT)

N = 190 (Male = 99, Female = 91). Male: Mean = 5.39, SD = 1.602. Female: Mean = 5.71, SD = 1.455. Independent Samples t-test: $t(188) = -1.438$, $p = 0.152$.

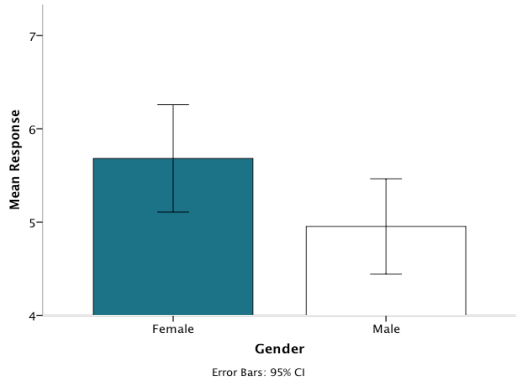

Fig. 5a: Brain in the Vat - One to Three Philosophy Courses (Mechanical Turk)

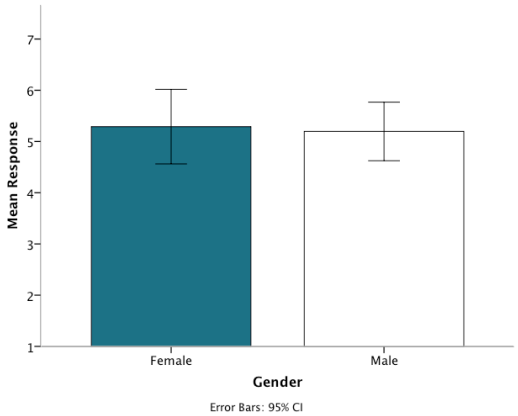

Fig. 5b: Twin Earth – One to Three Philosophy Courses (Mechanical Turk)

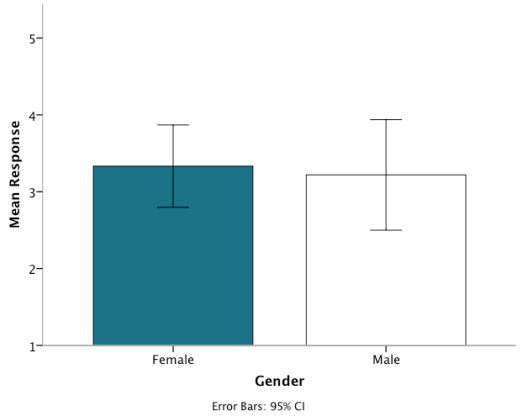

Fig. 5c: Chinese Room – One to Three Philosophy Courses (Mechanical Turk)

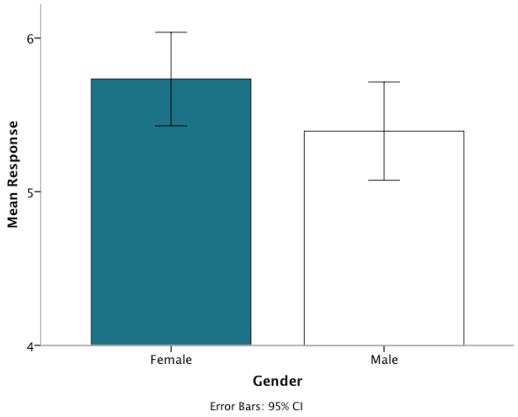

Fig. 5d: Plank of Carneades – One to Three Philosophy Courses (Mechanical Turk)

We will provide a brief discussion of these results in the concluding section of the paper. Next, we will discuss section 3.2 of Buckwalter & Stich (forthcoming) where the authors present results taken from Geoffrey Holtzman on Compatibalism, Materialism and Dualism.

3.1 Compatibalism, Materialism and Dualism

For the scenarios in this section we collected data through SurveyMonkey. The method of data collection is the same as described in section 2.1.

Compatibalism

The first case presented by Buckwalter & Stich is a scenario eliciting intuitions on a compatibilism thought experiment. The scenario reads as follows.

> Suppose Scientists figure out the exact state of the universe during the Big Bang, and figure out all the laws of physics as well. They put this information into a computer, and the computer perfectly predicts everything that has ever happened. In other words, they prove that everything that happens, has to happen exactly that way because of the laws of physics and everything that's come before. In this case, is a person free to choose whether or not to murder someone?

Respondents could select either answer choice 'Yes' or 'No'. Holtzman only included participants with no prior background in formal philosophy in the data analysis. The outcome Buckwalter & Stich report for Fisher's Exact Test comparing women and men is $p < 0.0005$, $N = 192$ (102 male, 90 female) and $d = 0.58$. Furthermore, 63% of women responded that in this scenario a person is free to choose to murder, whereas only 35% of men gave this answer.

Replication Results (SurveyMonkey):

Using the same filter as Holtzman we failed to attain a significant difference among men and women. Our sample consisted of 92 participants with 50 of those being female and 42 male. A Chi-Square test yielded $\chi 2 = 0.652$, $p = 0.419$.[14]

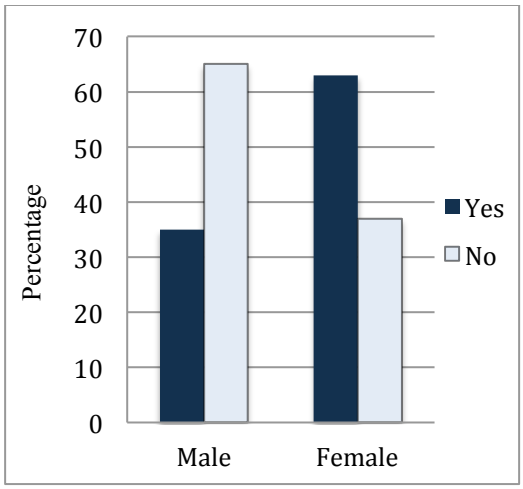

Fig. 6a: Compatibalism – Original Results

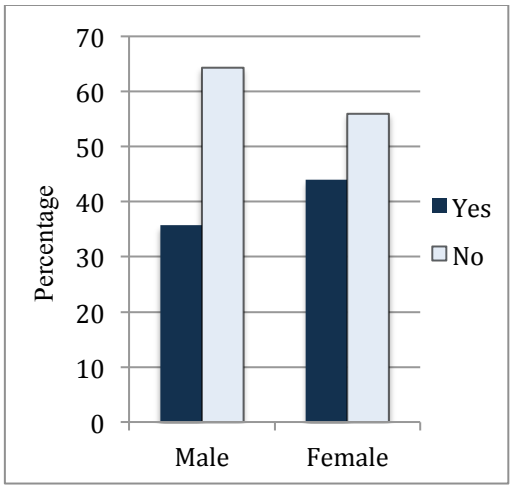

Fig. 6b: Compatibalism – Replication Results: SurveyMonkey

In our sample the percentage of men answering yes was also 35, however, the percentage of women who answered yes was 45. That is, women still had a higher percentage of 'yes' responses, however, not by as much as in Holtzman's data. Also, for our sample, both groups had a majority of 'no' responses as opposed to Holtzman's sample where women had a higher percentage of 'yes' than 'no' responses.

Physicalism

The next case that Buckwalter & Stich discuss reads as follows.

> Suppose you meet a man from the future who knows everything there is to know about science. He tells you that he doesn't like apples, and says that though he has never eaten one, he has figured out what apples taste like just by studying the relevant science. Could he know what apples taste like without ever having eaten one?

Again, the possible answer choices were 'Yes' or 'No'.

Buckwalter & Stich report a Fisher's Exact Test with $p < 0.005$, $d = 0.50$ and $N = 195$ (93 women and 102 men). Thirty-nine percent of male participants answered 'Yes' but only 17% of women answered so.

Our Results (SM):

As before we excluded from analysis participants who had taken one or more philosophy courses. The data yielded no statistically significant difference among women and men (at the typical levels).

$N = 101$ (49 Male, 52 Female), Fisher's Exact Test yielded $p = 0.518$ (one cell had expected count < 5).

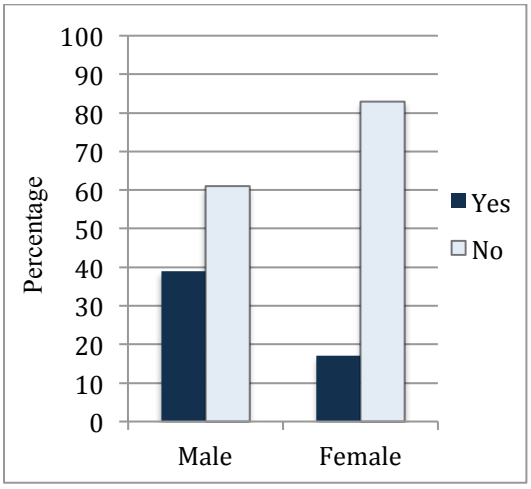

Fig. 7a: Physicalism – Original Results

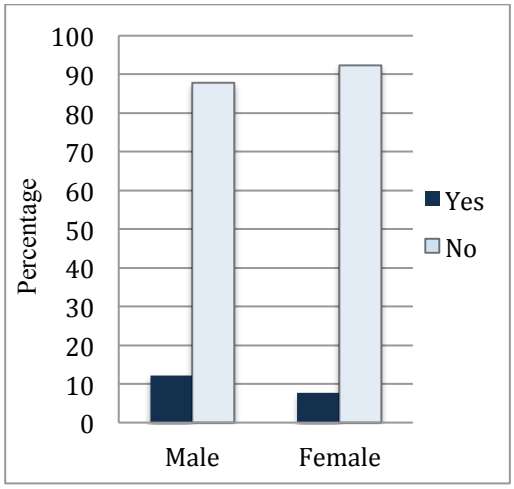

Fig. 7b: Physicalism – Replication Results: SurveyMonkey

Dualism

The dualism scenario Holtzman presented to participants reads as follows:

> Suppose neurologists are able to identify every part and every connection in the human brain. Working with a team of computer scientists, they then build a robot

that has a complete electronic replica of the human brain. Could this robot experience love?

The results presented by Buckwalter & Stich are the following: N = 185 (87 women, 98 men) Fisher's Exact Test yielded p = 0.016 (d = 0.37).

Replication Results (SM):

A Chi-Square test for 137 participants (65 Male, 72 Female) yielded $\chi 2 = 0.090$, p = 0.764.

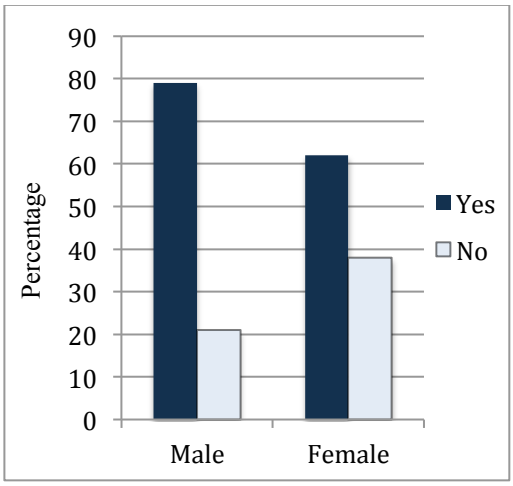

Fig. 8a: Dualism – Original Results

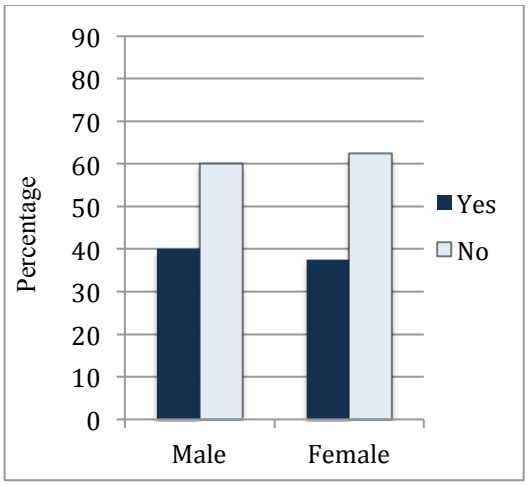

Fig. 8b: Dualism – Replicaiton Results: SurveyMonkey

Our samples for all three scenarios were smaller than those of Holtzman. It may be possible that the effect sizes are relatively small and that our data did not provide the necessary power to detect a difference.

We had used these scenarios as filler or dummy questions in some other surveys that we collected at the LSE in a computer lab setting. The samples were too small to be meaningful when we filtered out non-native speakers. However, when we included in

our analysis all who indicated their level of English as fluent the Dualism (N = 98) and the Compatibalism (N = 57) scenarios came close to significance both with a p-value of 0.052.[15] The Physicalim case (N = 57) had a p-value of 0.20.

We are not sure why there are these differences between our own samples and also between our samples and those of Holtzman. It may really be a difference between native and non-native English speakers. Also, the format between our two surveys was different using different questions in the SurveyMonkey questionnaire and the computer lab study. So, either the format could have made a difference or order effects could have played a role. For future studies of these cases it certainly makes sense to collect data where participants only see one of the scenarios in order to rule out any order effects.

3.2: Some Philosophy Background but Cases Not Seen Before

We ran a similar analysis as in section 2.2 where we filtered for respondents who had taken one to three philosophy courses but who indicated that they had not seen the scenarios before (and whose native language was English).

Once again there was no significant difference between women and men on any of the three scenarios though the samples for the Compatibalism and Physicalism cases were relatively small after filtering. See below for details.

Compatibalism:

A Chi-Square test yielded $\chi 2 = 1.227$, $p = 0.268$; N = 53 (Male = 30, Female = 23)

Physicalism

N = 58 (24 Male, 34 Female), Fisher's Exact Test yielded p = 0.432 (two cells had expected count < 5).

Dualism

N = 111 (54 Male, 57 Female), a Chi-Square test yielded $\chi 2 = 0.021$, p = 0.789.

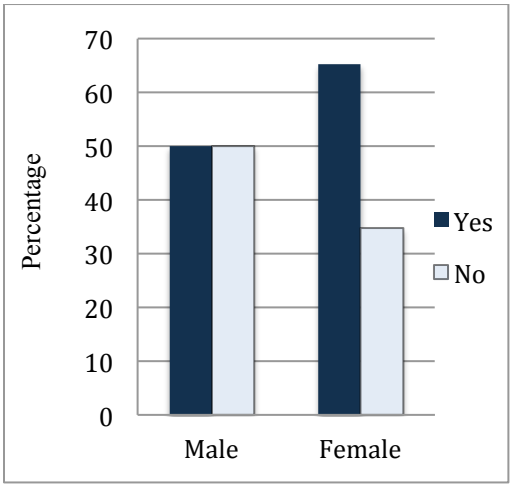

Fig. 9a: Compatibalism – One to Three Philosophy Courses (SurveyMonkey)

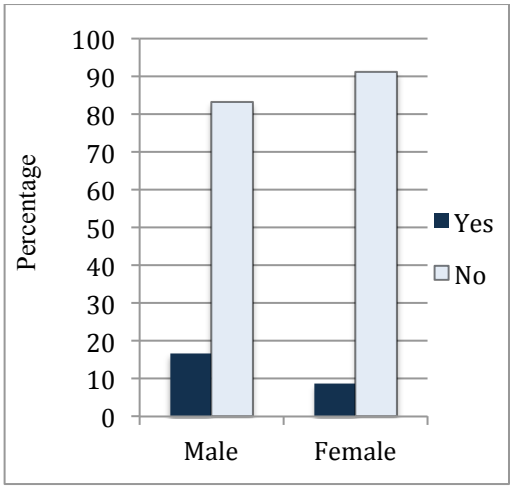

Fig. 9b: Physicalism – One to Three Philosophy Courses (SurveyMonkey)

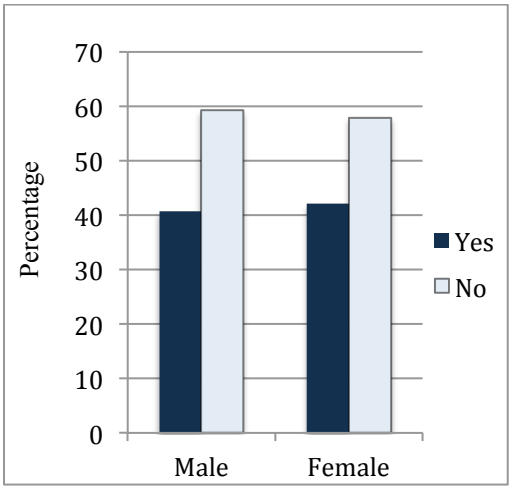

Fig. 9c: Dualism – One to Three Philosophy Courses (SurveyMonkey)

4. Gettier-style Scenarios

In section 3.1 of their paper Buckwalter & Stich present data for experiments conducted by Christina Starmans & Ori Friedman (2009) on Gettier-style cases. Although we did not collect data on the exact same scenarios, we had conducted surveys on four other Gettier type questions for a different study. We did not find significant differences among women and men in these experiments.

Procedures

For this section we collected data mainly through SurveyMonkey. However, for one of the scenarios (Gettier) we also collected data in classes at the LSE and online through Harvard University's Moral Sense Test (MST) website.[16] The procedures and methods for data collection for the SurveyMonkey samples were the same as described in section 2.1.

The in-class procedure was relatively straightforward. With the permission of class teachers we visited classes in the departments of Philosophy and Government and after a brief introduction handed out a short one-page questionnaire. Participation was voluntary although no one refused to answer. The whole procedure took about five minutes.

The procedure for the MST data was as follows. MST is setup so that people visit the site without an invitation or otherwise being solicited. After some initial instructions participants were forwarded to the questionnaires. The data presented in this section was drawn from several different surveys. The Gettier scenario was used as a filler question for surveys where we were testing several different effects.

Original Results:

The scenario that Starmans & Friedman presented to respondents reads as follows.

> Peter is in his locked apartment, and is reading. He decides to have a shower. He puts his book down on the coffee table. Then he takes off his watch, and also puts it on the coffee table. Then he goes into the bathroom. As Peter's shower begins, a burglar silently breaks into Peter's apartment. The burglar takes Peter's watch, puts a cheap plastic watch in its place, and then leaves. Peter has only been in the shower for two minutes, and he did not hear anything.
>
> Does Peter really know that there is a watch on the table, or does he only believe it?

The answer choices available were 'really knows' and 'only believes'. Starmans & Friedman report that whereas 71% of women choose 'really knows' only 41% of men choose this answer ($p < 0.05$, Fisher's exact test)[17].

Starmans & Friedman ran a variation on the above scenario where they changed the gender of the protagonist to female out of concern that this detail may have had an effect on responses and again attained a significant difference with $p < 0.01$ for $N = 112$ (52 men and 56 women) where 75% of women answered 'really knows' and only 36% of men answered so. For further details, see Buckwalter & Stich (forthcoming) and Starmans & Friedman (2009).

Replication Scenarios[18] and Results:

Gettier (SM)

The first scenario we examined was the following.

> Bob has a friend, Jill, who has driven a Buick for many years. Bob therefore thinks that Jill drives an American car. He is not aware, however, that her Buick has recently been stolen, and he is also not aware that Jill has replaced it with a Pontiac, which is a different kind of American car. Does Bob really know that Jill drives an American car, or does he only believe it?
>
> REALLY KNOWS ONLY BELIEVES

For our sample of 105 individuals (54 Male, 51 Female), a Chi-Square test yielded $\chi^2(1) = 0.108$, $p = 0.742$; (minimum expected count 10.9).[19]

Truetemp (SM)

The next scenario we examined is the Truetemp case, which we presented as follows:

> One day Charles is suddenly knocked out by a falling rock, and his brain becomes re–wired so that he is always absolutely right whenever he estimates the temperature where he is. Charles is completely unaware that his brain has been altered in this way. A few weeks later, this brain re–wiring leads him to believe that it is 71 degrees in his room. Apart from his estimation, he has no other reasons to think that it is 71 degrees. In fact, it is at that time 71 degrees in his room. Does Charles really know that it was 71 degrees in the room, or does he only believe it?
>
> REALLY KNOWS ONLY BELIEVES

The statistical analysis for N = 105 (Male = 54, Female = 51) yielded $\chi^2(1) = 0.382$, p = 0.536; (minimum expected count 13.6). There were two further Gettier type questions termed Zebra Case and Smoking Conspiracy Case for which we had previously collected data. For the exact wording of the cases, see Appendix C. The summary statistics for these two cases are as follows:

Zebra Case: N = 105 (54 Male, 51 Female). $\chi^2(1) = 0.654$, p = 0.419; (minimum expected count 10.7).

Smoking Conspiracy: N = 105 (54 Male, 51 Female). $\chi^2(1) = 0.153$, p = 0.696; (minimum expected count 10.2).

Below is a graph depicting the outcomes for all the Gettier-style experiments conducted on SM.

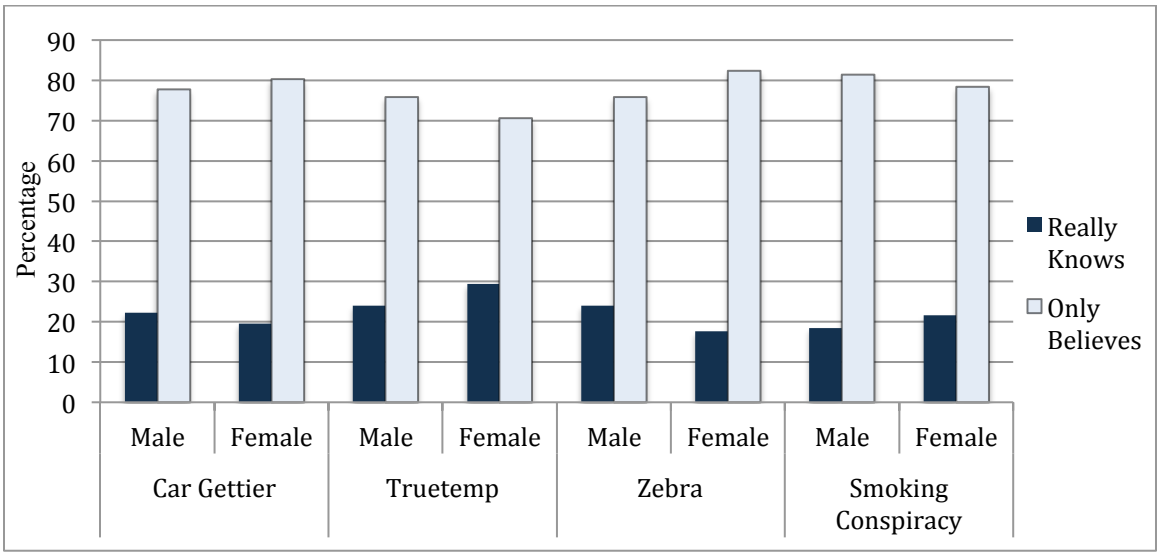

Fig. 10: Gettier Style Cases - SurveyMonkey

In addition to the tests above where the only filter used was for English as native language, we also ran two further analyses. In one, we further filtered out respondents who were not of a 'Western' background as there has been a question whether individuals

from 'Western' and non-Western backgrounds answer these scenarios differently.[20] Furthermore, in addition to native language and ethnic background filters, we also filtered out individuals whose highest level of education attained was below college. Again, this is because there has been a question whether individuals from different socioeconomic statuses (measured by an education proxy) answer Gettier type questions differently.[21] None of the tests yielded a significant difference among men and women.

In Class Data

As mentioned before, for the Car scenario we also collected data in two different ways; one in classroom settings and one through the Moral Sense Test (MST) website. The below summaries are for participants whose native language was English. The in-class data yielded a significant difference between men and women, the MT data, however, did not.

In-Class Gettier Results

N = 137 (71 Male, 66 Female). $\chi^2(1) = 4.222$, $p = 0.040$; (minimum expected count 9.1), p-exact = 0.049.

MST Car Gettier Results

N = 78 (44 Male, 34 Female). $\chi^2(1) = 0.608$, $p = 0.435$; (minimum expected count 7.4), p-exact = 0.582.

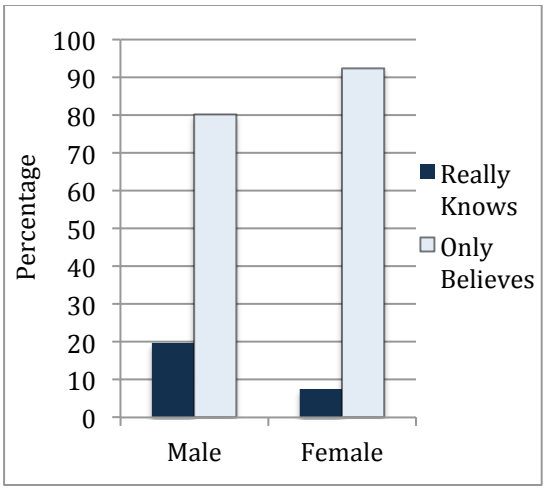

Fig. 11a Gettier (Car) – In Class

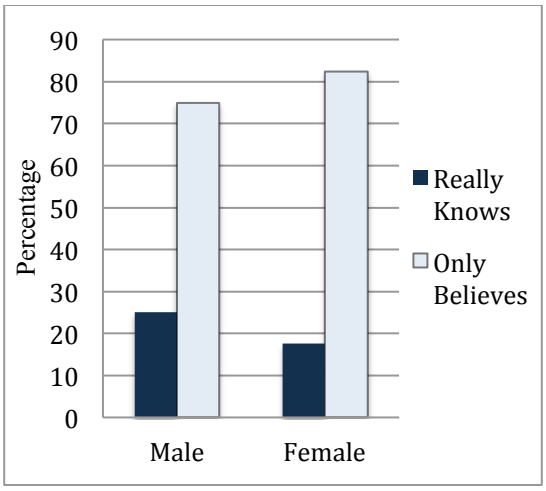

Fig. 11b Gettier (Car) – MST

Miscellaneous Points

There were several other scenarios for which we had collected data throughout the past couple of years and which we examined for differences between women and men that also did not yield any differences. Examples include Goedel-type scenarios and other Compatibalism scenarios, however, for the sake of brevity we will omit a formal discussion and restrict this paper to the cases that were presented by Buckwalter & Stich (forthcoming).

Apart from the negative results we reported above we also collected some data on the scenario of section 3.6 of Buckwalter & Stich (forthcoming) on the so-called Epistemic Side-Effect Effect.[22] Although our sample size was small and does not warrant a formal presentation, judging by the limited data we have, the outcome points to robustness of the results reported by Buckwalter & Stich.

5. Concluding Remarks

Overall, we were surprised by the outcomes presented in this paper. It is not very likely that our surveys were heavily distorted through excessive spamming as we took the same precautions as Buckwalter & Stich. Furthermore, we used more than one source (including classroom settings) for our data collection and the outcomes were comparable.

We hope that other researchers will find it worthwhile to carry out replications of the cases presented here. At this point, however, we do not believe that there is strong evidence that women's and men's responses differ significantly for the cases examined in this paper.

We are not too sure about the reasons for the different outcomes in our experiments and those reported in Buckwalter & Stich (forthcoming) and it is likely that there are different reasons for different studies.

A general problem is that Buckwalter & Stich asked many researchers to examine their data and naturally those who happened to have differences in their data responded. Others have pointed this out and although true, this explanation is obviously not a satisfactory one for the classical scenarios as Buckwalter & Stich collected the data themselves. A possible reason for the difference for the Mechanical Turk experiments could be that depending on when the HITs were published female and male respondents could have had different motivations for filling out surveys. For example, after working hours women may predominantly complete Mechanical Turk HITs for an alternative source of income and men may complete HITs to pass time, or vice versa. However, given that we collected data through several sources, which yielded similar outcomes, this also may not be a satisfactory explanation.

A factor that may have played a role in the Holtzman studies may be order effects. Given that each participant saw several scenarios, there could have been scenario interactions.

Finally, Buckwalter & Stich themselves point out that the robustness of the cases they discuss needs further investigation. For example, for the Holtzman cases Buckwalter & Stich note that Holtzman examined nine scenarios for which three yielded significant differences. Furthermore, Ori Friedman let us know via email that they themselves have been unable to replicate the results of their Gettier scenario and that the make-up of that particular sample may have been unusual.[23]

Naturally, we do not believe that our data gives a definite answer on whether women and men have different intuitive responses on the cases examined here. We made all possible efforts to make sure that our surveys were conducted correctly, that we did not bias participants in any way and that our data analysis was carried out properly. However, if we made a mistake, we are likely to have repeated it in all of our surveys. Hence, we hope that other researchers will attempt replication of the cases. The importance of the subject matter certainly merits further investigation.

Notes

[1] The author would like to thank Wesley Buckwalter and Stephen Stich for providing the details of the procedures and methods of their experiments and answering any questions we had. We would also like to thank Donal Cahill for his help with the Moral Sense Test and all class teachers at the University of London who provided us with class time to collect data and all students who participated.
[2] Download available from PhilPapers, http://philpapers.org/rec/BUCGAP
[3] It may be possible to bias individuals in philosophy courses other than through direct exposure of some cases. Nevertheless, by restricting samples as described we could at least rule out that participants had been influenced directly.
[4] For examples, see Buckwalter & Stich (forthcoming) or Nagel (2012), amongst others.
[5] Taken from Buckwalter & Stich (forthcoming).
[6] Summary taken from Buckwalter & Stich (forthcoming).
[7] We will refer to the groups as women/men and female/male interchangeably as female/male is how we asked for gender in the demographic part of our surveys.
[8] See Appendix B for the breakdown of the individual surveys.
[9] We used the same scaling for the charts as in Buckwalter & Stich (forthcoming) for better comparison.
[10] Chart taken from Buckwalter & Stich (forthcoming).
[11] Taken from Buckwalter & Stich (forthcoming).
[12] Taken from Buckwalter & Stich (forthcoming).
[13] Taken from Buckwalter & Stich (forthcoming).
[14] Throughout we will report the results for Chi-Square tests when none of the cells have an expected count of less than five and will conduct Fisher's Exact tests otherwise.
[15] Where we conducted a Chi-square test for the Dualism scenario and a Fisher's Exact test on the Compatibalism case.
[16] http://moral.wjh.harvard.edu/index2.html
[17] Taken from Buckwalter & Stich (forthcoming).
[18] All Gettier-style scenarios in this section were taken from Weinberg, Nichols & Stich (2001).
[19] All individuals were native English speakers.
[20] Weinberg, Nichols & Stich (2001).
[21] Weinberg, Nichols & Stich (2001).
[22] See Beebe & Buckwalter (2010).
[23] Personal correspondence 5/1/2012
[24] Taken from Weinberg, Nichols & Stich (2001).
[25] Taken from Weinberg, Nichols & Stich (2001).

References

Beebe, J.R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language, 25,* 474-498.

Buckwalter, W., & Stich, S. (forthcoming). Gender and Philosophical Intuition. In Joshua Knobe & Shaun Nichols (eds.), Experimental Philosophy, Vol.2. Oxford University Press.

Knobe, J. & Nichols, S. (forthcoming). *Experimental Philosophy, Vol.2*. Oxford: Oxford University Press.

Nagel, J. (2012). Intuitions and Experiments: A Defense of the Case Method in Epistemology. *Philosophy and Phenomenological Research, 85*(3), 495-527

Starmans, C., & Friedman, O. (2009). Is knowledge subjective? A sex difference in adults' epistemic intuitions. Poster presented at the 6th Biennial Meeting of the Cognitive Development Society, San Antonio, TX, October 16-17, 2009. Abstract available at http://www.cogdevsoc.org/prog2009/CDS09Program.pdf.

Weinberg, J.S., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics, 29(1), 429-460.*

Appendix A

This appendix contains the analyses as carried out in section 2.1 with the exception that participants who indicated that they had seen the scenarios were excluded. The data presented here is from the Mechanical Turk samples. Our samples from the SurveyMonkey data sets were not large enough after filtering.

Brain in the Vat

N = 108 (Male = 52, Female = 56). Male: Mean = 5.12, SD = 2.27. Female: Mean = 5.93, SD = 1.76. Independent Samples t-test: $t(96) = -2.07$ (equal variance not assumed), $p = 0.041$.

Twin Earth

N = 114 (Male = 63, Female = 51). Male: Mean = 5.22, SD = 2.38. Female: Mean = 5.43, SD = 2.12. Independent Samples t-test: $t(112) = -0.490$, $p = 0.625$.

Chinese Room

N = 99 (Male = 46, Female = 53). Male: Mean = 3.41, SD = 2.19. Female: Mean = 3.30, SD = 2.00. Independent Samples t-test: $t(97) = 0.264$, $p = 0.792$.

Plank of Carneades

N = 141 (Male = 64, Female = 77). Male: Mean = 5.23, SD = 1.55. Female: Mean = 5.48, SD = 1.47. Independent Samples t-test: $t(139) = 0.335$, $p = 0.335$.

Appendix B

Breakdown of the individual SurveyMonkey Surveys

Brain in the Vat

Survey 1 (longer survey)

N = 56, Male = 26, Female = 30. Male: Mean = 6.12, SD = 1.71. Female: Mean = 5.50, SD = 1.78. Independent Samples t-test: $t(54) = 1.317$, $p = 0.193$

Survey 2 (shorter survey)

N = 44, Male = 23, Female = 21. Male: Mean = 5.39, SD = 1.994. Female: Mean = 5.76, SD = 1.921. Independent Samples t-test: $t(42) = -0.627$, $p = 0.534$

Twin Earth

Survey 1

N = 54, Male = 26, Female = 28. Male: Mean = 6.00, SD = 1.81. Female: Mean = 5.07, SD = 2.62. Independent Samples t-test: $t(48) = 1.522$, $p = 0.134$

Survey 2

N = 31, Male = 14, Female = 17. Male: Mean = 5.64, SD = 2.53. Female: Mean = 5.47, SD = 2.53. Independent Samples t-test: $t(29) = 0.189$, $p = 0.852$.

Chinese Room

Survey 1

N = 49 (Male = 21, Female = 28). Male: Mean = 3.62, SD = 2.61. Female: Mean = 3.39, SD = 2.32. Independent Samples t-test: $t(47) = 0.320$, $p = 0.750$.

Survey 2

N = 31 (Male = 14, Female = 17). Male: Mean = 3.71, SD = 2.64. Female: Mean = 4.53, SD = 2.38. Independent Samples t-test: $t(29) = -0.904$, $p = 0.374$.

Plank of Carneades

Survey 1

N = 54 (Male = 26, Female = 28). Male: Mean = 6.04, SD = 1.43. Female: Mean = 5.54, SD = 1.71. Independent Samples t-test: $t(52) = 1.168$ (equal variances not assumed), $p = 0.248$.

Survey 2

N = 44 (Male = 22, Female = 22). Male: Mean = 5.64, SD = 1.50. Female: Mean = 5.73, SD = 1.75. Independent Samples t-test: $t(42) = -0.185$, $p = 0.854$.

Appendix C

Zebra Case

Mike is a young man visiting the zoo with his son, and when they come to the zebra cage, Mike points to the animal and says, "that's a zebra." Mike is right — it is a zebra. However, as the older people in his community know, there are lots of ways that people can be tricked into believing things that aren't true. Indeed, the older people in the community know that it's possible that zoo authorities could cleverly disguise mules to look just like zebras, and people viewing the animals would not be able to tell the difference. If the animal that Mike called a zebra had really been such a cleverly painted mule, Mike still would have thought that it was a zebra. Does Mike really know that the animal is a zebra, or does he only believe that it is?

REALLY KNOWS ONLY BELIEVES[24]

Conspiracy Case

It's clear that smoking cigarettes increases the likelihood of getting cancer. However, there is now a great deal of evidence that just using nicotine by itself without smoking (for instance, by taking a nicotine pill) does not increase the likelihood of getting cancer. Jim knows about this evidence and as a result, he believes that using nicotine does not increase the likelihood of getting cancer. It

is possible that the tobacco companies dishonestly made up and publicized this evidence that using nicotine does not increase the likelihood of cancer, and that the evidence is really false and misleading. Now, the tobacco companies did not actually make up this evidence, but Jim is not aware of this fact. Does Jim really know that using nicotine doesn't increase the likelihood of getting cancer, or does he only believe it?

REALLY KNOWS ONLY BELIEVES[25]