Christina Easton

Women and 'the philosophical personality': evaluating whether gender differences in the Cognitive Reflection Test have significance for explaining the gender gap in philosophy

Article (Published version) (Refereed)

Synthese
https://doi.org/10.1007/s11229-018-01986-w

Women and 'the philosophical personality': evaluating whether gender differences in the Cognitive Reflection Test have significance for explaining the gender gap in Philosophy

Christina Easton[1]

Received: 2 March 2018 / Accepted: 13 October 2018

Abstract
The Cognitive Reflection Test (CRT) is purported to test our inclination to overcome impulsive, intuitive thought with effortful, rational reflection. Research suggests that philosophers tend to perform better on this test than non-philosophers, and that men tend to perform better than women. Taken together, these findings could be interpreted as partially explaining the gender gap that exists in Philosophy: there are fewer women in Philosophy because women are less likely to possess the ideal 'philosophical personality'. If this explanation for the gender gap in Philosophy is accepted, it might be seen to exonerate Philosophy departments of the need to put in place much-needed strategies for promoting gender diversity. This paper discusses a number of reasons for thinking that this would be the wrong conclusion to draw from the research. Firstly, the CRT may not track what it is claimed it tracks. Secondly, the trait tracked by the CRT may not be something that we should value in philosophers. Thirdly, even if we accept that the CRT tracks a trait that has value, this trait might be of limited importance to good philosophising. Lastly, the causal story linking the gender gap in CRT score and the gender gap in Philosophy is likely to be far more complex than this explanation implies.

Keywords Philosophical methods · Rationality · Women · Gender · Cognition · Intuition

✉ Christina Easton
C.E.Easton@lse.ac.uk

[1] London School of Economics and Political Science, LSE, Houghton Street, London WC2A 2AE, UK

Published online: 31 October 2018

1 Introduction

The Cognitive Reflection Test (CRT) is purported to test our inclination to overcome impulsive, intuitive thought with effortful, rational reflection. Research suggests that philosophers tend to perform better on this test than non-philosophers, and that men tend to perform better than women. Taken together, these findings could be interpreted as partially explaining the gender gap that exists in Philosophy: there are fewer women in Philosophy because women are less likely to possess this aspect of the ideal philosophical personality. If this explanation for the gender gap in Philosophy is accepted, it might be seen to exonerate Philosophy departments of the need to put in place much-needed strategies for promoting gender diversity. This paper discusses a number of reasons for thinking that this would be the wrong conclusion to draw from the research. Firstly, the CRT may not track what it is claimed it tracks. The dominant interpretation of the CRT is that it tracks an aspect of rationality, but it may be that the CRT tracks numeracy or confidence instead. Secondly, the trait tracked by the CRT may not be something we should value in philosophers. Even if we currently select for this trait in Philosophy, it may be that this trait is not, in fact, an asset to good philosophising. Thirdly, even if we accept that the CRT tracks a trait that has value, this trait might be of *limited* importance to good philosophising. A whole range of virtues and skills can plausibly be postulated as part of the ideal philosophical personality, and it is not clear to what extent the trait tracked by the CRT is an *important* philosophical virtue or skill. Lastly, the causal story linking the gender gap in CRT score and the gender gap in Philosophy is likely to be far more complex than this explanation implies. If the CRT gender gap is explanatory for the Philosophy gender gap, it is likely that it will be one of several, interacting causal factors.

The research at present does not allow us to draw clear conclusions over which route (or combination of routes) we should take in response to the findings. However, one response is clear. Even if it is the case that the CRT gender gap is somewhat explanatory of the Philosophy gender gap, and even if it is right that the CRT tracks a trait that is conducive to good philosophising, this does not justify inaction on the part of Philosophy departments or wider society. Rather, it points to the need for a self-conscious analysis of the discipline, including looking at what other obstacles there may be to women's participation in Philosophy. Additionally, since gender differences in CRT score are likely to be (at least partly) the result of environmental factors, it points towards the need for action aimed at rectifying injustices in wider society, so that women can develop their skills at whatever it is that the CRT tracks.

2 The Cognitive Reflection Test

In 2005, Shane Frederick proposed the CRT as a measure of one type of cognitive ability. A participant's CRT score is the number of questions that he or she answers correctly on the following, three-item test:

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? *Answer: ___ cents.*

2. If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets? *Answer: ___ minutes.*
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? *Answer: ___ days.*

The test questions have been chosen because they invite an intuitive, wrong answer. For example, consider the first question. The answer '10 cents' springs to mind, but this "impulsive" answer is incorrect (Frederick 2005, p. 26). Further reflection on the problem leads one to realise that the difference between $1.00 and 10 cents is only 90 cents, not $1.00, and "catching that error is tantamount to solving the problem, since nearly everyone who does not respond "10 cents" does, in fact, give the correct response: "5 cents."" (Frederick 2005, p. 27)

In his original paper, Frederick discusses how CRT score is positively and significantly correlated with various other measures of cognitive ability (for example, the Wonderlic Personnel Test and the Scholastic Achievement Test). However, he argues that the CRT tests something distinctive—"cognitive reflection"—which he defines as "the ability or disposition to resist reporting the response that first comes to mind" (Frederick 2005, p. 35).

Frederick links performance on the CRT with the distinction between two types of cognitive processing, referred to by Stanovich and West (2000) as "System 1" and "System 2". Nobel Prize winner Daniel Kahneman has brought this dual-process model of decision-making (as well as the CRT itself) to the attention of the public through his internationally bestselling *Thinking, Fast and Slow*.[1] System 1 operates quickly and automatically, with little effort and no sense of voluntary control (Kahneman 2011, p. 20). It is what gives us the immediate, wrong answers to the CRT questions. System 2 involves slower, more deliberate and effortful thinking (2011, p. 13). It has a "supervisory function" (2011, p. 48), monitoring and controlling the thoughts and actions being 'suggested' by System 1 (2011, p. 44). Thus, if System 2 is activated in response to a CRT question, it can override System 1 to give the right answer.

Kahneman cautions us against interpreting 'systems' too literally: the terms do not describe two parts of the brain enacting distinct functions (2011, p. 29). Writing with Frederick, he clarifies that:

> [The terms System 1 and System 2] may suggest the image of autonomous homunculi, but such a meaning is not intended. We use *systems* as a label for collections of processes that are distinguished by their speed, controllability, and the contents on which they operate. (Kahneman and Frederick 2002, p. 51)

Different test scores on the CRT are said to indicate individual differences in the way this dual-system functions. Performing poorly on the test indicates a "lazy" System 2 that relies on System 1 to do the work (2011, p. 48). In Kahneman's words, these individuals are "impulsive, impatient, and keen to receive immediate gratification"

[1] Dual-process models come in various flavours, but all distinguish between cognitive operations that are quick and associative and those that are slow and rule-governed (Kahneman and Frederick 2002, p. 51). For an overview of dual-processing accounts, see Evans (2008). Though this model has dominated cognitive style research over the last 50 years, we should note that it has also been the subject of criticism (e.g. van Mulukom 2018; Wang et al. 2017; Keren and Schul 2009).

(2011, p. 48). In contrast, avoiding the intuitive incorrect answer indicates a "more active" mind (2011, p. 45). These individuals are more likely to invest the effort required to check their intuitions in other circumstances, and more likely to defer gratification (2011, p. 48).

The CRT has since become a "tremendously influential measure of reflective thinking" (Thomson and Oppenheimer 2016, p. 107) and has been utilised in dozens of research studies. At the time of writing, Frederick's paper has been cited 2835 times (Google Scholar, 5 October 2018). The dominant view remains that the CRT measures something unique (Szaszi et al. 2017, p. 207), marking it out from other cognitive tests. For example, Toplak et al. (2011, p. 1275) concluded that "the CRT was a unique predictor of performance on heuristics-and-biases tasks" and that it tracks "miserly processing" in a way no other test does. The quest for identifying correlations between CRT scores and other individual differences has continued until the present day.[2]

2.1 Interpreting the CRT

The standard interpretation of the CRT has been that it tracks 'reflectivity', understood as an inclination to stop and reflect on one's intuitions. In his original paper, Frederick gives a number of reasons to support this, including that many of the participants who gave right answers had scribbles in the margin or gave verbal reports indicating that the wrong, intuitive answer was considered first (Frederick 2005, p. 27).

However, some recent studies have questioned this standard interpretation. It makes sense to assume that 'stopping and reflecting' will take additional time, yet Stupple et al. (2017) found only a weak correlation between CRT response times and accuracy. In their 'thinking aloud' study, Szaszi et al. (2017) found that a significant proportion (77%) of 'Correct Answer' respondents started their thought process with a 'Correct Start' thinking lead (as opposed to the wrong intuition being reflected upon and corrected). Moreover, 39% of 'Wrong Answer' respondents did attempt to reflect on their answers. Perhaps these latter participants suffered not from a lack of reflectivity, but from a 'mindware problem':

> individuals lack the declarative knowledge and strategic rules that are needed to solve some problems. Consequently, even when individuals put considerable mental effort into the problem-solving process, the lack of this necessary knowledge can lead to thinking failures. (Szaszi et al. 2017, p. 223)

These studies raise important doubts over the standard interpretation of what the CRT tracks, and this should be kept in mind in the discussion that follows (especially Sect. 6.1.1). However, these studies are far from conclusive. Jimenez et al. (2018) also tested response times and got the reverse conclusion to Stupple et al. (2018): "impulsive subjects complete the test quicker than reflective subjects" (Jimenez et al. 2018, p. 41). Szaszi et al.'s study was a 'thinking aloud' study, relying on participant self-reports of

[2] To give just a few examples, correlations have been found between CRT performance and time and risk-preferences (Frederick 2005, p. 36), likelihood of voting Leave or Remain in the British EU referendum (House 51), utilitarian moral judgement (Paxton et al. 2012; Baron et al. 2015), belief in the supernatural (Gervais and Norenzayan 2012) and paranormal beliefs, conspiracy beliefs, and conspiracy mentality (Stahl and van Prooijenb 2018).

the thought process, which may be incomplete. Their data can be explained in ways that are consistent with the standard interpretation. For example, the 'Correct Start' participants may have considered the intuitive, wrong answer but never voiced this.[3]

3 The CRT and gender

Frederick's original study indicated a significant gender gap in CRT score. The average score for men was 1.47 and for women was 1.03 (out of a maximum score of 3). The significance of group difference was calculated as $p < 0.0001$ (Frederick 2005, p. 38). Frederick suggests that the test maps "something that men have more of" and concludes that "men are more likely to reflect on their answers and less inclined to go with their intuitive responses" (2005, p. 37).

The finding that there is a gender gap in CRT score has been reliably replicated ever since (e.g. Szaszi et al. 2017, p. 216; Zhang et al. 2016; Thomson and Oppenheimer 2016, p. 106; Livengood et al. 2010), including in studies with participants from different age groups, educational levels and countries, and using the original CRT as well as some modified versions (Primi et al. 2018, p. 259). For example, a 2016 study gave typical results when it found that women are more likely than men to answer all three questions incorrectly and that the average CRT score of men is significantly higher than women (1.12 vs. 0.58, $p < 0.001$) (Cueva et al. 2016, p. 82). One 2017 study reported that "males scored 83.8% higher on the CRT" than females (Agnew 2017, p. 8). In their meta-analysis of 118 CRT studies (comprising of 44,558 participants across 21 countries), Brañas-Garza et al. (2015) found a negative correlation between being female and giving correct answers to the CRT test questions. Amongst 'Wrong Answer' respondents, women are more likely than men to give the intuitive response (e.g. Frederick 2005; Cueva et al. 2016; Pennycook et al. 2016).

4 The CRT and Philosophy

The research on the gender gap in CRT score makes for some rather uncomfortable reading, especially in light of the association of the CRT with an aspect of rationality. It appears to support the stereotype that women are more guided by intuition than rationality. On the face of it, it lends credibility to the view that "rationality is masculine", a view that forms "a backdrop to common Western conceptions of gender difference that have a deep influence on everyday life" (Haslanger 2012, p. 47).

These feelings of discomfort escalate when we pair these findings with a look at how CRT scoring has been used in some recent research in experimental philosophy. In a study involving data on 4472 participants, Livengood et al. (2010) investigated the relationship between philosophical training and CRT score. They found that the mean CRT score for participants with some training in Philosophy (0.98) was more than double the mean CRT score for participants with no training in Philosophy (0.44). Further, the mean CRT score for participants with some graduate training in Philos-

[3] See Szaszi et al. (2017, pp. 225–226) for a discussion of the limitations of their study, some of which also apply to the Stupple et al. study.

ophy (1.32) was triple the mean CRT score for those with no training in Philosophy (Livengood et al. 2010, p. 316).

Those with more training in Philosophy tend to be better educated than those with no training in Philosophy, and so Livengood et al. sought to isolate 'training in Philosophy' as a factor. They found that even when controlling for relevant factors such as levels of education and gender, people with more philosophical training tend to exhibit higher CRT scores.[4] For example, out of those participants who reported having had some college education, the mean CRT score of those who had taken some Philosophy courses was nearly 70% higher than that of those who had not taken Philosophy courses (Livengood et al. 2010, p. 316).

The authors comment that their data suggests that there are some "deep commonalities" among philosophers. They hypothesise that philosophers share a "philosophical temperament"—a "cluster of dispositions that distinguishes philosophy from other intellectual endeavours" (Livengood et al. 2010, p. 318). Livengood et al. do not suggest what other dispositions might make up this 'philosophical temperament', instead focusing their discussion entirely on the "single, but important aspect of philosophical temperament" that they suppose is tracked by the CRT (2010, p. 318). This aspect is "cognitive reflectivity", understood as "a disposition to challenge one's own intuitions whenever presented with a novel problem, rather than simply relying on whatever first comes to mind" (2010, p. 314). They suggest that "philosophers are less likely to blindly accept their intuitions and more likely to submit those intuitions to scrutiny" (2010, p. 319). They conclude that their data suggests that this reflectivity is "an important facet of philosophical personality" (2010, p. 314).

More recently, Justin Sytsma (2016) has used CRT scoring to hypothesise that religious philosophers are "less analytic", perhaps explaining the alleged "poor health" of the sub-discipline of Philosophy of Religion. Sytsma implies that the CRT tracks what type of 'thinking style' you have ("analytic" versus "intuitive") and speculates that this might correlate with an ability to evaluate arguments. Putting his controversial hypotheses to one side, we can note that there are two assumptions at work here. Firstly, the CRT tracks analyticity (understood as a propensity to think analytically as opposed to intuitively). Secondly, possessing the feature tracked by the CRT contributes to 'healthy' philosophising. Both of these assumptions will be questioned in this paper.

4.1 Criticism of the idea of a 'philosophical personality'

The view taken by Livengood et al.—that philosophers are in some sense 'expert-intuiters'—is a version of what has become known as the 'expertise defence'.[5] Whilst

[4] This finding has been replicated by Byrd (2014, p. 31), who found that "training and selection in philosophy resulted in better performance on the CRT".

[5] It is a 'defence' because it has usually been discussed as a response to "the restrictionist challenge" (Alexander and Weinberg 2007), which says that since the findings of experimental philosophers call into question the truth-tracking features of many philosophically relevant intuitions (Feltz and Cokeley 2012), the current reliance on intuitions in Philosophy should be radically restricted. Roughly, the expertise defence replies that since philosophers are experts, their intuitions are not subject to the sorts of distortions that have been seen in experiments with non-philosophers, and therefore can be trusted for use in philosophical theorising. Rini (2015, p. 434) discusses two versions of the expertise defence: one that says that philosophers

some have seen philosophical expertise as lying in having *better intuitions*, Livengood et al.'s understanding seems to be that philosophers possess a special trait enabling them to *overcome* biases of judgement and to reflect appropriately on intuitions. They talk about it being part of "expert philosophizing" to employ "intuition-poking practices" (2010, p. 319), "a range of possible practices, which all have in common that they are meant to determine whether the intuition is trustworthy and should thus be endorsed" (2010, p. 318). Philosophers are "more reflective than their peers: they are less likely than their peers to embrace what seems obvious without questioning it, and they are disposed to submit to scrutiny their intuitive inclination to judge that something is the case" (2010, p. 314).

Recently, a number of empirical studies have led to the expertise defence coming under scrutiny (Tobia et al. 2013; Schulz et al. 2011; Horvath and Wiegmann 2016). Notably, Schwitzgebel and Cushman (2012)'s research suggests that philosophers are as easily trapped by unreflective intuitions as non-philosophers. Earlier studies had found that how people judge a hypothetical moral scenario is affected by the order in which these scenarios are presented. That is, moral judgements are subject to 'order effects'. Since order of presentation is a factor that seems irrelevant to the rightness or wrongness of a scenario, we would hope that philosophers would be protected against this source of bias. Yet Schwitzgebel and Cushman found that philosophers judging moral scenarios were also subject to these order effects. Moreover, the order in which scenarios were presented substantially influenced which general moral principles the philosophers then endorsed. Contra Livengood et al., this suggests that there is no distinctive personality trait of 'reflectivity' that gives philosophers a special ability to overcome biases of judgement.

It might be thought that these new findings prevent the idea of a 'philosophical personality' from getting off the ground. However, there are a number of ways of explaining Schwitzgebel and Cushman's results that leave the expertise defence intact (Rini 2015). For example, it might be that Schwitzgebel and Cushman get the results that they do for philosophers because they are forcing them to make a binary choice on a question which the philosopher believes does not have a clear 'yes' or 'no' answer. This is supported by Bourget and Chalmers's (2014) study of professional philosophers, which indicated that philosophers are disinclined to make binary judgements on moral principles similar to those asked for in Schwitzgebel and Cushman's study. Forcing a binary judgement in response to a moral scenario or principle already identified by philosophers as problematic is therefore unlikely to reveal much about ordinary philosophical practice (and correspondingly, about what philosophical expertise consists in), since ordinary philosophical practice seems to involve *refraining* from forming these judgments (Rini 2015, p. 445). Thus it might still be the case that philosophers, when acting *qua philosophers* (engaged in their ordinary philosophical practice), are

Footnote 5 continued
simply have better intuitions than non-philosophers, and one that says that philosophers make better use of their intuitions. Note that both of these are different from how I formulate the defence here. I try to do so in a way that is consistent with the strong performance of philosophers on the CRT, which seems to be about simply *discarding* intuitions, rather than starting off with correct intuitions or making good use of our intuitions. For defences of philosophical expertise, see Singer (1972), Ludwig (2007), Grundmann (2010), Williamson (2007) and Williamson (2011). For a theoretical challenge to the expertise defence, see Weinberg et al. (2010).

particularly careful when drawing conclusions based on certain intuitions, and here lies an element of their expertise.

More recently, Drożdżowicz (2018) has argued that even in light of the empirical studies, there remains room for a task-based version of the expertise defence, where philosophical expertise lies in (i) devising and discussing arguments, (ii) proposing, modifying, and refuting theories, and (iii) articulating and applying distinctions. She actually cites the Livengood et al. study as an example of one potentially fruitful way of testing whether philosophers have this kind of expertise. If philosophers have "extensive training in argumentation", which plausibly involves "evaluating one's intuitions as premises and blocking them, if needed, then it could perhaps be hypothesized that philosophers will score better in the CRT than non-philosophers…" (Drożdżowicz 2018, p. 268). Since this is precisely what was found in Livengood et al.'s study, the idea that such a disposition might be part of the philosophical personality remains somewhat plausible.

5 How do these findings bear on the under-representation of women in Philosophy?

The empirical research on the CRT and its relation to gender and Philosophy appears to be telling us two things: Women tend to perform worse on the CRT than men, and philosophers tend to perform better than non-philosophers. These purported facts could be interpreted as shedding light on a further fact: women are under-represented in Philosophy.[6] Whilst most career paths and subject areas have seen a steady increase in women's participation, often to the point of equal representation or over-representation, there remains a lack of gender parity in Philosophy, comparable to the under-representation of women in 'STEM' subjects (Science, Technology, Engineering and Mathematics). A steady decline (often referred to as a 'leaky pipeline') can be seen in women's participation in Philosophy as we move 'up the stages'. For example, in the UK, a drop was seen from 46% at undergraduate level, through to 31% at PhD level, to 24% of permanent staff and 19% of professors (Beebee and Saul 2011). In the US, women make up about 30% of those earning Philosophy PhDs, far less than the average for all disciplines (Figdor and Drabek 2016). According to the Survey of Earned Doctorates in the US in 2009, Engineering, Computer Science and Physics are the only subjects where women earn fewer PhDs than in Philosophy (Healy 2011). Women are also poorly represented in the highest-ranked Philosophy journals, even when compared to the number of women working in elite universities. Sally Haslanger's survey of seven top Philosophy journals from 2002 to 2007 found that 12.4% of all authors were women (Haslanger 2008).

Greeted with these three streams of research (on the CRT and gender, on the CRT and Philosophy, and on the Philosophy gender gap), one might be tempted to propose something like the following: Perhaps women are less likely to possess the aspect of the ideal philosophical personality that is tracked by the CRT, and this contributes to

[6] There is a growing literature documenting this trend. For detailed discussions, see Beebee and Saul (2011), Figdor and Drabek (2016) and Thompson (2017).

the gender gap in Philosophy.[7] If this is right, then it might be thought that the current trend towards encouraging Philosophy departments to engage in affirmative action strategies is misguided.[8] Since this is a natural (perhaps 'intuitive') response to the research, I call this the 'Quick Conclusion'.

> *Quick Conclusion*: The CRT tracks something valuable in Philosophy – an aspect of the ideal philosophical personality – which women tend to lack. The gender gap in the CRT is therefore explanatory for the gender gap in Philosophy. Philosophy departments and wider society are therefore exonerated of the need to institute or maintain practices intended to decrease the gender gap in Philosophy.

Many readers will find this a highly unpalatable explanation for the gender gap in Philosophy. At its most crude, this view suggests that there are fewer female philosophers because women are less rational. Livengood et al. and Sytsma do *not* draw this conclusion (in fact, they do not discuss the implications of their findings for women and Philosophy). But there is a risk that others who encounter these findings will do, for it is hard to deny that this kind of explanation for the gender gap has at least some prima facie plausibility.

First, it is suggested by the ideas presented above: the important role that intuitions play in philosophical practice, the dominance of intuitions in discussions of philosophical expertise, and the CRT as a particularly potent measure of how people tend to respond to intuitions.[9] As has been discussed, high CRT score is said to indicate a predisposition to careful reflection rather than reliance on intuitions. Some see intuitions as the "raw data" of Philosophy, with the role of the philosopher being to rigorously analyse these intuitions (Hutchison 2013, p. 112). Kahneman talks of System 2 as sometimes acting as an "apologist" for the automatic responses provided by System 1 (2011, p. 103). Thus we might see the practice of Philosophy as hyper-exercise of System 2, in order to scrutinise, justify and in some cases, override the intuitions provided by System 1. If women are more inclined to simply go with the first intuition

[7] Someone pushing a philosophical personality explanation of the gender gap would need to propose other aspects of the philosophical personality that women supposedly lack—perhaps an ability to withstand harsh criticism and aggressive questioning (Beebee 2013) or a propensity to enjoy topics that do not appear practically useful or relevant to one's life (Thompson et al. 2016). Following Livengood et al. (2010), my focus throughout this paper is solely on the one aspect of the philosophical personality apparently tracked by the CRT. There are a number of reasons for this. First and foremost, it is because the fundamental question of this paper arose when I encountered three apparent facts—the low CRT score amongst women, the high CRT score amongst philosophers, and the low representation of women in Philosophy—and had the 'Quick Conclusion' offered by my interlocutors in response to these facts. Thus the paper's main aim is to make sense of this combination of facts together. Second, the idea that a special skill relating to intuitions is central to good philosophical practice has prima facie plausibility and has been a dominant position in Philosophy, evidenced by the wide amount of discussion of the role of intuitions in Philosophy and of the idea that philosophers are 'expert-intuiters'. Third, there is a general consensus that this skill is particularly well-tracked by the CRT, and there is a vast amount of evidence, interest and literature surrounding the CRT to draw on, including in the experimental philosophy and philosophical methodology literature. Thanks to an anonymous reviewer for Synthese for pressing me on this issue.

[8] For examples of this trend, see APA (2017), BPA/SWIP (2011), and Hassoun (2017).

[9] On the important role that intuitions play in Philosophy, as well as the debates surrounding this question, see Feltz and Cokeley (2012, pp. 229–231), Pinillos et al. (2011, p. 116) and Sosa (2009). In opposition to this dominant view, see Cappelen (2012), who argues that it is not characteristic of philosophers to rely on intuitions as evidence.

that pops into their heads rather than employing System 2 processes, then perhaps this amounts to being less inclined to philosophical thinking.

Second, appeals to a cognitive gap—to a cognitive trait that women tend to have less of than men and which arguably has an important role in philosophical practice—have some explanatory merit over other explanations that have dominated the literature on the gender gap, such as stereotype threat and implicit bias.[10] These other explanations require generalising from research conducted in other fields or in the laboratory, and it is not yet clear to what extent it is legitimate to extrapolate to Philosophy. According to a recent review of the research into stereotype threat and implicit bias "there is little empirical evidence of their effects within Philosophy" (Thompson 2017, p. 5). It also remains unclear why these mechanisms would have had more of an effect on women in the field of Philosophy than in other disciplines. In contrast, appealing to a cognitive gap helps to explain the distinctive situation of Philosophy. Indeed, it fits with the finding by Livengood et al. that the opposite pattern can be found in Psychology (a field where women are significantly over-represented): those with more psychological training tend to exhibit lower CRT scores (2010, p. 328, n. 10). Livengood et al. do not attempt to explain this finding, nor do they report data for other disciplines. However, a defender of the Quick Conclusion might hypothesise that whilst women trickled into Psychology as the negative effects of discrimination and stereotype threat were gradually overcome, a matching trend did not happen in Philosophy because additional obstacles were (and remain) present. One such additional obstacle might be that women tend to lack an important aspect of the personality required to engage properly in philosophical practice.

Moreover, the research by Livengood et al. removes one obstacle to pursuing cognitive gap explanations for the under-representation of women in Philosophy. It has been argued by Thompson (2017, p. 3; 10, n.5) and Lemoine (2017) that it is not worth pursuing cognitive gap explanations for the gender gap in Philosophy because we do not know *which* cognitive abilities are correlated with philosophical aptitude. But since the research by Livengood et al. suggests one such correlation, this particular obstacle to pursuing cognitive gap explanations is now removed.

If this explanation for the gender gap in Philosophy is accepted, it might be seen to exonerate Philosophy departments of the need to put in place much-needed strategies for promoting gender diversity. If women are simply not up to doing Philosophy, there is little point in investing time and effort into making Philosophy departments more hospitable places for women. My view is that this would be the wrong response to the empirical research, since there are many plausible interpretations of the findings that avoid this implication. In the remainder of this paper, I show how thoughtful reflection on the research points against the Quick Conclusion, towards other interpretations that would necessitate different practical responses.

In order to properly assess the Quick Conclusion, it will be helpful to disaggregate it into several different claims that are at stake. To begin with, we can note that talk

[10] For discussions of the impact of stereotype threat and implicit bias in Philosophy, see Beebee (2013), Beebee and Saul (2011), Goguen (2016) and Saul (2013). For criticism of appeals to stereotype threat and implicit bias, see Hermanson (2017). In support of the possibility that a cognitive gap might explain some gender and race differences, see Summers (2005) and Winegard et al. (2017).

of the 'ideal philosophical personality' and a trait being 'valuable in Philosophy' is ambiguous, allowing for either a descriptive or normative interpretation:

> *Descriptive Ideal Philosophical Personality Hypothesis* (IPP^D): The CRT tracks something that is currently valued within the discipline of Philosophy—an aspect of what is (consciously or unconsciously) viewed as part of the 'ideal philosophical personality'—which women tend to lack. The gender gap in the CRT is therefore explanatory for the gender gap in Philosophy.[11]

> *Normative Ideal Philosophical Personality Hypothesis* (IPP^N): The CRT tracks something that (as a matter of fact) is a valuable philosophical trait – an aspect of the ideal philosophical personality – which women tend to lack. The gender gap in the CRT is therefore explanatory for the gender gap in Philosophy.

Even if both of these claims were true, it would not necessarily result in the following, action-guiding claim that is part of the Quick Conclusion:

> *Inaction Conclusion*: Philosophy departments and wider society are exonerated of the need to institute or maintain practices intended to decrease the gender gap in Philosophy.

In what follows, I will give reasons to question all three of these claims. However, the only claim that we can dismiss with confidence is the Inaction Conclusion. As I will discuss below, neither IPP^D nor IPP^N entails the Inaction Conclusion. It is worth at the outset pointing to one reason why this is so: the gender gap in CRT may be caused by environmental factors (as opposed to it being part of the 'female nature' that there is a tendency to exhibit less of the trait(s) tracked by the CRT). If this is the case, then action is still required. This would primarily need to take place *outside* Philosophy departments, in wider society, in order to rectify widespread, far-reaching structural injustices that result in women's poorer performance at this cognitive skill. The reason that changes *within* Philosophy capture more of my attention in what follows is simply that as philosophers, there is more that we can do to make an impact within the discipline than we can in society as a whole.

6 Responses

6.1 Does the CRT track what it is claimed to track?

The CRT has been seen as an indicator of rationality (Stanovich 2011; Toplak et al. 2011, p. 1283), reflectivity (Livengood et al. 2010; Szaszi et al. 2017, p. 208) and

[11] The 'ideal philosophical personality' in the sense of IPP^D is somewhat similar to the idea of the "philosophical personality" discussed by Peña-Guzmán and Spera (2017). They see the "philosophical personality" as "the profile of the contemporary philosopher that emerges from the organization and interaction of two specific forces" (Peña-Guzmán and Spera 2017, p. 911). First, the philosopher as *imago*—"the figure of the professional philosopher who has succeeded by the standards established by his field" (2017, p. 914). Second, the philosopher as *idea(l)*—the mental representation that philosophers have of 'the philosopher'. Since both philosopher as *imago* and philosopher as *idea(l)* are dictated by current sociological trends, neither term captures my normative understanding of the ideal philosophical personality (IPP^N).

analyticity (Sytsma 2016; Stahl and van Prooijenb 2018). These are all traits that have prima facie plausibility as part of the ideal philosophical personality.[12] Yet it is far from clear whether we can straightforwardly associate CRT performance with these traits. Two alternative possibilities for what the CRT tracks stand out in the literature: numeracy and/or confidence.

6.1.1 Numeracy

Numeracy is one's ability to store, represent and process mathematical operations (Peters, 2012). It has been widely discussed how difficult it is to disentangle cognitive reflection from numeracy (Thomson and Oppenheimer 2016, p. 101). All three test questions involve numbers, lending prima facie plausibility to the suggestion that the CRT tracks numeracy. There also exists a large body of research suggesting that the CRT measures both cognitive reflection *and* numeracy.[13]

In Frederick's original study, only one other cognitive test showed a gender difference—the SAT maths scores (Frederick 2005, p. 37). Frederick comments that "men generally score higher than women on math tests" and he cites various studies from the 80s and 90s to support this claim. As I will discuss below, there is now strong counter-evidence to this. However, some studies do continue to point towards gender differences in maths ability, particularly as age of participants and complexity of the test increases (e.g. Ganley and Vasilyevam 2014; Lindberg et al. 2010, p. 1132; Benbow et al. 2000). Primi et al. (2018, pp. 261–262) suggest that the strongest available evidence for gender differences in maths performance comes from the Programme for International Student Assessment (PISA), which assesses the competencies of 15 year old students from 65 countries in various subjects, including Mathematics. On average across OECD countries, boys outperform girls in Mathematics by eight score points. The difference is most notable amongst the highest achieving students: the highest-scoring 10% of boys score 16 points higher than the best-performing 10% of girls (OECD 2016, p. 196).[14]

If there is a numeracy gender gap, it seems plausible that this might be explanatory for the CRT gender gap. This explanation is supported by research by Thomson and Oppenheimer (2016). They piloted the 'CRT-2', a test designed to measure cognitive reflection whilst avoiding conflation with numeracy. The CRT-2 uses "trick questions" that "do not require a high degree of mathematical sophistication" (2016, p. 101).[15] 200 participants were tested on both the CRT and the CRT-2 and it was found that the gender gap significantly lessened on the CRT-2. Whilst men (M = 65.9% correct) significantly

[12] When I refer to the 'ideal philosophical personality' or the 'Ideal Philosophical Personality Hypothesis' without making specific reference to IPP^D or IPP^N, my comment applies to both versions.

[13] In the studies listed by Thomson and Oppenheimer (2016, p. 101) to support this point, correlations between the CRT and numeracy ranged from 0.31 to 0.51. More recently, Szaszi et al. (2017, p. 207) have argued that "the CRT is a multi-faceted construct: both numeracy and reflectivity account for performance".

[14] The mean score for Mathematics across all countries was 490 (OECD 2016). About two-thirds of all students across OECD countries score between 400 and 600 points (OECD 2018).

[15] Here is one question from the test: "Emily's father has three daughters. The first two are named April and May. What is the third daughter's name? (intuitive answer: June; correct answer: Emily)" (Thomson and Oppenheimer 2016, p. 101).

outperformed women (M = 36.0% correct) on the original CRT ($p < 0.001$), men (M = 60.5% correct) and women (M = 53.3% correct) were not reliably different on the CRT-2 ($p > 0.05$) (Thomson and Oppenheimer 2011, pp. 106–107). This finding is consistent with differences in numeracy being a cause of the gender gap on the original CRT.

This explanation is further supported by a recent study by Primi et al. (2018), which found that the direct effect of gender was no longer statistically significant once the variables of mathematical reasoning and maths anxiety were taken into account. Additionally, Szaszi et al. (2017) suggest that we simply cannot separate numeracy from reflectivity on the CRT, since good numeracy is likely to deliver you the right intuitions from the start. Indeed, it is notable from reading their examples of participants' vocalised thought processes that 'Correct Answer' respondents often recognised that there was an equation that needs solving in the bat and ball question (Szaszi et al. 2017, p. 218).

The research at present does not, however, lead us to a position where we can say that gender differences in the CRT can be *entirely* explained via gender differences in numeracy. Firstly, we should note that Thomson and Oppenheimer's CRT-2 has not gained popularity, nor is it agreed whether it tests the cognitive skill that behavioural economists and psychologists have become so interested in. As Primi et al. (2018, p. 274) point out, the correlations between the CRT-2 and various measures of rational thinking and decision-making skills were generally weaker than the correlations between these measures and the original CRT. Secondly, other studies, including Frederick's original study, claim to have controlled for numeracy and yet found that a significant gender gap remains (Frederick 2005, p. 37; Agnew 2017, p. 12). Thirdly, it is far from clear to what extent there is, in fact, a gender numeracy gap. In their meta-analysis of 242 studies published between 1990 and 2007, representing the testing of 1286,350 people, Lindberg et al. (2010, p. 1131) conclude that "there is no longer a gender difference in mathematics performance". This is consistent with Hyde et al.'s (2008) study, which (using data from over seven million students) found no evidence of gender differences on US state math tests among students between Grade 2 and Grade 11. Where gender differences in favour of males are seen (for example, in complex problem-solving at high school level), these differences appear to be attributable to multiple possible environmental explanations (for example, that parents and teachers give higher ability estimates to boys than girls, and that patterns of interest are affected by cultural influences) (Lindberg et al. 2010, p. 1132). This latter possibility would also help explain why gender differences in maths differ across countries, as well as the fact that these differences correlate with gender inequality measures for those countries (Else-Quest et al. 2010; Guiso et al. 2008; Penner 2008).

Nevertheless, a consensus does seem to have developed that numeracy is at least one component in performance on the CRT (Thomson and Oppenheimer 2016, p. 101; Szaszi et al. 2017, p. 207; Primi et al. 2018). What is the significance of this for explaining the gender gap in Philosophy?

It might be that the CRT tracks numeracy, and numeracy is required for success in Philosophy.[16] This fits with the high regard that philosophers have historically

[16] Talking of 'success in Philosophy' is ambiguous. It could refer to staying on in the discipline through the levels, eventually becoming a professional philosopher with publications and a permanent job. Or, it

held for mathematics. It may be that maths skills are closely related to philosophical skills, particularly those required for Logic, which is often a compulsory component of Philosophy programmes. Evidence suggests that studying advanced mathematics develops some aspects of conditional reasoning, including the ability to reject invalid inferences (Inglis and Attridge 2016, p. 130), and so there is good reason to think that maths skills and logical skills are linked. Some have even argued that mathematical competence is crucial to good Philosophy. Boghossian and Lindsay (2016) declare that "If you want to be a good philosopher, don't rely on intuition or comfort. Study maths and science." Their reason is that "Philosophers who can think like mathematicians are better at clear thinking, and thus philosophy."

However, evidence supporting this view seems rather sparse. As Thompson (2017, p. 3) says, the extent to which maths skills are *required* for success in Philosophy is not yet clear. Moreover, evidence of good numeracy is rarely, if ever, an entry requirement for university Philosophy programmes.[17]

Given the research at present, it is unclear (i) whether women are worse at numeracy, (ii) the extent to which the CRT measures numeracy and (iii) whether numeracy is required for success in Philosophy. We therefore cannot adequately justify the conclusion that women's tendency towards a low CRT score represents low numeracy, which contributes to their low participation in Philosophy.

6.1.2 Confidence

Some have praised the CRT for being a "performance measure rather than a self-report measure" (Toplak et al. 2011, p. 1275), but this neglects the important effect that self-perception of one's abilities can have on performance. It may be that the CRT tracks *confidence* in numerical abilities rather than (or in addition to) *actual* cognitive abilities. Zhang et al. (2016, p. 427) found that when differences in quantitative self-efficacy (perceived fluency with numerical information) are controlled for, gender differences on the CRT disappear. They conclude that "men perform better on the CRT because they are more confident in their quantitative abilities" (2016, p. 427).

This is consistent with research on maths anxiety and gender differences, which has found that females suffer more from maths anxiety than males (Else-Quest et al. 2010; Devine et al. 2012). Ganley and Vasilyevam (2014)'s research suggests that female's heightened worry on maths tests utilizes their visuospatial working memory resources, leading to poorer performance. This would fit with Szaszi et al.'s (2017) suggestion (discussed in Sect. 2.1) that those answering the CRT questions incorrectly may be failing to bring to mind the strategic rules needed to solve the questions.

Footnote 16 continued
could refer to producing new, plausible ideas that take us closer to the truth, or inspiring others to engage thoughtfully in philosophical issues, or some other measure of what it means to be a successful philosopher that is not dictated by one's success in the academy. These two kinds of success could, and perhaps sometimes do, come apart. Where I wish to distinguish between these two kinds of success, I refer to the first kind of success as 'successful progression in the field' and to the second kind of success as 'good philosophising'.

[17] Of course, this does not mean that it is *right* that maths skills are ignored as a selection criterion in Philosophy. In the UK, the largest 'drop-off' of women tends to happen between undergraduate and Masters level (BPA 2011, p. 9). One (amongst many) possible explanations of this is that some female undergraduates find that they are just not 'up to it', because of reasons linked with their poorer numeracy.

It also fits with the wider picture given by research on confidence, which has suggested that women tend to have lower levels of self-confidence than men.[18] We might hypothesise that pursuing Philosophy to higher levels requires a degree of confidence in one's abilities that women are less likely to possess. There is at least some prima facie reason to think that confidence contributes to successful progression within the field. For example, the level of confidence with which you deliver your question or paper, or the conviction with which you profess your conclusion, is likely to affect the way that it is received by others (see Schwitzgebel (2010) on the potential effects of "being good at seeming smart"). Additionally, effectively 'batting away' opponents requires not just intellect, but an element of performance (Larvor 2015). As Justin Weinberg (2015) comments, most graduate students in Philosophy are advised to "project confidence". Perhaps women's poorer performance on the CRT tracks their high anxiety and low confidence, and these traits affect their levels of participation and performance in Philosophy.

However, a concern with this line of reasoning is that Zhang et al.'s study, like many others, does not account for the possibility that people's beliefs about ability are accurate (Lemoine 2017; Jussim 2012). That is, the self-report measure of quantitative self-efficacy may track numeracy, because the people that lack confidence in their quantitative abilities do so because they are, as a matter of fact, less competent at numeracy. This is consistent with research by Primi et al. (2018, p. 273), which found a direct link between maths anxiety and cognitive reflection, but found that the effect of maths anxiety on cognitive reflection was partially mediated by mathematical reasoning.

If quantitative self-efficacy is strongly linked with actual mathematical ability, then we are back to our unanswered question of whether numeracy is relevant to success in Philosophy.

6.1.3 Implications

The research discussed in this section does not point to clear conclusions about what the CRT tracks. Nor is it clear what the relevance to explaining the gender gap in Philosophy would be. However, it does suggest that we should, at the least, be sceptical about a straightforward equating of CRT score with rationality, reflectivity or analyticity. It therefore attacks a version of IPP that suggests that it is a lack of these particular traits that holds women back in Philosophy.

The discussion so far has not attempted to deny that there may be traits that women tend to lack which might help explain the gender gap in Philosophy. Rather, it has explored the possibility that the CRT tracks numeracy or confidence. The absence of relevant empirical research on the roles that numeracy and confidence play in Philosophy means that we are unable to say to what extent these attributes are currently valued in Philosophy and whether they contribute to successful progression in the field as it stands. There is, however, some anecdotal evidence suggesting that confidence

[18] For example, Bleidorn et al. (2016) found that across 48 countries, males consistently reported higher self-esteem than females. Thompson et al. (2016, p. 9) found that female students taking Philosophy classes reported feeling less confident in their ability to do well in Philosophy than did men.

might contribute to successful progression in the field, lending at least some, limited support to IPP^D.

6.2 Is the trait tracked by the CRT something we should value in philosophers?

There is clearly a question mark over *what* the CRT tracks. But whatever it tracks, this is something that women tend to have less of than men and philosophers tend to have in abundance. So, we can raise a second question asking why we should think that the CRT tracks something that we should value in philosophers. That is, even if IPP^D is true, why should we think that IPP^N is true?

The idea that the CRT tracks something we should value in philosophers seems to be assumed by Livengood et al. When talking of the 'philosophical personality', the authors say that they seek only to describe "who philosophers are" (2010, p. 314). But at points they slip from this descriptive exercise by implicitly adopting the normative assumption that they have identified a philosophical virtue. For example, they imply that what the CRT tracks is part of the *expertise* of philosophers (2010, pp. 319, 320).

But who philosophers *are* and who philosophers *should* be are different questions. The fact that some norm exists amongst philosophers which correlates with their good performance on CRTs does not, in itself, tell us that this trait is an asset to philosophising. Imagine that there was evidence suggesting that philosophers are more likely to exhibit social awkwardness than non-philosophers. It would be wrong to conclude from this research that social awkwardness is part of the ideal philosophical personality (even in the sense of IPP^D, for this trait might appear accidentally, rather than being (consciously or unconsciously) selected for). Rather, this trait is irrelevant (or even detrimental) to good philosophising.

Similarly, we might generalise from the finding about CRT tracking quantitative self-efficacy to say that philosophers have a tendency to be more confident about their cognitive abilities. But again, this attribute does not necessarily make for better philosophising. The philosophers discussed in the previous section who have offered anecdotal support for the role of confidence in Philosophy have tended to see this as a *flaw* in currently philosophical practice—a mark of a deep methodological problem with the way that Philosophy currently operates (Larvor 2015). Indeed, one might even think that those with lower confidence actually make for better philosophers, because they may be more open to counter-arguments. When evaluating IPP^N, the salient question in assessing the relevance of the CRT should be whether whatever it tracks is an epistemologically relevant trait, one that we should value as conducive to the pursuit of knowledge (or whatever we see as the aim of Philosophy).

This idea that certain traits might be dominant in Philosophy without necessarily being conducive to good philosophising becomes more plausible when we consider the flaws in the supposedly meritocratic system used to select philosophers (onto courses, and into posts). It has been well-discussed that meritocratic selection may be subject to biases at the level of deciding whether a candidate fulfils certain criteria.[19] But

[19] For an overview of some core studies on bias in appointing for faculty and leadership roles, see Corrice (2009).

it may also be that there is bias present in deciding *what these criteria are*.[20] The 'success criteria' of what it is to be a good philosopher are (at least partially) decided by those already successful in the discipline, so that the norms and values of these individuals are reproduced in those selected, in a kind of feedback loop (Jenkins 2013). For example, Haslanger (2008, p. 217) and others have expressed concern over the dominance of a hyper-rational norm in Philosophy, which is often taken to represent the high-end of the discipline, but which may not necessarily contribute towards good philosophising.

So, it may be that the CRT tracks trait T, and those possessing T are (intentionally or unintentionally) more likely to be recruited to Philosophy. But this does not, in itself, tell us that T is important for good philosophising. This 'irrelevant trait hypothesis' resists the move from IPP^D to IPP^N, as it suggests that although the trait(s) tracked by the CRT may be part of the philosophical personality, this does not mean that they are part of the *ideal* philosophical personality. It suggests that the CRT tracks a trait that is not relevant to good philosophising, but either (1) just so happens to be well-represented in philosophers, despite not being selected for (as in the social awkwardness example) or (2) is unconsciously or consciously selected for because it is mistakenly thought to be part of the ideal philosophical personality (as in the confidence and hyper-rationality examples). If this were the case, then we certainly should not settle for the Inaction Conclusion. Rather, we should seek changes to the status quo in the discipline, such as re-evaluation of the criteria used when assessing applicants for Philosophy jobs.

This response has flagged that there is an open question as to whether we should be valuing whatever it is that the CRT tracks. But there are difficulties with pursuing this 'irrelevant trait hypothesis'. Though there may be scope for debate over the purposes and methodology of the discipline, there is also wide agreement that Philosophy aims at the truth. The person who does badly in the CRT gets the wrong answers, and philosophers are after right answers. Moreover, as has been discussed, it seems plausible to say that it is part of good philosophising to engage in careful reflection over one's intuitions, and to be especially immune to biases of judgement. We therefore might not want to press too hard with the idea that there is *nothing* of value in what is tested by the CRT.

6.3 How important is this trait to good philosophising?

We might concede that the CRT tracks something of value, but argue that it is only one small part of the cognitive skills that contribute to good philosophising.

Imagine a test used to assess physical fitness for the military that has press-ups as the key element. Since women tend to have lower levels of arm strength than men, they might find it harder to pass this test. But it would be wrong to conclude that the women who fail this test are 'physically unfit'. Arm strength is just one small part of

[20] Studies suggest that people alter what criteria they say are relevant for a particular job according to the characteristics of the person that they want to hire (Uhlmann and Cohen 2005; Luzadis et al. 2008). On the inherent difficulties with neutrally assessing merit in specific domains, see Crosby et al. (2003), Kane (1998), and Cicchetti (1994).

physical fitness; core strength and endurance also have an important role. In the same way, we might allow that the CRT tracks one aspect of rationality that women tend to have less of than men, but without drawing any conclusions about *overall* levels of rationality.

In our military fitness example, the important practical question is whether a certain level of arm strength is required for success in the military. Analogously, the salient question for us is whether the aspect of rationality potentially tracked by the CRT is an essential element in good philosophising. There seem to be good reasons to think that it is not, and rather, that the type of reflectivity tracked by the CRT is only one, fairly minor skill utilised by philosophers. It might make for a good start to one's philosophical project to begin with sound thoughts that have already been subject to some System 2 scrutiny, but it seems that the bulk of philosophical work comes later.

Consider how Livengood et al. set the scene for explaining the aspect of the philosophical personality that they are interested in:

> An intuition is a spontaneous intellectual sensation: p seems to be true without being consciously inferred. In considering the first question of the CRT, for example, it intuitively seems that the answer must be 10 cents. Similarly, in the Gettier case, it intuitively seems that the agent does not have any knowledge… (Livengood et al. 2010, p. 318)

There seems something odd about this analogy. In the CRT, intuition delivers the wrong answer, and getting the right answer requires *overriding* intuition (rather than making use of it). In the Gettier case, we have an intuition which then becomes the subject of further philosophical exploration. By presenting thought experiments invoking certain intuitions, Gettier's (1963) paper far from closed the question of whether knowledge is justified true belief. Rather, it was the starting point of an ongoing philosophical project. Further philosophical work has consisted in: (i) suggesting additional conditions that might be added in order to avoid Gettier cases, such as the 'no false lemmas condition' (e.g. Armstrong 1973, p. 152; Clark 1963), (ii) engaging in further thought experiments to question the conditions for knowledge (e.g. Goldman's (1976) 'fake barn' cases), (iii) making distinctions within 'justification' and exploring what it takes for a belief to be justified (e.g. Feldman and Conee 2001), and (iv) suggesting alternative accounts of what constitutes knowledge, such as reliabilism (e.g. Nozick 1981). If it makes sense to talk of 'getting the right answer' to a Gettier case, arriving at this 'right answer' when the case is first presented seems to be only a small and insignificant part of the process, and it is not clear to what extent getting the answer wrong at the start would be damaging to the long-term philosophical project.[21] Philosophers have far more than the few seconds or minutes spent on the CRT questions to properly evaluate Gettier cases and come to a judgement on what knowledge really consists in. Thus although there might be *something* in the reflectivity that is tested in the CRT, it seems like there is another, broader type of reflectivity that is of more value and importance to

[21] One way it could be damaging is if further reflection (or use of System 2 judgement) is used as a kind of 'press secretary' for the initial intuition (or System 1 judgement), such that the philosophical project consists in providing justifications for our initial, unconsidered judgements. On the idea of reason as a 'press secretary' for our existing judgements, see van Mulukom (2018) and Haidt (2012).

the long-term philosophical project—perhaps one involving an indefatigable pursuit of answers, even where these are particularly hard to find.[22]

The above discussion has given just one reason to question the relative value of the trait(s) tracked by the CRT compared to other traits that are potentially part of the ideal philosophical personality. Given the precise nature of the CRT questions, set against the range of virtues and skills that we might plausibly postulate as part of the ideal philosophical personality, my view is that we probably need not hang too much on whatever the CRT tracks. Not all philosophers perform well on the CRT and so it is, at the least, possible to successfully progress in the field whilst lacking this particular skill. And even if the trait that the CRT tracks *contributes* to good philosophising, it is far from clear that this trait is *essential* to good philosophising and therefore to the ideal philosophical personality in the sense of IPP^N.

Moreover, it could be that there is a correlation between possessing above-average levels of analyticity (or whatever we suppose it is that the CRT tracks) and lacking other skills that are valued amongst philosophers, such as creativity. This is purely speculative, but it is conceivable that a high CRT score comes at the expense of other virtues that we need more of in Philosophy.[23] Kahneman says that "absence of bias is not always what matters most" (2011, p. 192), and this surely applies to Philosophy. It could be that relief from the constraints of analyticity allows for more creative thinking, increasing the likelihood of hitting upon unusual, divergent ideas. If this were true, low CRT score should not be viewed as indicative of a poor philosopher.

Reflecting back on the military fitness example may be helpful here. Let us say that (1) men have more arm strength than women, (2) military personnel have more arm strength than those outside the military, (3) there are more men than women in the military and (4) there is a good prima facie case for thinking that arm strength contributes to doing your military service well. This state of affairs is perfectly consistent with there being other attributes that contribute to success in the military that women have more of than men (for example, endurance or emotional literacy). If this were the case, in addition to checking that entry tests for the military are not overly-focused on arm strength, it would also be important to look at how other factors such as discrimination and unconscious bias might be contributing to the under-representation of women.

Applying the same reasoning to our case: Even if it is true that the trait T tracked by the CRT is currently valued amongst philosophers (i.e. there is some truth to IPP^D), and even if possessing T does, as a matter of fact, make *some* contribution to good philosophising (i.e. there is some truth to IPP^N), it would *still* be wrong to think that

[22] The ideas in this paragraph are heavily influenced by Emily Perry's excellent response to this paper at the London-Berkeley Graduate Conference 2018.

[23] The idea that diversity amongst participants will benefit the discipline itself has been argued for in relation to other disciplines, particularly Science (Rubin and O'Connor 2018; Harding 1991). Rubin and O'Connor (2018) outline the potential benefits of diverse collaborations in Science, pointing to research by Zollman (2010) suggesting that a diversity of beliefs within an epistemic community is key to ensuring that the group eventually arrives at true beliefs. Moreover, Page (2008) and Phillips et al. (2006) have found that a diversity of perspectives can aid complex problem-solving, as well as creative work. It seems plausible that these arguments can be extended to Philosophy, in order to say that a diversity of philosophers may positively influence the methodology, content, outcomes and practice of the discipline. For examples of this kind of argument applied to Philosophy, see Wylie (2011). For criticisms of using this kind of argument to justify affirmative action, see Anderson (2002).

the empirical research *entirely* explains the gender gap in Philosophy. Since T is one amongst many possible philosophical virtues and skills, women tending to exhibit less of T should not be having such a dramatic effect as to produce the wide gender gap we see in Philosophy. If that is not the case, and in fact it *is* a significant factor, because possessing trait T is wrongly being prioritised as an important selection criterion, we might speculate that this is detrimental to the discipline of Philosophy, since prioritisation of T might come at the expense of other valuable philosophical virtues and skills. Regardless of the truth of this last hypothesis, the Inaction Conclusion would be unjustified. Rather, as in the military case, we should turn a critical eye to entry criteria, as well as onto whether there are other obstacles to women's participation such as discrimination and unconscious bias.

6.4 How should we understand the causal story?

Lastly, and importantly, we should note that the direction of causation has not been established between CRT scores and philosophical training. We cannot say whether it is philosophical training that leads to the increased CRT score amongst philosophers, or whether possessing the trait(s) tracked by the CRT to a high degree leads people to undergo more philosophical training.[24]

At least three possibilities explain the current data. Firstly, it may be that people with higher CRT score are more likely to take up further philosophical training (Fig. 1).[25] Since women tend to have lower CRT score, fewer women continue in Philosophy.

Secondly, it might be that the two facts are independent and we should draw no conclusions from the gender gap in CRT score and the increased CRT score of philosophers (Fig. 2).

However, given all that has been said so far, it seems unlikely that these facts are *entirely* independent. Imagining our social awkwardness example to be true, we would probably want to posit at least *some* causal relation between the phenomena. For example, we might hypothesise that you need to be clever to be a philosopher and being clever makes it harder to talk to other people. Analogously, there is likely to be *some* causal story that can be told between the gender gap in CRT score and the gender gap in Philosophy.

Thirdly, it could be that practising Philosophy brings up your CRT score, but fewer women are continuing with Philosophy (for reasons unrelated to CRT score) (Fig. 3). Women may be put off staying in Philosophy by contingent features of the discipline in its present state, features that are amenable to change by the actions of university faculties.[26] If this were the case, it would be an injustice that we should actively seek to rectify, for women would be missing out on opportunities to develop their capacities in whatever it is that the CRT tracks.

[24] Livengood et al. say that "our data do not tell us how philosophers came to be more reflective than their peers" (2010, p. 319).

[25] I borrow the terms 'selectionist' and 'educationist' from Livengood et al.'s (2010) discussion of the causal relationship between philosophical training and cognitive reflectivity.

[26] For discussion of potential ways that women are deterred from continuing in Philosophy and some suggestions for interventions, see Demarest et al. (2017), Figdor and Drabek (2016), Saul (2013) and Thompson et al. (2016).

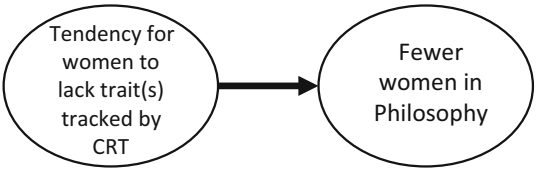

Fig. 1 Simple selectionist

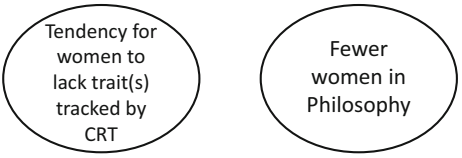

Fig. 2 Independent

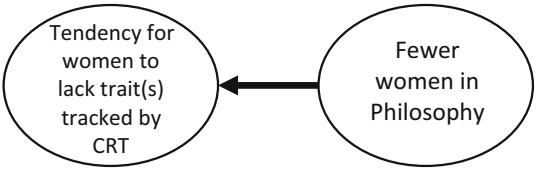

Fig. 3 Simple educationist

This third explanation denies both IPP^D and IPP^N, since it denies that the gender gap in the CRT is explanatory for the gender gap in Philosophy. It says that although the CRT may track a trait that is part of the philosophical personality, it is the study of Philosophy that nurtures this trait, and so we must look elsewhere for explanations of why women tend not to continue studying Philosophy beyond their tendency to have a lower CRT score.

However, given the small number of participants in Philosophy, clearly the gender gap in Philosophy cannot account for the gender gap in CRT score on its own. If the hypothesis that the direction of causation runs this way is to be at all plausible, we would need to speculate that Philosophy is one of a number of disciplines or activities that improve CRT score and which men are more likely to engage in than women (Fig. 4).

The causal story behind the gender gap in Philosophy is likely to be far more complex than is allowed by any of the possibilities discussed so far. An intelligent supporter of the Ideal Philosophical Personality Hypothesis would not claim that the *only* cause of the gender gap in Philosophy is that women lack the aspect of philosophical personality tracked by the CRT. A 'perfect storm' explanation of the gender gap seems more plausible, where many factors combine to produce the dramatic gender gap we see in Philosophy (Antony 2012). Figure 5 illustrates this with some hypothetical (but plausible) examples of other causal factors.

The interesting question is then whether the tendency for women to exhibit less of the trait(s) tracked by the CRT is one cause amongst many. Where a factor F is one cause amongst several (mutually independent) causes, we should be able to vary the

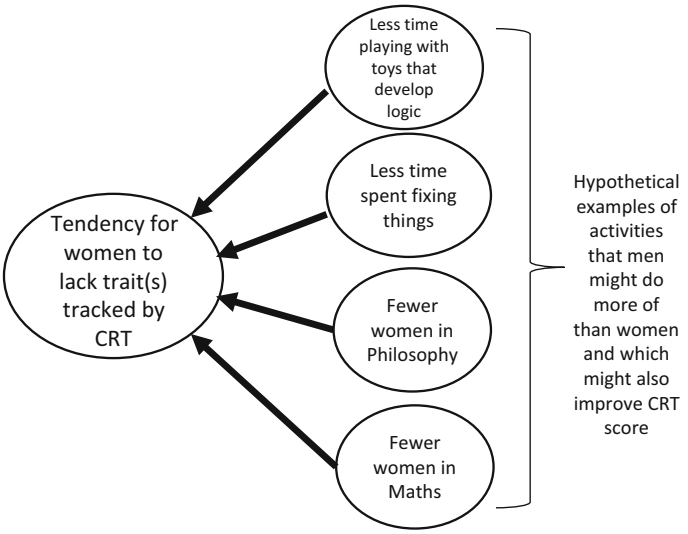

Fig. 4 Complex educationist

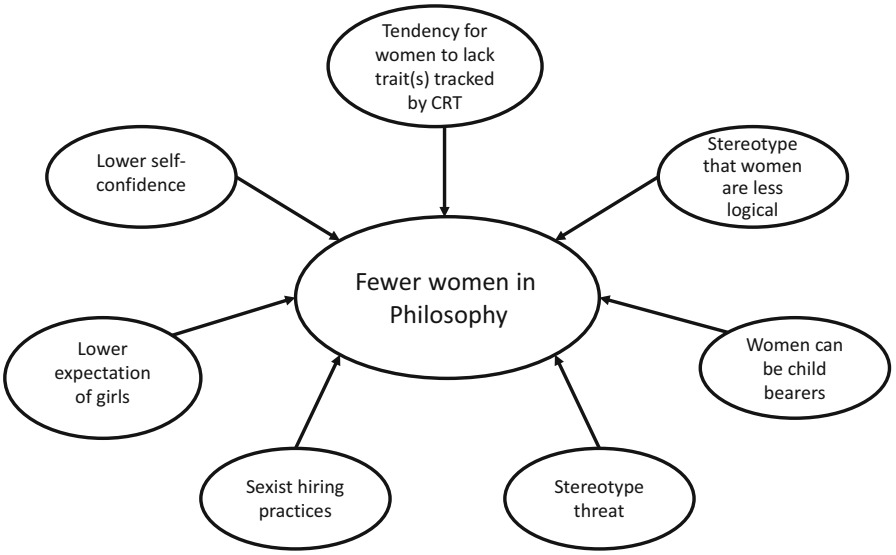

Fig. 5 Complex selectionist

other causes without this leading to a change in F. And yet, it is not clear that this would be the case here. For example, it seems plausible that were we to vary one of the social norms that contribute to the gender gap in Philosophy, this would *also* lead to a change in the CRT gender differential. In that case, this social norm would be a common cause of both the CRT gender differential and the gender gap in Philosophy, and the tendency for women to exhibit less of the trait(s) tracked by the CRT would not be a mutually independent cause of the gender gap in Philosophy.

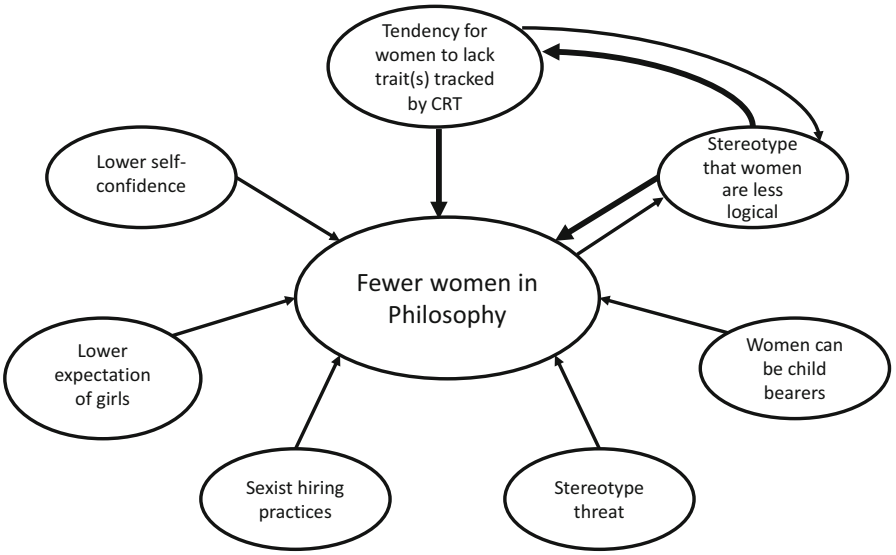

Fig. 6 Example of a more complex causal story

To make this thought more concrete, we can take as an example the stereotype that women are more intuitive and less logical. This stereotype might make women less likely to imagine themselves as philosophers, with the consequence that they are less likely to continue in the discipline (Demarest et al. 2017). In that case, this stereotype is a causal factor in the gender gap in Philosophy. But the stereotype might also contribute by a more indirect route. For example, the stereotype may have the effect that adults are less likely to give girls toys that develop logic (Oksman 2016), with the consequence that girls have fewer opportunities to develop skills at whatever the CRT tracks. In that case, the stereotype acts as a causal factor in the CRT gender differential, which then feeds into women appearing less likely to have the 'philosophical personality' and there being fewer women in Philosophy. The gender gap in Philosophy, as well as the poorer performance of women on tasks like the CRT, would then provide further evidence for the stereotype. So, the stereotype would be causing several environmental interventions, which have effects that validate the stereotype (Fig. 6).[27]

Stories like this, where there is a kind of causal feedback loop operating between different factors, seem plausible. To endorse this particular story would be to endorse a version of the Ideal Philosophical Personality Hypothesis, since it allows that there are fewer women in Philosophy partially as a result of women tending to lack an aspect of the philosophical personality. But it is a story that points against the Inaction Conclusion, because it blames women's low CRT score not on innate differences in aptitude, but on contingent structural norms and cultural practices that would lessen

[27] The purpose of this diagram is to illustrate a possible causal feedback loop; it does not represent anything like the true level of causal complexity. For example, the fact that women are potential child-bearers might contribute to the stereotype that women are less logical, because being intuitive and 'following your instincts' is associated with pregnancy and childbirth. This stereotype might then feed into the lower expectations that society has of girls, which then feeds into women's lack of self-confidence.

or disappear in a fairer, more equal society. An implication of this is that tackling the structural injustice leading to the gender gap in CRT score would require far more than simply making changes within the discipline of Philosophy.

Whatever we think of that story, it should at least be clear that a straightforward causal arrow from CRT aptitude to the gender gap in Philosophy is highly implausible. The causal story is likely to be far more complex than any of the initial hypotheses allowed.

7 Conclusion

We have seen that there are several routes by which we can argue against the Quick Conclusion. Firstly, we can dispute whether the CRT tracks what it is claimed it tracks. Secondly, we can question whether the trait tracked by the CRT is something we should value in philosophers. Thirdly, even if we allow that the CRT does track a valuable trait, we can question how important this is when compared to all the other traits that contribute to good philosophising. Lastly, we should question the implausibly simplistic causal story that crude versions of the Ideal Philosophical Personality Hypothesis imply.

The empirical research in this area is still in its infancy and there remain many unanswered questions. It is unclear exactly what the CRT tracks, and the extent to which whatever it tracks is currently selected for when recruiting philosophers onto courses and into posts. It is therefore impossible to draw firm conclusions about the truth of IPP^{D}. However, if IPP^{D} is true, it is only plausible to claim that the gender gap in the CRT is *somewhat* explanatory for the gender gap in Philosophy. Women exhibiting less of the trait tracked by the CRT will be one amongst many factors, and it is likely that there will be a number of interaction effects between these causal factors, resulting in a complex causal story where causal connections run in multiple directions.

A plausible case can be mounted against IPP^{N}, particularly when we think about the range of philosophical virtues and skills that plausibly might constitute the ideal form of the philosophical personality. This, however, is a matter for debate; there is, at the least, a prima facie case for thinking that the skill tracked by the CRT has at least *some* value for good philosophising (though this may be offset if it comes at the expense of other, valuable traits). What we can say with confidence is that the interpretation of the empirical research which says that there are fewer women in Philosophy because women *naturally* lack the personalities required to be good philosophers is unconvincing.

Rather than endorsing one particular response, the intention of this paper is to open up discussion of how best to make sense of the research as it currently stands, and to prompt reflection on what practical responses are appropriate in light of the different hypotheses. For example, if it is right that poor CRT performance is an indicator of low confidence, then this would add urgency to the already growing cries for finding ways to increase self-confidence amongst women. Even simple interventions, such as giving more explicit encouragement to undergraduates (Saul 2013, p. 51) or emphasising the

importance of effort rather than 'brilliance' (Thompson et al. 2016), might partially stem the flow out of Philosophy's leaky pipeline.[28]

Although not all the routes discussed have dismissed the part of the Quick Conclusion that claims that the gender gap in the CRT is explanatory for the gender gap in Philosophy, all routes imply that it would be the wrong response to the empirical findings for Philosophy departments to simply relax and take no action aimed at narrowing the gender gap in Philosophy. Instead, the discussion points towards the view that making relevant structural changes to the environment (both inside and outside of Philosophy) should remain our focus when thinking about the gender gap in Philosophy.

Acknowledgements Special thanks must go to Luc Bovens, who provided helpful feedback on many different versions of this paper. Thanks to David Kinney for his help thinking about causal relationships. Emily Perry gave an excellent response to this paper at the London-Berkeley Graduate Conference 2018. Justin Sytsma engaged in email correspondence over his research, sent me his unpublished PowerPoint, and drew my attention to the Zhang et al. study. Participants of the LSE Choice Group, the Joint Session 2017, and the LSE Philosophy Research seminar all provided useful feedback, as did Bastian Steuwer, Paul Davis, Jonathan Birch and Alex Voorhoeve. I am also indebted to three anonymous reviewers for this journal who provide invaluable comments on earlier drafts of this paper.

References

Agnew, A. (2017). The role of gender, cognitive attributes and personality on willingness to take risks. *Business and Economic Research, 7*(1), 1–16.

Alexander, J., & Weinberg, J. M. (2007). Analytic epistemology and experimental philosophy. *Philosophy Compass, 2,* 56–80.

Anderson, E. S. (2002). Integration, affirmative action, and strict scrutiny. *New York University Law Review, 77*(5), 1195–1248.

Antony, L. (2012). Different voices or perfect storm: Why are there so few women in philosophy? *Journal of Social Philosophy, 43*(3), 227–255.

APA (2017). Good practices guide (Spring 2017). https://cdn.ymaws.com/www.apaonline.org/resource/resmgr/docs/Good_Practices_Guide.pdf. Accessed 02/07/2018.

Armstrong, D. M. (1973). *Belief, truth, and knowledge*. Cambridge: Cambridge University Press.

Baron, J., Scott, S., Fincher, K. S., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284.

Beebee, H. (2013). Women and deviance in philosophy. In K. Hutchison & F. Jenkins (Eds.), *Women in philosophy: what needs to change?* (pp. 61–80). Oxford: Oxford University Press.

[28] Leslie et al. (2015) found that disciplines where success in the field is viewed as requiring natural brilliance and who 'idolise' genius had lower proportions of women obtaining PhDs. Notably, Philosophy was the subject which had the highest emphasis on brilliance. See also Storage et al. (2016) and Cimplian and Leslie (2015). Thompson et al. (2016) found that women believing that success in Philosophy is rooted in brilliance affects how interested and willing to continue in Philosophy they are, in a way that it does not for men who possess similar beliefs. Thompson (2017, p. 7) cites research suggesting that the prevalence of brilliance-based beliefs about success in Philosophy increases from the first day to the end of the course, whereas the reverse is true in Psychology.

Beebee, H., & Saul, J. (2011). Women in philosophy in the UK: A Report by the British Philosophical Association and the Society for Women in Philosophy. September 2011. http://www.swipuk.org/notices/2011-09-08/Women%20in%20Philosophy%20in%20the%20UK%20(BPA-SWIPUK%20Report).pdf. Accessed 08/08/2017.

Benbow, C. P., Lubinski, D., Shea, D. L., & Eftekhari-Sanjani, H. (2000). Sex differences in mathematical reasoning ability at age 13: Their status 20 years later. *Psychological Science, 11*(6), 474–480.

Bleidorn, W., Arslan, R. C., Denissen, J. J. A., Rentfrow, P. J., Gebauer, J. E., Potter, J., et al. (2016). Age and gender differences in self-esteem—A cross-cultural window. *Journal of Personality and Social Psychology, 111*(3), 396–410.

Boghossian, P., & Lindsay, J. (2016). Want to be good at philosophy? Study maths and science. *The Philosophers' Magazine*. 29 May 2016. http://www.philosophersmag.com/essays/131-want-to-be-good-at-philosophy-study-maths-and-science. Accessed 10/07/2018.

Bourget, D., & Chalmers, D. J. (2014). What do philosophers believe? *Philosophical Studies, 170*(3), 465–500.

BPA/SWIP (2011). Good practice scheme. http://bpa.ac.uk/uploads/Good%20Practice%20Scheme/General%20guidance.pdf. Accessed 02/07/2018.

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2015). Cognitive reflection test: Whom, how, when. Middlesex University, London. https://mpra.ub.uni-muenchen.de/68049/. Accessed 01/07/2018.

Byrd, N. (2014). Intuitive and reflective responses in philosophy. *Philosophy Graduate Theses and Dissertations* 6. https://scholar.colorado.edu/phil_gradetds/6/. Accessed 02/07/2018.

Cappelen, H. (2012). *Philosophy without intuitions*. Oxford: Oxford University Press.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290.

Cimplian, A., & Leslie, S. (2015). Response to comment on 'Expectations of brilliance underlie gender distributions across academic disciplines'. *Science, 349*(6246), 391c.

Clark, M. (1963). Knowledge and grounds. A comment on Mr. Gettier's paper. *Analysis, 24*(2), 46–48.

Corrice, A. (2009). Unconscious bias in faculty and leadership recruitment: A literature review. *Analysis in Brief (Association of American Medical Colleges), 9*(2), 1–2.

Crosby, F. J., Iyer, A., Clayton, S., & Downing, R. A. (2003). Affirmative action. Psychological data and the policy debates. *The American Psychologist, 58*(2), 93–115.

Cueva, C., Iturbe-Ormaetxe, I., Mata-Péreza, E., Ponti, G., Sartarelli, M., Yu, H., et al. (2016). Cognitive (ir)reflection: New experimental evidence. *Journal of Behavioral and Experimental Economics, 64,* 81–93.

Demarest, H., Robertson, S., Haggard, M., Martin-Seaver, M., & Bickel, J. (2017). Similarity and enjoyment: Predicting continuation for women in philosophy. *Analysis, 77*(3), 525–541.

Devine, A., Fawcett, K., Szucs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, *8*(1), 33.

Drożdżowicz, A. (2018). Philosophical expertise beyond intuitions. *Philosophical Psychology, 31*(2), 253–277.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103–127.

Evans, J. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255–278.

Feldman, R., & Conee, E. (2001). Internalism defended. *American Philosophical Quarterly, 38*(1), 1–18.

Feltz, A., & Cokeley, E. T. (2012). The philosophical personality argument. *Philosophical Studies, 161*(2), 227–246.

Figdor, C., & Drabek, M. L. (2016). Experimental philosophy and the underrepresentation of women. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 590–602). West Sussex: Wiley Blackwell.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Ganley, C. M., & Vasilyevam, M. (2014). The role of anxiety and working memory in gender differences in mathematics. *Journal of Educational Psychology, 106*(1), 105–120.

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science, 336*(6080), 493–496.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis, 23*(6), 121–123.

Goguen, S. (2016). Stereotype threat, epistemic injustice, and rationality. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 1: Metaphysics and epistemology* (pp. 216–237). Oxford: Oxford University Press.

Goldman, A. I. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy, 73*(20), 771–791.

Grundmann, T. (2010). Some hope for intuitions: A reply to Weinberg. *Philosophical Psychology, 23*(4), 481–509.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880), 1164–1165.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York: Random House.

Harding, S. G. (1991). *Whose science? Whose knowledge? Thinking from women's lives*. Ithaca, NY: Cornell University Press.

Haslanger, S. (2008). Changing the ideology and culture of philosophy: Not by reason (alone). *Hypatia, 23*(2), 210–223.

Haslanger, S. (2012). *Resisting reality: Social construction and social critique*. Oxford: Oxford University Press.

Hassoun, N. (2017). How to improve the situation for women in philosophy. *Blog of the APA* [Blog]. 25 May 2017. https://blog.apaonline.org/2017/05/25/how-to-improve-the-situation-for-women-in-philosophy/. Accessed 02/07/2018.

Healy, K. (2011). Gender divides in philosophy and other disciplines. *Kieran Healy* [Blog]. 4 February 2011. https://kieranhealy.org/blog/archives/2011/02/04/gender-divides-in-philosophy-and-other-disciplines/. Accessed 02/07/2018.

Hermanson, S. (2017). Implicit bias, stereotype threat, and political correctness in philosophy. *Philosophies, 2*(2), 12.

Horvath, J., & Wiegmann, A. (2016). Intuitive expertise and intuitions about knowledge. *Philosophical Studies, 173*(10), 2701–2726.

House 51 (n.d.). Brexit bites. Data collected Aug 16. http://www.house51.co.uk/mishmash/brexit-bites-part-1-the-cognitive-refection-test/. Accessed 27/06/2018.

Hutchison, H. (2013). Sages and cranks: The difficulty of identifying first-rate philosophers. In K. Hutchison & F. Jenkins (Eds.), *Women in philosophy: What needs to change?* (pp. 103–126). Oxford: Oxford University Press.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321,* 494–495.

Inglis, M., & Attridge, N. (2016). *Does mathematical study develop logical thinking? Testing the theory of formal discipline*. London: World Scientific Publishing Europe Ltd.

Jenkins, F. (2013). Singing the post-discrimination blues: Notes for a critique of academic meritocracy. In K. Hutchison & F. Jenkins (Eds.), *Women in philosophy: What needs to change?* (pp. 81–102). Oxford: Oxford University Press.

Jimenez, N., Rodriguez-Lara, I., Tyran, J., & Wengström, E. (2018). Thinking fast, thinking badly. *Economics Letters, 162,* 41–44.

Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. Oxford: Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. London: Penguin.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge: Cambridge University Press.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment, 5,* 129–145.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two system theories. *Perspectives on Psychological Science, 4*(6), 533–550.

Larvor, B. (2015). Performance. *Show and Tell* [Blog]. 26 April 2015. https://manifestvirtue.wordpress.com/2015/04/26/performance/. Accessed 30/06/2018.

Lemoine, P. (2017). Why are women underrepresented in philosophy and should we care? *Nec Pluribus Impar* [Blog]. 4 June 2017. https://necpluribusimpar.net/women-underrepresented-philosophy-care/. Accessed 08/08/2017.

Leslie, S., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science, 347*(6219), 262–265.
Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123–1135.
Livengood, J., Sytsma, J., Feltz, A., Scheines, R., & Machery, E. (2010). Philosophical temperament. *Philosophical Psychology, 23*(3), 313–330.
Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies In Philosophy, 31*(1), 128–159.
Luzadis, R., Wesolowski, M., & Snavely, B. K. (2008). Understanding criterion choice in hiring decisions from a prescriptive gender bias perspective. *Journal of Managerial Issues, 20*(4), 468–484.
Nozick, R. (1981). *Philosophical explanations*. Cambridge, MA: Harvard University Press.
OECD (2016). *PISA 2015 Results (Volume 1): Excellence and equity in education*. Paris: OECD Publishing. https://www.oecd-ilibrary.org/docserver/9789264266490-en.pdf?expires=1538575255&id=id&accname=guest&checksum=62915A75FCB8315CD5E0A148216AF2CC. Accessed 03/10/2018.
OECD (2018). *PISA FAQ*. http://www.oecd.org/pisa/pisafaq/. Accessed 03/10/2018.
Oksman, O. (2016). Are gendered toys harming childhood development? *Guardian Online*. 28 May 2016. https://www.theguardian.com/lifeandstyle/2016/may/28/toys-kids-girls-boys-childhood-development-gender-research. Accessed 08/08/2017.
Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.
Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–177.
Peña-Guzmán, D. M., & Spera, R. (2017). The philosophical personality. *Hypatia, 32*(4), 911–927.
Penner, A. J. (2008). Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. *American Journal of Sociology, 114,* S138–S170.
Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods, 48*(1), 341–348.
Peters, E. (2012). Beyond comprehension the role of numeracy in judgments and decisions. *Current Directions in Psychological Science, 21*(1), 31–35.
Phillips, K. W., Northcraft, G. B., & Neale, M. A. (2006). Surface-level diversity and decision-making in groups: When does deep-level similarity help? *Group Processes and Intergroup Relations, 9*(4), 467–482.
Pinillos, N. A., Smith, N., Nair, G. S., Marchetto, P., & Mun, C. (2011). Philosophy's new challenge: Experiments and intentional action. *Mind and Language, 26*(1), 115–139.
Primi, C., Donati, M. A., Chiesi, F., & Morsanyi, K. (2018). Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Thinking & Reasoning, 24*(2), 258–279.
Rini, R. A. (2015). How not to test for philosophical expertise. *Synthese, 192*(2), 431–452.
Rubin, H., & O'Connor, C. (2018). Discrimination and collaboration in science. *Philosophy of Science, 85*(3), 380–402.
Saul, J. (2013). Implicit bias, stereotype threat, and women in philosophy. In K. Hutchison & F. Jenkins (Eds.), *Women in philosophy: what needs to change?* (pp. 39–60). Oxford: Oxford University Press.
Schulz, E., Cokely, E. T., & Feltz, A. (2011). Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense. *Consciousness and Cognition, 20*(4), 1722–1731.
Schwitzgebel, E. (2010). On being good at seeming smart. *The Splintered Mind* [Blog]. 25 March 2010. http://schwitzsplinters.blogspot.com/2010/03/on-being-good-at-seeming-smart.html. Accessed 25/06/2018.
Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind and Language, 27*(2), 135–153.
Singer, P. (1972). Moral experts. *Analysis, 32*(4), 115–117.
Sosa, E. (2009). A defense of the use of intuitions in philosophy. In M. Bishop & D. Murphy (Eds.), *Stich and his critics* (pp. 101–112). Oxford: Wiley.
Stahl, T., & van Prooijenb, J. (2018). Epistemic rationality: Skepticism toward unfounded beliefs requires sufficient cognitive ability and motivation to be rational. *Personality and Individual Differences, 122,* 155–163.
Stanovich, K. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.
Stanovich, K., & West, R. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 22,* 645–726.

Storage, D., Horne, Z., Cimpian, A., & Leslie, S. (2016). The frequency of 'brilliant' and 'genius' in teaching evaluations predicts the representation of women and African Americans across fields. *PLoS ONE, 11*(3), e0150194.

Stupple, E. J. N., Pitchford, M., Ball, L. J., Hunt, T. E., & Steel, R. (2017). Slower is not always better: Response-time evidence clarifies the limited role of miserly information processing in the Cognitive Reflection Test. *PLoS ONE, 12*(11), e0186404.

Summers, L. H. (2005). Remarks at NBER conference on diversifying the science and engineering workforce. NBER conference on diversifying the science & engineering workforce, 14 January 2005. https://www.harvard.edu/president/speeches/summers_2005/nber.php. Accessed 02/07/2018.

Sytsma, J. (2016). Are religious philosophers less analytic? *LSE Popper Seminar,* 15 November 2016. Abstract http://www.lse.ac.uk/philosophy/events/justin-sytsma-are-religious-philosophers-any-less-analytic/. Accessed 30/01/17.

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: Exploring the ways individuals solve the test. *Thinking & Reasoning, 23*(3), 207–234.

Thompson, M. (2017). Explanations of the gender gap in philosophy. *Philosophy Compass, 12,* 1–12.

Thompson, M., Adleberg, T., Sims, S., & Nahmias, E. (2016). Why do women leave philosophy? Surveying students at the introductory level. *Philosophers' Imprint, 16*(6), 1–36.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making, 11*(1), 99–113.

Tobia, K., Buckwalter, W., & Stich, S. (2013). Moral intuitions: Are philosophers experts? *Philosophical Psychology, 26*(5), 629–638.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition, 39,* 1275–1289.

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science, 16*(6), 474–480.

van Mulukom, V. (2018). Is it rational to trust your gut feelings? A neuroscientist explains. *The Conversation.* 16 May 2018. http://theconversation.com/is-it-rational-to-trust-your-gut-feelings-a-neuroscientist-explains-95086. Accessed 01/07/2018.

Wang, Y., Highhouse, S., Lake, C. J., Petersen, N. L., & Rada, T. B. (2017). Meta-analytic investigations of the relation between intuition and analysis. *Behavioral Decision Making, 30*(1), 15–25.

Weinberg, J. (2015). Confidence and performance in philosophy. *Daily Nous* [Blog]. 30 April 2015. http://dailynous.com/2015/04/30/confidence-performance-in-philosophy/. Accessed 30/06/2018.

Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters? *Philosophical Psychology, 23*(3), 331–355.

Williamson, T. (2007). *The philosophy of philosophy*. Oxford: Blackwell.

Williamson, T. (2011). Philosophical expertise and the burden of proof. *Metaphilosophy, 42*(3), 215–229.

Winegard, B., Winegard, B., & Boutwell, B. (2017). Human biological and psychological diversity. *Evolutionary Psychological Science, 3*(2), 159–180.

Wylie, A. (2011). Women in philosophy: The costs of exclusion—editor's introduction. *Hypatia, 26*(2), 374–382.

Zhang, D. C., Highhouse, S., & Thaddeus, B. R. (2016). Explaining sex differences on the cognitive reflection test. *Personality and Individual Differences, 101,* 425–427.

Zollman, K. J. S. (2010). The epistemic benefit of transient diversity. *Erkenntnis, 72*(1), 17–35.